

MOLECULAR ANALYSIS OF HUMAN UBIQUITIN GENES

A thesis submitted for the degree of
Doctor of Philosophy
in the
Australian National University

by

ROHAN THOMAS BAKER

John Curtin School of Medical Research
The Australian National University
Canberra

March 1988

This thesis describes the results of research
conducted in the Human Genetics Group, Division of
Clinical Research, John Curtin School of Medical Research,
Australian National University, Canberra, between March
1985 and March 1988, while I received a Commonwealth
Postgraduate Research Award.

"One of the advantages of being disorderly is that
one is constantly making exciting discoveries."

A.A. Milne

STATEMENT

This thesis describes the results of research undertaken in the Human Genetics Group, Division of Clinical Sciences, John Curtin School of Medical Research, Australian National University, Canberra, between March 1985 and March 1988, while I received a Commonwealth Postgraduate Research Award.

The results and analysis presented in this thesis are my own original work accomplished under the supervision of Dr Philip G. Board, except when otherwise acknowledged in the text or Acknowledgements.



Rohan Baker

ACKNOWLEDGEMENTS

I am deeply grateful to my supervisor Dr Philip G. Board for his advice and encouragement throughout the course of this study, and especially during the preparation of this thesis.

I also wish to thank Drs Robert L. Kirk and Susan W. Serjeantson as past and present head of this department for their interest and help.

I am indebted to Drs Don Anson, David Bentley, R.H. Don, and Geoffrey Howlett for their generous gifts of recombinant libraries, which made these studies possible.

I also thank the following colleagues and friends for their willingness and help in many ways: Dr Graham Webb for his expert in situ hybridisation studies, Dr Simon Easteal for assistance with statistical analysis, Ms Marjorie Coggan for help with population studies, thesis proof-reading and for the smooth running of the lab., and Ms Catherine Merritt, Ms Kerrie Pierce, Ms Robin Chapple, Dr Pa-thai Yenchitsomanus, Mr Peter Johnston and Mr Brett White for help on many occasions.

Thanks are due to the School technical support and photographic staff for their excellent services. I also gratefully acknowledge the financial support of a Commonwealth Postgraduate Research Award.

Finally, I thank my parents and family for their continued support, while the largest debt of all is to my loving wife Chelsey, for her patience, love and care, and lastly for the typing of this thesis.

ABSTRACT

1. The aim of this study was to characterise the structural organisation of the human ubiquitin gene family, in particular the UbB and UbA₅₂ subfamilies, about which no information was available. This characterisation was performed by the isolation and sequence analysis of ubiquitin cDNA and genomic clones.

2. A UbB cDNA clone was isolated by immunological screening of a liver cDNA library. This clone contained two complete, and a portion of a third, ubiquitin coding units in a tandem repeat structure. This cDNA was used to isolate the corresponding UbB gene from a genomic library. This gene contained three tandem repeats of the ubiquitin coding unit and a single 715 base pair intron within the 5' non-coding region. A heat shock promoter region was also identified upstream of the gene, suggesting its involvement in the stress response. Genomic hybridisation analysis revealed the presence of several UbB-related sequences in the genome. Three of these were isolated and sequenced to reveal UbB processed pseudogenes, which differed from the gene in their number of coding units: one contained two, while the other two only contained one coding unit. A fourth related sequence appears to be a duplicated UbB gene, but was not isolated.

3. The studies on the UbB subfamily also resulted in the isolation of a pseudogene of the ubiquitin/tail fusion gene subfamily, UbA₅₂. This pseudogene was used to isolate UbA₅₂

cDNA clones from placental and adrenal gland cDNA libraries to reveal that the sequence of the human 52 amino acid tail protein was very similar to the known yeast and slime mould proteins and also contained a putative metal-binding, DNA-binding domain. These cDNA clones were used to isolate the corresponding UbA₅₂ gene. While the characterisation of the 5' non-coding region of this gene is incomplete, it contains at least 5 exons separated by 4 introns distributed over 3400 base pairs. One intron is in the 5' non-coding region, two interrupt the single ubiquitin coding unit, while the fourth is within the tail coding region. Genomic hybridisation revealed the presence of several UbA₅₂-related sequences in the genome. One such sequence was isolated and sequenced to reveal a UbA₅₂ processed pseudogene, while re-examination of the first pseudogene revealed that it too was processed but had suffered a deletion of its 3' portion.

4. A putative variation in the number of coding units in the UbC gene was identified, with 7, 8, and 9 coding unit alleles postulated. This variation is exhibited as a restriction fragment length polymorphism at the DNA level which correlates with a UbC mRNA length polymorphism, both of which are inherited in Mendelian fashion. It is likely that the RFLPs result from unequal crossover events giving rise to varying numbers of ubiquitin coding unit repeats.

5. In situ hybridisation with the UbB gene intron specifically localised the duplicated UbB genes to chromosome band 17p11.2. Similar hybridisation with the intact UbB cDNA

clone also labelled this region, but also indicated several other ubiquitin loci on other chromosomes which may correspond to UbB processed pseudogenes and members of other ubiquitin subfamilies.

6. Several members of the Alu family of repetitive DNA sequences were identified in association with some members of the UbB and UbA₅₂ subfamilies during sequence analysis. The significance of this association is presently unknown.

7. These studies have provided a detailed characterisation of the human ubiquitin UbB and UbA₅₂ subfamilies, which like many other higher vertebrate gene families appear to be composed mainly of processed pseudogenes. They also provide an example of a member from each ubiquitin subfamily in a species other than yeast, which will allow a study of the mechanisms of ubiquitin gene evolution and expression.

PUBLICATIONSOriginal Papers

- Baker, R.T. and Board, P.G. (1987). The human ubiquitin gene family: structure of a gene and pseudogenes from the UbB subfamily. *Nucl. Acids Res.* 15, 443-463.
- Baker, R.T. and Board, P.G. (1987). Nucleotide sequence of a human ubiquitin UbB processed pseudogene. *Nucl. Acids Res.* 15, 4352.
- Baker, R.T. and Board, P.G. (1988). An improved method for mapping recombinant λ phage clones. *Nucl. Acids Res.* 16, 1198.

Abstracts of papers presented at Conferences:

- Baker, R.T. and Board, P.G. (1986). The human ubiquitin gene: a novel structure for one of evolution's most conserved proteins. (The Genome Conference, 8th Annual Meeting, February 17-21, Lorne, Victoria.) Programme and Abstracts, p. 46.
- Baker, R.T. and Board, P.G. (1986). Novel structural features of the human ubiquitin gene family. (Human Genetics Society of Australasia, 10th Annual Scientific Meeting, May 8-10, Canberra.) Programme and Abstracts, p. 16.
- Baker, R.T. and Board, P.G. (1986). The ubiquitin multigene family: analysis of the UbB subfamily. (Australian Biochemical Society, 30th Annual Meeting, May 13-16, Melbourne.) *Proc. Aust. Biochem. Soc.* 18, p. 48.
- Baker, R.T. and Board, P.G. (1987). The human ubiquitin gene family: structure of a gene and pseudogenes of the UbB subfamily. (The Genome Conference, 9th Annual Meeting, February 16-20, Lorne, Victoria.) Programme and Abstracts, p P45.
- Baker, R.T., Webb, G.C. and Board, P.G. (1987). The human ubiquitin gene family: chromosomal location and molecular characterisation of the UbB subfamily. (Genetics Society of Australia, 34th Annual Conference, August 26-29, Canberra.) Programme and Abstracts, p33.

TABLE OF CONTENTS

STATEMENT.....	i
ACKNOWLEDGEMENTS.....	ii
ABSTRACT.....	iii
PUBLICATIONS.....	vi
TABLE OF CONTENTS.....	vii
ABBREVIATIONS.....	xii
<u>CHAPTER 1: INTRODUCTION</u>	
1.1 Ubiquitin	1
1.2 Ubiquitin-Protein Conjugates	4
1.3 Ubiquitin-Mediated Proteolysis	8
1.4 Ubiquitin and Histone Conjugation	12
1.5 Ubiquitin and the Stress Response	16
1.6 Ubiquitin as a Functional Component of Cell Surface Receptors	18
1.7 Other Ubiquitin Conjugates	21
1.8 Ubiquitin Gene Structure	21
1.9 Ubiquitin Gene Expression	25
1.10 Aims	28
<u>CHAPTER 2: MATERIALS AND METHODS</u>	
2.1 MATERIALS	
2.1.1 Reagents and Materials	30
2.1.2 Enzymes	31
2.1.3 Media for Bacterial Growth	31
2.1.4 Bacterial Strains, Bacteriophage, Plasmids and Recombinant Libraries	32
2.2 METHODS	
2.2.1 Routine Laboratory Procedures	33
2.2.2 Biological Containment and Radiation Safety	34
2.2.3 Gel Electrophoresis	34
2.2.4 Preparation of Plasmid and Bacteriophage DNA	35
2.2.5 Restriction Endonuclease Digestion ...	36
2.2.6 Subcloning	37

2.2.7 Preparation and Transformation of Competent Cells	38
2.2.8 Identification and Characterisation of Recombinants	38
2.2.9 Construction and Screening of a Human Liver cDNA Expression Library	39
2.2.10 Radioactive Labelling and Probe Preparation	40
2.2.11 Phage Library Screening and Hybridisation	41
2.2.12 Preparation of DNA and RNA from Human Tissues	42
2.2.13 Nucleic Acid Capillary Transfer and Hybridisation	43
2.2.14 Restriction Mapping of Recombinant λ Phage Inserts	43
2.2.15 DNA Sequencing and Computer-Assisted DNA Sequence Analysis	44
2.2.16 S1 Nuclease Mapping and Primer Extension Analysis	45
 <u>CHAPTER 3: THE HUMAN UBIQUITIN UbB SUBFAMILY</u>	
3.1 Isolation and Sequence Analysis of a Human Liver Ubiquitin cDNA Clone.	47
3.2 Isolation and Characterisation of Human Ubiquitin Genomic Clones	49
3.3 Sequence Analysis of a UbB Processed Pseudogene EHB4	50
3.4 Sequence Analysis of a UbB Processed Pseudogene EHD1	51
3.5 Sequence Analysis of a UbB Processed Pseudogene EHB7	52
3.6 Sequence Analysis of a Human UbB Gene EHB8.	
3.6.1 Characterisation of Coding and 3' Non- coding Regions	53
3.6.2 Characterisation of the 5' Non-coding Region	56
3.6.3 Characterisation of an Intron Within the 5' Non-coding Region	57

3.6.4 Determination of the Transcription Initiation Site	58
3.6.5 Specific Hybridisation of the EHB8 Gene to the UbB mRNA	60
3.6.6 Identification of Promoter and Enhancer Elements Upstream of the UbB Gene	61
3.6.7 Identification of <u>Alu</u> Repetitive DNA Sequences Upstream of the UbB Gene	62
3.7 Isolation of Full-Length Human Liver UbB cDNA Clones	63
3.8 Hybridisation Analysis of Total Human Genomic DNA	64
3.9 UbB Gene and Pseudogene Comparisons	
3.9.1 Comparison of Coding Regions	68
3.9.2 Comparison of Non-Coding Regions	68
3.9.3 Results of UbB Gene and Psuedogene Comparisons	69
3.10 Structural Similarities of the Human UbB and Chicken UbI Ubiquitin Genes	72
3.11 Discussion	
3.11.1 UbB Gene Structure	73
3.11.2 Polyubiquitin Gene Structure	76
3.11.3 UbB Translation and Post- translational Processing	79
3.11.4 UbB Processed Pseudogene Structure and Creation	82
3.11.5 Ubiquitin cDNA Clone Structure.....	86
3.11.6 The UbB gene and Heat Shock	89
3.11.7 The UbB Subfamily	91
 CHAPTER 4: THE HUMAN UBIQUITIN UbA₅₂ SUBFAMILY	
4.1 Isolation and Sequence Analysis of a Human UbA ₅₂ Pseudogene, EHD5	93
4.2 Isolation and Sequence Analysis of a Human Placental UbA ₅₂ cDNA Clone.	94
4.3 Isolation and Sequence Analysis of Human Adrenal Gland UbA ₅₂ cDNA Clones.	95

4.4 Specific Hybridisation of UbA ₅₂ to the UbA mRNA	96
4.5 Isolation of UbA ₅₂ -Homologous Genomic Clones.	96
4.6 Sequence Analysis of a UbA ₅₂ Processed Pseudogene, λUA4.	97
4.7 Characterisation of a Human UbA ₅₂ Gene	
4.7.1 Sequence Analysis of a UbA ₅₂ Gene, λUA1	98
4.7.2 Characterisation of the UbA ₅₂ Gene 5' Non-Coding Region	100
4.7.3 Identification of Putative Promoter Elements Upstream of Exon 1	103
4.8 Comparison of UbA ₅₂ Gene and Pseudogene Regions	
4.8.1 Comparison of Coding and Non-Coding Regions	105
4.8.2 Features of the UbA ₅₂ Tail Protein ...	108
4.9 Hybridisation Analysis of Human Genomic DNA	109
4.10 Discussion	
4.10.1 UbA ₅₂ Gene Structure - Exons and Introns	111
4.10.2 UbA ₅₂ Gene Structure - Promoter Elements	115
4.10.3 Features of the Fused Ubiquitin/Tail Protein	117
4.10.4 The UbA ₅₂ Subfamily.....	118

CHAPTER 5. LENGTH POLYMORPHISM AT THE HUMAN UbC LOCUS

5.1 Identification of a Putative UbC mRNA Length Polymorphism	121
5.2 Correlation of the UbC mRFLP with a Restriction Fragment Length Polymorphism (RFLP), Identification of a Second UbC mRFLP/RFLP and Demonstration of Mendelian Inheritance	122
5.3 Calculation of Gene Frequency	124
5.4 Discussion	125

<u>CHAPTER 6: CHROMOSOMAL LOCALISATION OF THE HUMAN</u>	
UBIQUITIN UbB GENE	
6.1 Introduction, Materials and Methods	130
6.2 <u>In situ</u> Hybridisation of a UbB cDNA to Human Chromosomes.....	131
6.3 <u>In situ</u> Hybridisation of the UbB Intron to Human Chromosomes	132
6.4 Discussion	132
 <u>CHAPTER 7: ASSOCIATION OF <u>Alu</u> REPETITIVE DNA</u>	
SEQUENCES WITH UBIQUITIN GENES AND PSEUDOGENES	
7.1 Analysis of <u>Alu</u> Repeats Associated with the UbB Gene and Pseudogenes	135
7.2 Analysis of <u>Alu</u> Repeats Associated with the UbA ₅₂ Gene	138
7.3 Association of <u>Alu</u> Repetitive DNA Sequences with the UbB and UbA ₅₂ Subfamilies	139
 <u>CHAPTER 8: GENERAL DISCUSSION AND CONCLUSIONS</u>	
8.1 The Human Ubiquitin Gene Family	142
8.2 Ubiquitin Gene Evolution	144
8.3 Future Studies	147
8.4 Conclusions	150
 <u>REFERENCES</u>	 151

ABBREVIATIONS

Abbreviations used were generally those listed in the "Policy of the Journal and Instructions to Authors", Biochem. J. 169, 1-27. (1978). The following list is of abbreviations commonly used in this thesis.

bp	base pair(s)
cDNA	DNA complementary to RNA
C-terminal	carboxyl terminal
DNaseI	deoxyribonuclease I
dNTP	3' deoxynucleoside triphosphate
ddNTP	2', 3' dideoxynucleoside triphosphate
DTT	dithiothreitol
EDTA	ethylenediaminetetra-acetic acid
HF	hybridising fragment
IAA	isoamyl alcohol
IPTG	isopropyl- β -D-thiogalactopyranoside
kb	10^3 base pair(s)
mRNA	messenger RNA
N,n	any nucleotide
NCR	non-coding region
nt	nucleotide
N-terminal	amino-terminal
PAGE	polyacrylamide gel electrophoresis
R,r	purine (A,G)
RE	restriction endonuclease
RF	replicative form
rRNA	ribosomal RNA
ss	single stranded
Tris	tris(hydroxymethyl)aminomethane
tRNA	transfer RNA
Y,y	pyrimidine (C,T)
X-GAL	5-bromo-4-chloro- β -D-galactoside

CHAPTER 1

INTRODUCTION

CHAPTER 1: INTRODUCTION

1.1 Ubiquitin

Ubiquitin is a small polypeptide which appears to be present in all eukaryotic cells. It was first isolated in 1975 from bovine thymus during studies on the polypeptide hormone thymopoietin (Goldstein et al, 1975). Goldstein and co-workers purified an 8500 Dalton polypeptide which induced the in vitro differentiation of both precursor T (thymus-derived) and B (bone-marrow-derived) cells into mature immunocytes, while thymopoietin specifically induces the in vitro differentiation of T cells. The same authors also reported the adenylate cyclase activating properties of the polypeptide, and its occurrence in a wide range of organisms (as determined by radioimmunoassay) including animal cells, higher plants, yeast and bacteria (Goldstein et al, 1975). This widespread occurrence, and its ability to induce immunocyte differentiation, led to the name ubiquitous immunopoietic polypeptide, or UBIP. However, it was concluded that UBIP was unlikely to function as a hormone in vivo, due to its presence in a wide range of animal tissues (Goldstein et al, 1975). The claims of UBIP-induced lymphocyte differentiation and adenylate cyclase activation have not been confirmed by other investigators (Low et al, 1979; Low and Goldstein, 1979).

Also in 1975, Goldstein's group reported the amino acid sequences of the bovine and human protein, and renamed it ubiquitin (Schlesinger et al, 1975; Schlesinger and Goldstein, 1975). From these studies, ubiquitin was found to

be a 74 amino acid (aa) polypeptide of identical sequence in these two species. This sequence identity confirmed the previous observation of complete immunological cross-reactivity between the human and bovine proteins (Goldstein et al, 1975). Later studies revealed that ubiquitin consisted of 76aa, being extended at the C-terminus by two glycine residues, which were presumably lost as a proteolytic artefact during its isolation (Wilkinson and Audhya, 1981). The sequence of ubiquitin has been determined in cattle, man, trout (Watson et al, 1978), fruit fly (Gavilanes et al, 1982), oat (Vierstra et al, 1986) and yeast (Wilkinson et al, 1986), and has been deduced from cDNA and genomic sequences from Xenopus (Dworkin-Rastl et al, 1984), yeast (Ozkaynak et al, 1984; 1987), chicken (Bond and Schlesinger, 1985; 1986), man (Lund et al, 1985; Wiborg et al, 1985), barley (Gausung and Barkardottir, 1986), mouse (St John et al, 1986), slime mould (Westphal et al, 1986; Giorda and Ennis, 1987), Drosophila (Arribas et al, 1986), and pig (Einspanier et al, 1987a). These sequences are presented in Figure 1.1.1. Remarkably, the aa sequence is identical in all animal species ranging from fly to man, while yeast, plant and slime mould ubiquitin differ from the animal protein at 3, 3, and 2 or 3 positions respectively, with slime mould exhibiting heterogeneity at one position. The yeast and plant ubiquitins differ at only two positions, while the slime mould protein differs at 3 or 4 and 4 or 5 positions from yeast and plant ubiquitin respectively. These observations have led to the conclusion that ubiquitin is perhaps the most conserved protein known, even more so than histone 4, generally

5 10 15
Met-Gln-Ile-Phe-Val-Lys-Thr-Leu-Thr-Gly-Lys-Thr-Ile-Thr-Leu-Glu-

20 25 30
Val-Glu-Pro-Ser-Asp-Thr-Ile-Glu-Asn-Val-Lys-Ala-Lys-Ile-Gln-Asp-
Ser(P,Y) Asn(SM) Asp(P,Y) Ser(Y)
Gly(SM) Thr(SM)

35 40 45
Lys-Glu-Gly-Ile-Pro-Pro-Asp-Gln-Gln-Arg-Leu-Ile-Phe-Ala-Gly-Lys-

50 55 60
Gln-Leu-Glu-Asp-Gly-Arg-Thr-Leu-Ser-Asp-Tyr-Asn-Ile-Gln-Lys-Glu-
Ala(P)

65 70 75
Ser-Thr-Leu-His-Leu-Val-Leu-Arg-Leu-Arg-Gly-Gly

Figure 1.1.1: Amino Acid Sequence of Ubiquitin.

The amino acid sequence of ubiquitin from several species as determined by protein sequencing and derived from cDNA and genomic sequences from insect to human is listed. Differences in ubiquitin from other organisms are shown under the sequence and are from oat and barley (P), yeast (Y) and the slime mould Dictyostelium discoideum (SM). The slime mould variant Thr-28 residue (underlined) is present in three coding units of the transcriptionally active five coding unit polyubiquitin gene (see Chapter 1.8). See text for references.

considered the most slowly evolving protein (Sharp and Li, 1987a). This extraordinary conservation implies extreme functional constraint on the protein. Ubiquitin's functions are discussed in the following sections.

The three-dimensional structure of human ubiquitin has been determined at 2.8Å resolution by X-ray diffraction studies (Vijay-Kumar et al, 1985), and has since been refined by ¹H NMR assignment (Di Stefano and Wand, 1987; Weber et al, 1987) and further crystal diffraction studies to 1.8Å resolution (Vijay-Kumar et al, 1987a). These studies reveal a very compact and tightly hydrogen-bonded protein, with ~87% of the residues involved in hydrogen-bonded secondary structure, comprised of a 5-stranded mixed β sheet, a 3.5 turn α helix, a short helical turn, and 7 reverse turns. Notable residues excluded from the secondary structure are the four C-terminal aa, which extend away from the molecule, are not involved in any secondary structure, and are very flexible in solution. The availability of the C-terminus is critical to ubiquitin's functions as described below. Conversely, the N-terminal methionine is involved in structural hydrogen-bonding and is buried in the molecule. The extensive secondary structure of ubiquitin is consistent with previous observations of its exceptional stability with respect to heat denaturation, pH extremes or denaturing agents. For example, ubiquitin maintains its structure between pH 1 and 13, and below 80°C (Cary et al, 1980).

Plant and yeast ubiquitin have also been studied by X-ray diffraction, to reveal three-dimensional structures very similar to the human protein, relatively unaffected by the aa

substitutions (Vijay-Kumar et al, 1987b). Notably, the observed interspecies aa differences are clustered in a small area on the surface of the molecule.

1.2 Ubiquitin-Protein Conjugates

To date, all of ubiquitin's known functions are mediated via its covalent attachment to other proteins (see Chapters 1.3 to 1.7). All such conjugates are formed by an isopeptide linkage between the C-terminal glycine of ubiquitin and a free amino group of the target protein, which may be its N-terminus and/or a lysine ϵ -NH₂ group. The enzymatic reactions leading to the formation of ubiquitin-protein conjugates have been characterised during the analysis of ubiquitin-dependent proteolysis in reticulocyte extracts (see Chapter 1.3) mainly through studies by Ciechanover, Haas, Hershko, Rose and co-workers. The system has also been recently studied in yeast by Varshavsky and colleagues as discussed later. The reactions of the ubiquitin-specific enzymes have been detailed in recent reviews by Finley and Varshavsky (1985), Hershko and Ciechanover (1986), and Rechsteiner (1987), and constitute a multi-step pathway. The first stage involves the activation of ubiquitin via a two-step reaction sequence catalysed by a ubiquitin-activating enzyme, E1 (Ciechanover et al, 1981). The first step involves the formation of a ubiquitin-adenylate-E1 complex and is the ATP-dependent step of the activation system. The adenylate is then transferred to an E1 sulfhydryl group to yield the activated E1-S-ubiquitin thioester. The ubiquitin aa residue thus activated is the C-terminal glycine (Hershko et al, 1981). This

reversible covalent linkage has been exploited in a single step "covalent-affinity" chromatography method employing ubiquitin-Sepharose to obtain near-homogenous E1 as a homodimer of 105,000 kDa subunits (Ciechanover et al, 1982). Studies involving purified E1 and other components of the system revealed that the activated E1-S-ubiquitin complex could not perform ubiquitin conjugation by itself, but required two other enzymes termed E2 and E3 (Hershko et al, 1983).

It is now known that E2 may be one of a family of enzymes which act as carrier proteins, transferring activated ubiquitin as a thioester from E1 to various target proteins. Five E2 species of 32, 24, 20, 17 and 14 kDa subunits have been identified and termed E2₁ to E2₅ respectively (Pickart and Rose, 1985a). Initial studies indicated that a third enzyme, E3, was required to catalyse the final transfer of activated ubiquitin from an E2 carrier to an amide linkage with target proteins destined for proteolysis (Hershko et al, 1983). However, it now appears that some E2s can transfer activated ubiquitin to certain proteins that are not substrates for proteolysis independently of E3 (Pickart and Rose, 1985a, and see below). Such proteins include specific histone types (see Chapter 1.4) and are generally monoubiquitinated, whereas E3-catalysed ubiquitination tends to produce multiubiquitin substrates for proteolysis. As such there is functional heterogeneity amongst the E2 species: E2₅ specifically ubiquitinates proteolytic substrates in an E3-dependent manner, while E2₃ can directly ubiquitinate histones (Pickart and Rose, 1985a).

Recently, investigation of the analogous enzyme system in the yeast Saccharomyces cerevisiae has been undertaken by Varshavsky and co-workers. Yeast possesses a single E1 and 5 E2 enzymes of practically identical sizes to their mammalian counterparts, a not unexpected similarity given the extraordinary inter-species conservation of the protein with which they interact (Jentsch et al, 1987). A most interesting discovery resulted from the cloning and sequencing of the gene coding for the 20kDa E2 species (the yeast analogue of mammalian E2₃): this E2 enzyme is the product of the previously identified yeast RAD6 gene. RAD6 is required for several cellular functions including DNA repair, induced mutagenesis, and sporulation (see Jentsch et al, 1987). Additionally, the RAD6/20kDa E2 protein mediated the specific in vitro ubiquitination of histone H2B in the presence of ATP and purified yeast E1 enzyme: ie, in an E3-independent fashion. The authors suggest that the multiple functions of RAD6 are mediated by its ubiquitin conjugating activity, for example ubiquitinating chromosomal proteins to induce the necessary alterations in chromatin structure required for DNA repair and sporulation. Also reported are: (i) homology between the E2 series of enzymes; (ii) homology between the yeast cell-division cycle gene CDC34 protein and the E2 enzymes, implying its membership in this family of enzymes, and (iii) the cloning of the genes for the yeast 30kDa E2 enzyme (mammalian E2₁ analogue) and the E1 ubiquitin-activating enzyme (Jentsch et al, 1987). The availability of these genes and their products, plus the ability of combining biochemical and molecular genetic analyses in the yeast

system, should now allow the characterisation of the remaining components of the ubiquitin activating and conjugating system.

Other investigators have begun studies aimed at determining the specific regions of the ubiquitin molecule which interact with the enzymes of the ubiquitin system. Using ubiquitin iodinated at His-69, Cox et al (1986) showed that this derivative did not influence ubiquitin's interaction with the E1, E2 or E3 enzymes, but apparently destabilised the subsequently-formed conjugates and increased their rate of degradation. Further studies by Duerksen-Hughes et al (1987) using ubiquitin chemically modified at Arg residues revealed that modification at any one of Arg-42, -72 or -74 prevented the protein from stimulating ATP dependent proteolysis. However, different derivatives caused blockage at different stages. Arg-42 and Arg-72 derivatives failed to be activated by E1, while the Arg-74 derivative formed normal levels of conjugated substrates. The subsequent failure of these conjugates to be degraded is attributed to the interference of the Arg-74 derivative with the conjugate-specific proteases. Preliminary results also indicate that E1 is sensitive to modification at Tyr-58 (Duerksen-Hughes et al, 1987). These studies begin to describe the regions and conformations of ubiquitin recognised by the enzymes of the ubiquitin-dependent proteolysis system.

An alternative approach to the investigation of the regions of ubiquitin of importance for its structure and function is that of Ecker et al (1987a), who have chemically synthesised a ubiquitin coding region and successfully expressed it in

yeast to produce biologically active ubiquitin. Eight unique restriction sites were engineered into the "gene" to allow in vitro mutagenesis studies by the replacement of one "mutagenesis module" with an altered module encoding an aa substitution. Several variant ubiquitins have been thus expressed in Escherichia coli, purified, and subjected to preliminary structural and functional studies (Ecker et al, 1987b). Clearly, this technique provides a versatile method of generating specific variant ubiquitin proteins for several types of study.

The fundamental importance of ubiquitin conjugation to the function of the eukaryotic cell is dramatically demonstrated by the mammalian cell line ts85. This cell line has a temperature-sensitive ubiquitin activating (E1) enzyme (Finley et al, 1984), and upon shift-up to a nonpermissive temperature (39°C), generation of new mitotic cells is completely inhibited, and the cells arrest primarily in G2 phase, with a subset in late S phase (Mita et al, 1980; Yasuda et al, 1981). The phenotype exhibited by ts85 at the nonpermissive temperature provides insights into the various functions of the ubiquitin system as discussed below.

1.3 Ubiquitin-Mediated Proteolysis

Ubiquitin has been most extensively characterised as an essential component of the eukaryotic non-lysosomal ATP-dependent proteolysis system, mainly through individual and collaborative studies by Ciechanover, Haas, Hershko, Rose, Rechsteiner, and co-workers. This ubiquitin-mediated pathway has recently been comprehensively reviewed by Hershko and

Ciechanover (1986); Ciechanover (1987) and Rechsteiner (1987) and a summary of ubiquitin's involvement in this process is presented here.

Ubiquitin was first identified as a component of the reticulocyte ATP-dependent proteolysis system when a previously identified essential component of the system, ATP-dependent proteolysis Factor 1 (APF1) was found to be ubiquitin (Ciechanover et al, 1980; Wilkinson et al 1980). Further studies revealed that ubiquitin became covalently conjugated to protein substrates destined for proteolysis, leading to the hypothesis that ubiquitin conjugation signals the degradation of the conjugated protein by an enzyme system specifically recognising ubiquitinated proteins (Hershko et al, 1980). This hypothesis is supported by several studies revealing a correlation between a protein's ubiquitination and rapid degradation (see Hershko and Ciechanover, 1986; Rechsteiner, 1987). One such case is the ts85 cell-cycle mutant, whereby inactivation of E1 upon shift-up results in the stabilisation of most of the short-lived intracellular proteins that are normally rapidly degraded. This also suggests that the ubiquitin system is responsible for the selective turnover of the short-lived, regulatory proteins (Ciechanover et al, 1984; Finley et al, 1984). Direct evidence was obtained by studies on the breakdown of purified preformed ubiquitin-protein conjugates in a reconstituted reticulocyte lysate system (Hershko et al, 1984a). These studies showed that proteolysis of the pre-formed conjugate was independent of further ubiquitin conjugation, whereas the breakdown of the free substrate required ubiquitin conjugat-

ion. A further aspect of the ubiquitin-mediated proteolysis system was also revealed: an absolute requirement of ATP for the degradation of the conjugated substrate, which was shown to be separate from the ATP-dependent step in ubiquitin activation (see Chapter 1.2). Thus the ATP-dependence of the ubiquitin-mediated proteolysis system is twofold: first, the formation of the conjugate, and second, the breakdown of the ubiquitinated substrate. Rapoport et al (1985) estimate that more than one ATP molecule is hydrolysed per peptide bond cleaved.

While some of the enzymes involved in ubiquitin activation and conjugation have been isolated and characterised (Chapter 1.2), much less is known of the conjugate-specific proteolytic enzyme(s). Rechsteiner and co-workers have recently reported the identification and purification of a large ATP/ubiquitin dependent reticulocyte protease of $\sim 1000\text{kDa}$ (Hough et al, 1986, 1987). These authors also report the co-purification of a smaller ($\sim 700\text{kDa}$) protease which appears to be ATP independent and is not specific for ubiquitinated substrates. Both proteases exhibit extensive subunit structure. In addition, Hershko and Ciechanover (1986) have fractionated three components of the reticulocyte protease system, all of which are absolutely required for the ATP dependent proteolysis of ubiquitin conjugates. These investigations are in their initial stages, and further studies should improve our knowledge of the proteolytic components of the system.

Other features of the proteolytic system include specific roles for the α -amino group and transfer RNA (tRNA), recently

reviewed by Ciechanover (1987). From several experimental results, Hershko et al (1984a) suggest that the availability of an N-terminal α -amino group may be a requisite signal for proteolysis via the ubiquitin system, while free ϵ -amino groups are not essential, but may accelerate the rate of proteolysis. Recent evidence indicates that the N-terminal aa itself may influence the in vivo rate of protein degradation, with the protein's half-life a function of this residue: the N-end rule (Bachmair et al, 1986). By this rule, short-lived proteins have destabilising N-terminal residues, and are rapidly ubiquitinated and degraded, while long-lived proteins are apparently not ubiquitinated and hence stable.

Less well characterised is a role for tRNA molecules in the proteolysis of some substrates, stemming from the observation that the breakdown of some exogenously-added proteins is sensitive to the presence of ribonucleases (Ciechanover et al, 1985). Experiments employing fractionated RNA species indicated that the addition of tRNAs specifically restored activity to RNA-depleted systems, with the active component narrowed down to tRNA^{His}. Further investigation revealed that tRNA was a requirement for the conjugation of ubiquitin to some proteolytic substrates (Ferber and Ciechanover, 1986). While the mechanism behind this tRNA-dependence is unknown, Ciechanover (1987) has noted that all known affected proteolytic substrates have acidic N-termini, and that this feature may serve as a recognition marker for the tRNA-dependent reaction. He further speculates that this reaction may be a tRNA-dependent modification of the N-terminus by the addition

of an aa residue, an event with a known precedent (see Ciechanover, 1987).

A further class of enzymes involved in the proteolysis system are the so-called ubiquitin-protein lyases. The first type has an isopeptidase activity, cleaving ubiquitin from histone conjugates (Matsui et al, 1982) and also from potential proteolytic substrates in the absence of proteolysis following ATP-depletion (Hershko et al, 1980, 1984b). Whether the same isopeptidase catalyses both reactions is not known. The latter authors suggest that this enzyme may have a correction function, releasing ubiquitin from incorrectly ubiquitinated proteins not destined for proteolysis. A second type of lyase is the ubiquitin carboxy-terminal hydrolase, which has both a thioesterase activity (Rose and Warms, 1976), and the ability to cleave amide linkages between ubiquitin and small amines (Pickart and Rose, 1985b). Again, this enzyme may perform a corrective function, while both may function to recover and recycle ubiquitin from the breakdown products of proteolysis, and in the reversible ubiquitination of other proteins (eg histones, Chapter 1.4).

1.4 Ubiquitin and Histone Conjugation

The first characterisation of a ubiquitin-protein conjugate was reported in 1977 during studies on the chromosomal protein A24. Both Goldknopf and Busch (1977) and Hunt and Dayhoff (1977) reported that A24 was a Y-shaped protein consisting of ubiquitin covalently linked to histone H2A. Subsequent investigations revealed that histone H2B is also ubiquitinated (West and Bonner, 1980), and these proteins are

now known as uH2A and uH2B. In higher eukaryotes, approximately 10% of H2A and 1-2% of H2B are ubiquitinated (eg. West and Bonner, 1980). Conjugation occurs near the C-terminus of the histone, with the C-terminal Gly-76 of ubiquitin linked to the ϵ -amino group of Lys 119 in H2A (Goldknopf and Busch, 1977) and Lys 120 in H2B (Thorne et al, 1987). Ubiquitin-histone conjugation is apparently a general feature of organisms possessing the ubiquitin system, as uH2A and uH2B have also been identified in the slime mould Physarum polycephalum (Mueller et al, 1985) and uH2A has been characterised in the protozoan Tetrahymena pyriformis, where ubiquitin is linked to a Lys residue near the H2A C-terminus (Fusauchi and Iwai, 1985). However, the relative proportions of Physarum uH2A and uH2B (respectively 7 and 6% of H2A and H2B) vary from the higher eukaryote values (Mueller et al, 1985). Ubiquitinated histones and other chromosomal proteins have also been detected in the yeast nucleus (Finley and Varshavsky, 1985).

The role of the ubiquitinated histones is presently unclear. The conjugates appear quite stable and do not result in the degradation of the histone moiety, making it unlikely that they are intermediates in histone breakdown (see Ciechanover et al, 1984). Investigation of nucleosome composition during mitosis has revealed that uH2A disappears from the nucleosomes prior to metaphase and reappears in the G₁ phase, prompting suggestions that mitotic chromosome condensation is in part triggered by the removal of uH2A from metaphase chromosomes (Matsui et al, 1979). Similar events have been observed by Wu et al (1981), and have been studied in detail

in Physarum by Mueller et al (1985). In the latter case, de-ubiquitination occurs during the 7 minute metaphase, while re-ubiquitination takes place in the subsequent 3 minute anaphase, out of a 9 hour nuclear division cycle. In mammalian cells, these events are specific to ubiquitinated histone species, as the conjugation of other ubiquitin substrates is relatively unaffected (Raboy et al, 1986). Whether histone de-ubiquitination actually triggers chromosome condensation or is simply a late obligatory event in this process, is not presently known.

A second possible function of ubiquitinated histones is their involvement in the regulation of gene expression. This possibility stems from the observations that transcribed chromosomal regions in Drosophila appear enriched in ubiquitinated histones, as does the 5'-end of the mouse dihydrofolate reductase gene (Levinger and Varshavsky, 1982; Barsoum and Varshavsky, 1985). However, this may not be a universal phenomenon, or may be restricted to certain genes, as the immunoglobulin κ -chain gene apparently lacks ubiquitinated histones (Huang et al, 1986). These initial observations led to an hypothesis of gene expression regulated by the chromosomal locus-specific ubiquitin-dependent degradation of histones (Levinger and Varshavsky, 1982). However, with regard to the apparent stability of ubiquitinated histones (see above), the hypothesis has been refined to suggest that histone ubiquitination modifies chromatin structure to allow interaction with other factors (Finley and Varshavsky, 1985). While several lines of evidence indicate that substitution of uH2A/uH2B in place of H2A/H2B has little effect on chromatin

structure at the level of the mononucleosome (see Hershko and Ciechanover, 1986), histone ubiquitination may affect interactions between nucleosomes, or between the nucleosomes and other factors.

A feature of the ts85 cell line relevant to histone ubiquitination is the observation that uH2A disappears from the chromatin of ts85 cells with a half-life of ~ 3 hours following the inactivation of the ubiquitin conjugating enzyme (Marunouchi et al, 1980; Matsumoto et al, 1983). This presumably occurs due to the inability to conjugate ubiquitin to histones, while the deubiquitinating isopeptidase(s) is still functional (Chapter 1.2). Notably this apparently complete deubiquitination of histones at the nonpermissive temperature does not trigger chromosome condensation, implying that this may not be an in vivo function of histone-ubiquitin conjugation.

Clearly, further investigations are required to characterise ubiquitin's possible role in marking specific chromosomal regions for recognition by other factors. In this light it is interesting to note that the functions of the yeast RAD6 gene product may be mediated by its ubiquitin-conjugating ability, including its in vitro ability to ubiquitinate H2B (see Chapter 1.2). One such function is in DNA repair, whereby ubiquitination of chromosomal proteins at a damaged site may produce structural change and enable access of the DNA repair system (Jentsch et al, 1987).

1.5 Ubiquitin and the Stress Response

The stress (or heat-shock) response refers to the induction of a specific set of proteins following exposure of cells to a variety of stresses, most commonly elevated temperatures, but also agents such as amino-acid analogues, starvation and heavy metals (for recent reviews see Burdon, 1986; Lindquist, 1986; Schlesinger, 1986; Bond and Schlesinger, 1987). While the exact function of most of these heat-shock proteins (hsps) is not known, current evidence suggests that they protect the cell from damage caused by abnormal proteins resulting from the initial stress event, perhaps by the involvement of hsps in their degradation.

Ubiquitin's known and postulated roles in the stress response stem from several observations. The most direct observation is that ubiquitin itself is a hsp. Ubiquitin genes containing the consensus heat-shock promoter element (HSE) have been identified in both chicken and yeast (Bond and Schlesinger, 1986; Ozkaynak et al, 1987). The HSE is a sequence element which is found in the promoter region of heat-shock genes and mediates their transcriptional activation following stress (reviewed by Bienz and Pelham, 1987). In addition, transcription of this chicken ubiquitin gene and synthesis of ubiquitin itself is elevated following heat-shock (Bond and Schlesinger, 1985) while this yeast ubiquitin gene has been shown to be specific for the stress response (Finley et al, 1987) (see Chapters 1.8 and 1.9). The most straightforward conclusion from these observations is that ubiquitin expression is elevated to cope with an increased turnover of abnormal proteins (generated by the stress conditions)

through the ubiquitin-dependent proteolytic pathway (Bond and Schlesinger, 1985).

Other studies indicate that the induction of the stress response may be linked to the ubiquitin system. This was first noted from studies on the mouse cell line ts85, defective in the ubiquitin activating enzyme. At the non-permissive temperature (39°C), ubiquitin conjugation (and hence ubiquitin-dependent proteolysis) ceases. At the same time, expression of heat shock genes is induced, at a temperature which is well below that required for the stress response in non-mutant cells (Finley et al, 1984). This observed correlation between a lack of ubiquitination and induction of the heat-shock response has led to the proposal that these two systems are closely coupled and may function in a complementary fashion (Finley et al, 1984). These authors speculate that the coupling agent may be a protein factor capable of activating heat-shock genes, which is normally ubiquitinated and degraded via the ubiquitin-dependent proteolytic system. However, overloading of the system by a build-up of heat-denatured proteins, or an inability to conjugate ubiquitin (eg ts85 at the non-permissive temperature) would stabilise the factor and activate the heat shock genes. A similar model proposed by Munro and Pelham (1985) requires that the factor is inactive when conjugated to ubiquitin in equilibrium with the free ubiquitin pool. By similar events as described above, the factor is deubiquitinated and hence activated. These authors further speculate that the factor may be the heat-shock transcription factor itself.

However, recent studies of the levels of free ubiquitin and ubiquitin conjugates in both normal and stressed chicken embryo fibroblasts do not support the hypothesised role of ubiquitin's direct involvement in the induction of the heat shock proteins (Bond et al, 1988). These authors found that while there was a shift in the ubiquitin pool from free to conjugated forms, the size of the shift was small relative to the size of the pool: thus, an ample supply of free ubiquitin is available during the induction of the stress response. An alternative proposal is that an increased level of ubiquitin conjugates may trigger the stress response (Bond et al, 1988).

1.6 Ubiquitin as a Functional Component of Cell Surface

Receptors

Recently another class of apparently stable ubiquitin conjugates has been identified - cell surface molecules. The first such example is the lymphocyte homing receptor, which mediates the specific recognition of, and entry into, the various lymphoid organs by circulating T- and B- lymphocytes (for a recent review see Gallatin et al, 1986). A monoclonal antibody (MEL-14) specific for the lymphocyte homing receptor (which mediates lymphocyte recognition of the peripheral lymph node) was isolated, and was found to block lymphocyte homing both in vitro and in vivo (Gallatin et al, 1983). In an attempt to characterise the receptor molecule, MEL-14 was used to screen a λ gt11 cDNA expression library. Sequence analysis of the three independent cDNA clones thus obtained revealed that all encoded ubiquitin (St. John et al, 1986). These studies also revealed that the MEL-14 antigenic deter-

minant was within ubiquitin's C-terminal 13aa, which were in a conformation different to that in native ubiquitin or in other cellular ubiquitin conjugates. Parallel experiments involving aa sequence analysis of the purified receptor produced sequences from two N-termini, one corresponding to ubiquitin, the other presumably the core polypeptide of the receptor (Siegelman et al, 1986). These authors also reported the immunoprecipitation of other cell surface proteins with monoclonal antiubiquitin antibodies, suggesting that ubiquitination of cell surface molecules may be a more general phenomenon.

These suggestions were confirmed by the discovery that ubiquitin is covalently bound to the core polypeptide of the murine platelet-derived growth factor receptor (Yarden et al, 1986). In this case, ubiquitin was again identified by N-terminal aa sequencing of the purified receptor, revealing two approximately equimolar sequences. However, it was not determined to which domain of the receptor (ie. extra- or intra-cellular) ubiquitin was conjugated.

Another possible example of cell surface molecule ubiquitination comes from reports that anti-ubiquitin antibodies inhibit the sodium-dependent uptake of several compounds in rat brain synaptosomes (Meyer et al, 1986, 1987). These authors suggest that neuronal transporters (or proximal sites) may be ubiquitin conjugates.

Observations of ubiquitinated proteins on the outer cell membrane poses the question of their origin. Three possible pathways exist. First, cytosolic conjugation of ubiquitin to the core polypeptide followed by transfer of the branched

protein through the endoplasmic reticulum (ER) membrane for transport to the outer surface. Second, transfer of an activated ubiquitin-E2 carrier enzyme into the ER and conjugation to the core protein within the ER lumen. Third, transfer of ubiquitin into the ER followed by activation and conjugation to the core protein. The first two possibilities require the transfer of a branched protein through a membrane, while the third requires the presence of the ubiquitin activating and conjugating enzymes within the ER lumen, both of which are unknown entities. However, the identification of the ubiquitin-tail fusion protein (Chapter 1.8) may provide a means for trans-membrane ubiquitin transfer.

A second puzzle concerns the role of cell surface ubiquitination. Presumably ubiquitin is too "ubiquitous" to confer functional specificity on a receptor, although St. John et al (1986) report that MEL-14 recognises an unusual ubiquitin conformation. Possibilities raised include a role in stabilising cell-cell interaction, or even as a pre-formed ubiquitin conjugate to promote rapid degradation once a receptor is internalised and exposed to the ubiquitin-specific proteases (Siegelman et al, 1986). However, it is noteworthy that ubiquitin appears to be at or close to the "active site" of the lymphocyte homing receptor and the neuronal transporters, considering their inactivation by ubiquitin-specific antibodies. Thus ubiquitin may play a significant role in cell surface receptor function.

1.7 Other Ubiquitin Conjugates

Two other examples of stable ubiquitin conjugates have recently been reported. The first case is the identification of ubiquitin as a component of paired helical filaments (PHF), a pathological neuronal fibre specific to the brains of Alzheimers disease patients (Mori et al, 1987). The function of ubiquitin here is presently unclear, although the authors suggest that its presence may reflect a failed attempt at proteolysis of PHF, perhaps due to defects in the ATP/ubiquitin-dependent protease system.

The second example concerns the insect flight muscle myofibrillar protein arthrin, which has been found to be a ubiquitinated form of the more common muscle protein actin (Ball et al, 1987). Again, the function of ubiquitination is unclear, although the authors consider it likely that arthrin plays a specific role in some aspect of the assembly or function of the insect flight muscle thin filaments.

Clearly, as methods become available for examining proteins in fine detail, ubiquitin-conjugation is becoming a more commonly observed event. While the exact role played by such conjugation is generally unclear at present, further studies will no doubt elucidate the functional aspects of this novel post-translational modification.

1.8 Ubiquitin Gene Structure

In the past four years several groups have reported nucleotide sequences of cDNAs and genes encoding ubiquitin from several eukaryotic organisms. These include the entire yeast ubiquitin gene family (Ozkaynak et al, 1984, 1987),

sequences from the slime mould Dictyostelium discoideum (Westphal et al, 1986; Giorda and Ennis, 1987), barley (Gausing and Barkardottir, 1986), Xenopus (Dworkin-Rastl et al, 1984), chicken (Bond and Schlesinger, 1985, 1986; Mezquita et al, 1987), mouse (St. John et al, 1986), pig (Einspanier et al, 1987a), and man (Lund et al, 1985; Wiborg et al, 1985; Baker and Board, 1987a; Einspanier et al, 1987b, Salvesen et al, 1987). Ubiquitin-coding regions have also been sequenced in Drosophila as genomic restriction fragments (Arribas et al, 1986), and thus the actual gene structure remains uncharacterised.

The ubiquitin genes thus identified have revealed a completely novel gene structure with two basic variations. The first structural type is the polyubiquitin gene, which consists of tandemly repeated 228bp (76aa) coding units uninterrupted by spacer peptides. Evidence for polyubiquitin genes exists for all of the abovementioned species. In the four species from which polyubiquitin genes have been sequenced (yeast, slime mould, chicken and man), introns are absent from gene coding regions, while an intron interrupts the 5' non-coding region (NCR) of one chicken and one human, but not the yeast, polyubiquitin gene, the only such genes presently completely characterised with respect to introns. The number of polyubiquitin loci vary between these four species: yeast has a single locus, while the others have at least two. A more variable feature is the number of coding units per locus: yeast has 5 or 6, slime mould 3 and 5, chicken 3 and 4, and man 3 and 9, while cDNA clones indicate that barley, Xenopus and mouse contain one locus of at least

3 coding units, with a 4 or more coding unit locus present in the pig. In addition, Xenopus contains one polyubiquitin locus with at least 12 coding units. The reason for this variability is presently unknown, but is assumed to result from unequal crossover during the evolution of the polyubiquitin gene (Sharp and Li, 1987b). Notably the number of coding units per locus does not appear critical, as the yeast polyubiquitin locus can be functionally replaced by an in vitro constructed single coding unit "minigene" (Finley et al, 1987).

Another feature common to polyubiquitin genes is the encoding of a single aa between the last coding unit and the termination codon. This feature has been noted in all polyubiquitin genes and cDNAs sequenced with the single exception of a Xenopus cDNA, where the stop codon is immediately adjacent to the last coding unit (Dworkin-Rastl et al, 1984). As with coding unit number, this extra aa exhibits both inter- and intra-species variation. Known residues are Asn (yeast), Asn and Leu (slime mould), Lys (barley), Tyr and Asn (chicken), Phe (pig), and Val and Cys (human). The structure of the polyubiquitin gene implies that its expression would produce a polyprotein product, which is then presumably cleaved at the Gly-Met junctions to produce monomeric ubiquitin. Thus, the presence of the extra C-terminal aa would prevent participation of unprocessed polyubiquitin in conjugation, which is most likely its in vivo function. While these processing enzymes have not been characterised, some of the enzymes involved in recycling ubiquitin from ubiquitin-protein

conjugates (Chapter 1.3) are potential candidates for such processing events.

The second ubiquitin gene structural type encodes a fusion protein of one ubiquitin moiety as the N-terminal domain, joined directly to an unrelated protein sequence known as a "tail" protein. These ubiquitin-tail fusions have only been revealed through DNA sequencing, as the tail protein has never been isolated, either individually or as a ubiquitin fusion. Such fusion genes/cDNAs have been reported in yeast, slime mould, mouse and man (Ozkaynak et al, 1987; Westphal et al, 1986; St. John et al, 1986; Salvesen et al, 1987), the latter three as partial cDNA sequences. These reports indicate two fusion protein subtypes varying in the length and sequence of the tail protein. One type has a 52aa tail in yeast, slime mould and man, which is conserved to a high degree between these species. The second type has a 76aa tail in yeast, and 80aa in man, again strongly conserved. The known mouse tail is of the former type, but the partiality of the cDNA prevents analysis of its complete length. These ubiquitin-tail fusion genes are presumably also present in other eukaryotes. For example, a probe specific for the human 80aa tail coding region hybridises to a similarly sized rat mRNA (Lund et al, 1985), while antibodies raised against a synthetic peptide corresponding to a portion of the human 80aa tail protein specifically identify a protein in chicken embryo fibroblasts and a wide variety of cells (see Bond et al, 1988).

The aa sequence composition of the tail proteins indicates several possible functions. First, in each tail type, about 1

in 3 residues are Lys or Arg, suggesting a nuclear function for this basic protein. This possibility is reinforced by the identification of a stretch of basic aa similar to the nuclear translocation signal of the SV40 virus T antigen in each of the tail types: at its N-terminus in the 76/80aa type, and the C-terminus in the 52aa tail protein (Lund et al, 1985; Ozkaynak et al, 1987). While these initial observations hint that the tail protein may function to transport ubiquitin to the nucleus (eg. for histone conjugation), the subsequent identification of a DNA binding domain within each tail type indicates that the tail proteins may function as DNA binding proteins, perhaps in the regulation of gene expression (Ozkaynak et al, 1987). Whether such a function involves the intact fusion protein, or a ubiquitin-free tail protein, is not known. Notably, cleavage of the ubiquitin-tail fusion protein is known to occur, as yeast mutants deleted at the polyubiquitin locus can still produce mature ubiquitin from (at least) one of the ubiquitin-tail fusion genes (Finley et al, 1987).

These known ubiquitin genes and cDNAs thus reveal a novel gene structure, whereby ubiquitin must always be produced by proteolytic cleavage of a fusion protein, consisting either of ubiquitin fused to itself as a polyprotein, or fused to an unrelated, but highly conserved, tail protein.

1.9 Ubiquitin Gene Expression

Most reports of ubiquitin cDNA and genomic sequences have been accompanied by Northern analysis of ubiquitin mRNAs. The general observation from these analyses is that several mRNAs

of different sizes exist in each species. The most complete Northern analysis has been conducted on the yeast S. cerevisiae, which has a ubiquitin gene family of four members: two 52aa tail fusion genes (UBI1 and UBI2), one 76aa tail fusion gene (UBI3) and one polyubiquitin gene (UBI4) (Ozkaynak et al, 1987). Using gene specific probes, expression of all four genes was detected in exponentially growing cells. However, in stationary phase cells, UBI1 and UBI2 expression falls to undetectable levels, UBI3 is expressed but produces a different size distribution of mRNAs, while UBI4 expression increases dramatically. The latter observation is in agreement with the known role of UBI4 in the stress response and the location of a heat shock promoter upstream of the gene (Finley et al, 1987; Ozkaynak et al, 1987).

Several ubiquitin genes have been identified as a result of their specific expression. For example, a chicken ubiquitin cDNA was sequenced during a study of heat-inducible mRNAs, while the subsequently cloned gene was found to contain a heat shock promoter and its transcription was induced ~5-fold upon heat shock (Bond and Schlesinger, 1985, 1986). In addition, ubiquitin was identified in both Xenopus and slime mould as a transcript of a developmentally regulated gene (Dworkin-Rastl et al, 1984; Giorda and Ennis, 1987).

Analysis of Xenopus ubiquitin mRNAs reveals both population polymorphism in mRNA length between individuals and a change in the hybridisation pattern during the development of each individual (Dworkin-Rastl et al, 1984). Similarly, the hybridisation pattern of slime mould ubiquitin mRNAs changes

relative to the developmental stage of the organism (Westphal et al, 1986; Giorda and Ennis, 1987). Drosophila also exhibits population polymorphism with respect to ubiquitin mRNA length (Arribas et al, 1986), while the shorter mRNA species in barley are developmentally regulated, being specifically expressed in dividing cells, possibly to supply extra ubiquitin needed for nuclear histone conjugation (Gausung and Barkardottir, 1986). This speculation by these authors is interesting as it is likely that the smaller mRNAs are transcripts of the ubiquitin-fusion genes, which may function to transport ubiquitin to the nucleus.

Less is known of the expression of ubiquitin genes in higher animal species. In the rat, four different sized ubiquitin mRNAs have been observed (Lund et al, 1985) while three are present in the pig (Wiborg et al, 1985; Einspanier et al, 1987a). A most interesting result is the recent finding that herpes simplex virus infection induces transcription of a hamster ubiquitin gene, with this induction dependent on a viral immediate-early protein ICP4 (Latchman et al, 1987). The exact mechanism and function of this induction is not known.

In man, three differently sized ubiquitin mRNAs can be resolved by Northern analysis. Their sizes have been estimated at 600, 900 and 2400nt, or 650, 1100 and 2500nt, and have been termed UbA, UbB and UbC respectively (Lund et al, 1985). In addition, the human and porcine species are very similar in length (Wiborg et al, 1985). Lund et al (1985) have demonstrated the specific hybridisation of the 80aa tail-fusion cDNA to the UbA species (and to a similarly

sized rat mRNA), while the UbC mRNA is of the correct size to correspond to the human 9 coding unit polyubiquitin gene (Wiborg et al, 1985). However, this latter conclusion has not been confirmed by the use of gene specific probes. This is a common feature of ubiquitin Northern analysis, where, with a few exceptions noted above, coding region probes have been used, which are unable to distinguish products of different ubiquitin genes. Thus the relationship of an organisms ubiquitin gene family to the transcription products of these genes has only been determined in yeast (Ozkaynak et al, 1987).

1.10 Aims

The aim of this study is to characterise the structural organisation of the human ubiquitin gene family, in particular the UbB and UbA₅₂ subfamilies, about which little information is presently available. At the initiation of these studies (March 1985), only three reports of ubiquitin cDNA or genomic sequences had appeared in the literature: a partial Xenopus cDNA (Dworkin-Rastl et al, 1984), a partial sequence of the yeast polyubiquitin gene (Ozkaynak et al, 1984), and the human 9 coding unit polyubiquitin gene uncharacterised with respect to the 5' non-coding region (Wiborg et al, 1985). While these studies had revealed the novel structure of the polyubiquitin gene, there was a clear need for further characterisation of ubiquitin genes to complement and possibly answer questions raised by studies on ubiquitin's functions in histone conjugation and protein degradation. The fundamental importance of the ubiquitin

system to the viability of the eukaryotic organism was by this stage well documented. Conversely, knowledge of the means of providing ubiquitin for this system through the expression of ubiquitin genes was practically non-existent, as no ubiquitin gene had yet been fully sequenced. With these facts in mind, this study was undertaken to improve our knowledge of ubiquitin gene structure, with the ultimate aim of determining how ubiquitin gene expression is regulated to satisfy the functional requirements of the ubiquitin system.

CHAPTER 2

MATERIALS AND METHODS

2.1 MATERIALS

2.1.1 Reagents and Materials

Acrylamide, N,N,N',N' tetramethylethylene diamine (TEMED), agarose (Type II, medium EEO), 5-bromo-4-chloroindolyl- β -D-galactoside (X-Gal), isopropyl- β -D-thiogalactopyranoside (IPTG), oligo(dT)-cellulose, and dextran sulfate were from Sigma Chemical Co., St Louis, MO, USA. N,N'-methylene-bis-acrylamide was from Bio-Rad Laboratories, Richmond, CA, USA. Low gelling temperature agarose was from FMC Bioproducts, Rockland, ME, USA. Lymphoprep density gradient medium was from Nyegaard and Co., Oslo, Norway. Nitrocellulose membranes were from Bio-Rad (Trans-Blot) and Amersham International (Hybond-C), Amersham, UK. GeneScreen Plus was from Du Pont, Boston, MA, USA. Pipes (1,4-Piperazine diethanesulfonic acid), 2'-deoxynucleotide triphosphates and 2',3'-dideoxynucleotide triphosphates were from Boehringer Mannheim GmbH, Mannheim, W. Germany. Oligonucleotide forward and reverse sequencing primers were from New England Biolabs, Beverly, MA, USA and BRESA, Adelaide, S. Australia. Radiochemicals - [α - 32 P]dATP and [α - 32 P]dCTP (3000 mCi/mMole) and [γ - 32 P]ATP (>5000 mCi/mMole) - were from Amersham. X-ray film (Fuji-RX and Fuji-NC II) was from Fuji Photo Film Co., Tokyo, Japan. Black and white instant print film (Land Film Type 57) was from the Polaroid Corporation, Cambridge, MA. USA. Tryptone, yeast extract and agar were from Difco Laboratories, Detroit, MI, USA. All other reagents were of Analytical or A grade.

2.1.2 Enzymes

Restriction endonucleases were from Boehringer Mannheim, New England Biolabs, and Pharmacia, Uppsala, Sweden. The large (Klenow) fragment of Escherichia coli DNA polymerase (EC 2.7.7.7) was from New England Biolabs and BRESA. Alkaline phosphatase (EC 3.1.3.1) and deoxyribonuclease grade II (DNase, EC 3.1.21.1) were from Boehringer Mannheim.

Ribonuclease A type I-As (RNase, EC 3.1.27.5) and Proteinase K (EC 3.4.21.14) were from Sigma. T₄ DNA Ligase (EC 6.5.1.1) was from BRESA, and Polynucleotide kinase (EC 2.7.1.78) was from Pharmacia. S1 nuclease (EC 3.1.30.1) and M-MLV Reverse transcriptase were from BRL, Bethesda, MD, USA.

2.1.3 Media for Bacterial Growth

E. coli strains were grown in the following media:

- 1) Minimal media (per litre): K₂HPO₄, 10.5g; KH₂PO₄, 4.5g; (NH₄)₂SO₄, 1.0g; trisodium citrate, 0.5g. After autoclaving, D-glucose, MgSO₄ and thiamine were added to concentrations of 0.2%, 0.02% and 0.01% (w/v) respectively (Miller, 1972).
 - 2) L-broth (per litre): Tryptone, 10g; yeast extract, 5g; NaCl, 5g (Lennox, 1955).
 - 3) Low salt L-broth: as above, but with 0.6g NaCl per litre.
- Media plates were 1.2% (w/v) agar or agarose. Soft or Top agar/agarose contained 0.6% (w/v) agar/agarose. Where required, ampicillin was added to a final concentration of 50 to 100µg/ml from a filter-sterilized stock solution.

2.1.4 Bacterial Strains, Bacteriophage, Plasmids and Recombinant Libraries

E. coli K12 JM103 (Messing et al, 1981) was used as the host for M13mp8, mp9 (Messing and Viera, 1982), mp18, mp19, pUC18 and pUC19 (Norrander et al, 1983) and their recombinants. E. coli K12 ED8655 (Murray et al, 1977) was the host for EMBL3A (Frischauf et al, 1983) and λ gt10 (Huynh et al, 1984) bacteriophage and their recombinants. Bacteriophage λ gt11 (Young and Davis, 1983a) and its recombinants were propagated in E. coli K12 1090 (Young and Davis, 1983b). The pEX expression plasmids (Stanley and Luzio, 1984) and their recombinants were propagated in E. coli K12 MC1061 (Casadaban and Cohen, 1980) superinfected with pcI857 (Remaut et al, 1983) as the source of temperature-sensitive cI repressor. Human genomic libraries "B" and "D", constructed by ligating partially Sau3A digested human genomic DNA between the BamHI sites of EMBL3A, were gifts of Dr D. Anson, Oxford. A human liver cDNA library constructed in pAT153/PvuII/8 (Giannelli et al, 1983) was the gift of Dr D. Bentley, Oxford (Reid et al, 1984). A human liver cDNA library constructed in λ gt11-amp3 (Kemp et al, 1983) was the gift of Dr G. Howlett, Melbourne. A human placental cDNA library in λ gt11 (Clonetech, Palo Alto, CA, USA) was the gift of Dr R. H. Don, Sydney. A human adrenal gland cDNA library in λ gt10 was from Clonetech.

2.2 METHODS

2.2.1 Routine Laboratory Procedures

Where required, sterilisation was performed by autoclaving at 121°C/100kPa for 15min, or by filtration through 0.2µm membranes. Heat-labile compounds (ie. 2-mercapto ethanol, dithiothreitol, ampicillin, thiamine) or those likely to precipitate during autoclaving (ie. MgSO₄) were added after autoclaving where required. Distilled deionized water was used for all solutions. For manipulations involving RNA, solutions were treated with 0.1% (v/v) diethyl pyrocarbonate overnight prior to autoclaving, and glassware was sterilised in a dry air oven at 180°C overnight. Disposable plastic labware not supplied sterile was autoclaved prior to use. Nucleic acid solutions were generally purified by successive extractions with an equal volume of phenol (saturated with TE buffer: 10mM Tris/Cl; 1mM EDTA, pH8.0) and an equal volume of chloroform/isoamyl alcohol (49:1, v/v). Nucleic acids were concentrated by ethanol precipitation, involving the addition of 1/10th volume of 3M NaOAc pH 5.6 and 2 volumes (DNA) or 2.5 volumes (RNA) of absolute ethanol. Following incubation at 0° or -20°C for 20 to 30 min, precipitates were collected by centrifugation, rinsed with 70% (v/v) ethanol and dried under vacuum. In some instances, a half volume of 7.5M NH₄OAc replaced NaOAc, and 3 volumes of ethanol were used. If volume was limiting, ethanol was replaced by 0.6 volumes (DNA) or an equal volume (RNA) of isopropanol. Nucleic acid concentration was estimated by measuring the absorbance of solutions at 260 nm (1 A₂₆₀ unit = 50µg/ml (DNA) and 40µg/ml (RNA)).

2.2.2 Biological Containment and Radiation Safety

Recombinant DNA procedures were performed according to C1 biological containment conditions stipulated by the Australian Recombinant DNA Monitoring Committee. Radioactive substances were used and disposed of in accordance with the Australian National University Radiation Safety Handbook.

2.2.3 Gel Electrophoresis

Acrylamide/N,N' methylene-bis-acrylamide (19/1, w/w) was prepared as a 40% (w/v) stock solution, deionized over mixed bed resin, filtered, and stored at 4°C. Ammonium persulfate (20%, w/v) was made fresh monthly. Non-denaturing gels (400x 200x0.4mm) were 4 to 20% (w/v) polyacrylamide, 0.06% (w/v) ammonium persulfate, 0.03% (v/v) TEMED, and contained the 89mM Tris base, 89mM Boric acid, 2.5mM EDTA (pH~8.3) electrophoresis buffer (TBE) of Peacock and Dingman (1967). Denaturing (eg DNA sequencing) gels contained 8M Urea, were 6 to 8% (w/v) polyacrylamide, and were generally poured as a "wedge", increasing in thickness from 0.4mm at the top to 1.2mm at the bottom. Gels were pre-electrophoresed for 0.5 to 1h prior to sample loading, and were electrophoresed at 25mA (single thickness) to 33mA (wedge) employing an aluminium heat dispersal plate. DNAs were generally detected by autoradiography. Sequencing gels were fixed in 10% acetic acid/10% methanol (v/v) for 10 to 20 min and dried under vacuum for 40 to 60 min at 80°C prior to autoradiography.

A variety of agarose gels were used for the electrophoresis of nucleic acids. Restricted total human genomic DNA (5µg/lane) and recombinant EMBL3A phage DNAs (1µg/lane) were

electrophoresed in 200x200x5mm 0.8% (w/v) gels in TAE buffer (40mM Tris base; 30mM CH₃COOH; 5mM CH₃COONa; 2mMEDTA; pH~8.0) at 1V/cm for 16h. Rapid analysis of various DNA samples was performed in minigels (60x60x5mm) of 0.6 to 1% (w/v) agarose in TBE buffer at 10V/cm. Restriction mapping gels were 200x200x3mm, 0.4% (w/v) agarose in 40mM Tris base; 30mM CH₃COOH; 20mM CH₃COONa; 2mM EDTA; pH~7.8 (Uher, 1986), and were run at 1V/cm for 42h, fixed in 12% (v/v) CH₃COOH for 10 min, and dried under vacuum at 60°C for 1h prior to autoradiography. Glyoxylated RNAs (Chapter 2.2.11) were electrophoresed in 200x200x5mm, 1% (w/v) gels made in 10mM sodium phosphate buffer (pH6.7) at 4V/cm with constant buffer recirculation. Low gelling temperature gels were 0.6 to 1.4% (w/v) agarose in TAE buffer. Nucleic acids were detected by U.V. irradiation after staining with ethidium bromide (0.5µg/ml of electrophoresis buffer) for 20 min following electrophoresis. For minigels, ethidium bromide was included in both gel and buffer at 0.5µg/ml. Where required, gels were photographed through an orange filter.

2.2.4 Preparation of Plasmid and Bacteriophage DNA

Plasmid DNA was prepared from infected cultures by the alkaline lysis method (Ish-Horowicz and Burke, 1981). Large scale preparations were from 100 to 500 ml cultures in L-broth plus ampicillin (100µg/ml) and were amplified for 16h with chloramphenicol (170µg/ml) upon reaching an A₆₀₀ of 0.8. Small scale preparations (mini-preps) were from 1ml of an overnight culture in L-broth/ampicillin without amplification.

Bacteriophage M13-derived single stranded (ss) DNA was prepared by the method of Sanger et al (1980) with the addition of a CHCl_3 extraction prior to ethanol precipitation. Double stranded replicative form (RF) DNA was prepared on a large scale by a modification of the method of Van Den Hondel and Schoenmakers (1975). Phage culture from a ssDNA preparation was diluted 1 in 250 into an E. coli JM103 culture in L-broth grown to an A_{600} of 1.0, incubated with shaking at 37°C for 4h, and RF DNA isolated from the cell pellet as for a plasmid preparation. RF DNA was purified by CsCl equilibrium gradient centrifugation (Maniatis et al, 1982) or polyethyleneglycol precipitation (Lin et al, 1985). RF DNA was also prepared on a small scale by a mini-prep of the cell pellet remaining after ssDNA preparation.

Bacteriophage λ DNA was prepared on a small scale from plate lysate stocks originating from single phage plaques (Maniatis et al, 1982) by the method of Ozaki and Sharma (1984). Large scale EMBL3A phage preparations were from 400ml liquid lysate cultures inoculated with 8×10^9 host cells pre-incubated with 1.6×10^8 phage pfu as described by Maniatis et al (1982).

2.2.5 Restriction Endonuclease Digestion

Restriction endonuclease (RE) digestions were performed under the conditions given by Farrel et al (1981) employing low, medium and high salt concentration buffers. Double or multiple digestions were generally performed simultaneously by selecting a mutually compatible buffer, or sequentially by adjusting buffer conditions or incubation temperature between RE additions where necessary. RE's were generally diluted at

least 1 in 20 to prevent inhibition or alteration of specificity due to high glycerol concentration. Digestions were terminated by phenol and chloroform extraction, or by the addition of loading buffers containing EDTA, and then heated at 65°C for 10min and quick-chilled prior to electrophoresis to dissociate annealed compatible ends.

2.2.6 Subcloning

The pUC plasmid series and the M13mp bacteriophage series were chosen as vectors for subcloning, DNA manipulation and DNA sequencing because of the versatility offered by the multiple cloning site and the ease of detecting recombinants. A typical preparation of a vector for ligation involved digesting approximately 500ng of vector with the relevant restriction endonuclease(s), removal of 5' phosphate groups with alkaline phosphatase (generally only for singly-digested vector), phenol and chloroform extraction, and ethanol precipitation. Cleaved vectors were resuspended in 10µl TE buffer (10mM Tris/Cl; 1mM EDTA, pH8.0) at a concentration of 40ng/µl (assuming 80% recovery), stored at 4°C and used for up to one year. Target fragments for subcloning were prepared by one of two methods. Wherever possible, the fragment source was cleaved with a combination of RE's so that only the target fragment was compatible with the cleaved vector. The digested source DNA was then phenol and chloroform extracted, ethanol precipitated and ligated with the vector. Where such forced cloning was not possible, target fragments were isolated from low gelling temperature agarose gels as described by Sanger et al (1980). Ligation was performed in 30mM Tris/Cl pH7.5;

10mM MgCl₂; 5mM DTT; 1mM ATP by combining 40ng vector with an approximately equimolar amount of target fragment(s) and 0.3U (cohesive ends) to 1.0U (blunt ends) of T₄ DNA ligase in a final volume of 10µl. Incubation was at room temperature for a minimum of 1 h (cohesive ends) to 6 h/overnight (blunt ends). Ligations were heated at 65°C for 10 min and quick-chilled on ice prior to transformation.

2.2.7 Preparation and Transformation of Competent Cells

Competent cells were prepared by the method of Dagert and Ehrlich (1979), and also by a modification of this procedure described in the "Manual for the M13 Cloning/Sequencing System" (Pharmacia, Uppsala, Sweden). The first resuspension was in 1/5th of the growth volume of 10mM NaOAc pH5.6; 50mM MnCl₂; 5mM NaCl, and the second resuspension was in 1/50th of the growth volume of 10mM NaOAc pH5.6; 70mM CaCl₂; 5mM MnCl₂ and 5% (v/v) glycerol. Competent cells thus prepared were snap-frozen on dry ice and stored at -70°C. Transformation of E. coli strains with plasmid and M13 RF DNA was as described by Yanisch-Perron et al (1985) except that L-broth was used throughout. A typical transformation employed half to all of the ligation mixture described in Chapter 2.2.6.

2.2.8 Identification and Characterisation of Recombinants

Ligations involving the pUC plasmids and M13mp bacteriophage were used to transform an appropriate δ lacpro host (E. coli K12 JM103) and grown in the presence of IPTG and X-GAL. Non-recombinant vectors produce blue colonies/zones of retarded growth ("plaques") by complementation of the host deficiency

to produce a lac⁺ phenotype. Recombinant vectors produce a lac⁻ phenotype due to the insertional inactivation of the β -galactosidase gene, resulting in white colonies/plaques. The efficiency of ligations involving the pEX plasmids was monitored by the increase in the transformation of host cells to ampicillin resistance by the vector-plus-insert ligation compared to the vector-alone control ligation. The presence of an insert in a recombinant plasmid was confirmed by comparing undigested parent and recombinant plasmids by the colony disruption method (Maniatis et al, 1982) and/or restriction analysis of a plasmid mini-prep (Chapter 2.2.4) on agarose mini-gels. Plasmid inserts were characterised by restriction mapping employing a range of enzymes, and where necessary, by Southern blotting and hybridisation of these mini-gels. Restriction fragments of less than 600bp in length were generally characterised by end-labelling (Chapter 2.2.10) and non-denaturing polyacrylamide gel electrophoresis (PAGE, Chapter 2.2.3).

Recombinant M13 phage ssDNA was analysed by direct gel electrophoresis (DIGE) of the phage culture supernatant (Chapter 2.2.4) as described by Messing (1983). Where appropriate, phage containing inserts in opposite orientations were detected by the "C-test" (Messing, 1983).

2.2.9 Construction and Screening of a Human Liver cDNA Expression Library

The cDNA inserts from a human liver cDNA library in pAT153/PvuII/8 (Chapter 2.1.4) were excised by BamHI and PstI digestion and extracted from low gelling temperature agarose

gel slices as described by Sanger et al (1980). Inserts were ligated with BamHI/PstI digested expression plasmid pEX₂ (Chapter 2.1.4) and this mixture was used to transform E. coli K12 MC1061 superinfected with pcI857 as the source of the temperature sensitive cI repressor (Chapter 2.1.4). The resulting colonies were screened by the colony blot procedure of Stanley (1983). The antiserum used to detect ubiquitin antigenic determinants was a gift of 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 as described by Board (1984). Areas of master plates corresponding to positive signals were replated at a lower density and re-screened. Second screen positives were streaked onto ampicillin plates, transferred to a master grid plate and re-screened. Third screen positives were characterised by restriction mapping and sequence analysis.

2.2.10 Radioactive Labelling and Probe Preparation

Restriction fragments were 3'-end-labelled by filling in a restriction site producing a 5' overhang, generally in the same reaction mixture as the digestion. A typical reaction involved the addition of 2.5 μ Ci [α -³²P]dATP, 1 μ l "-A" mix (0.17mM each dGTP, dTTP and dCTP) and 1U Klenow polymerase to the digested DNA (10 μ l) and incubation at room temperature for 15 min. End-labelled fragments were generally analysed by non-denaturing PAGE, employing end-labelled HinfI-digested plasmid pAT153/PvuII/8 as a size standard. Where required (ie. MspI or TaqI restricted DNA), [α -³²P]dCTP and an appropriate "-C" mix was used.

Where performed, 5'-end-labelling was by a Polynucleotide kinase forward reaction essentially as described by Richardson (1965) employing [γ - 32 P]ATP (>5000 Ci/mMole). Uniformly labelled ssDNAs for use as hybridisation probes and S1 protection probes were generated from M13ssDNA subclones of the region of interest by primer extension basically as described by Burke (1984). The 17mer sequencing primer was annealed as for a sequencing reaction and extended by Klenow polymerase in the presence of [α - 32 P]dATP and/or [α - 32 P]dCTP, followed by cleavage with a restriction enzyme cutting 5' to the insert and termination with EDTA. Probes thus prepared were used without further treatment for hybridisation to Southern blots of plasmid and phage digests. For hybridisation to genomic Southern or Northern blots, probe preparations were made 0.15M NaOH, heated at 90°C for 3 min, quick chilled, and electrophoresed in a 1.2% (w/v) low gelling temperature agarose mini-gel. Following autoradiography (15-20 min), probe slices were cut from the gel, boiled for 10 min and added to the hybridisation. Probes for S1 mapping studies were separated from the template DNA by denaturing PAGE, located by autoradiography, and eluted from gel slices by incubating the mashed slice at 55°C for 1 h in 500 μ l of 0.5M NaOAc, 10mM MgCl₂, 1mM EDTA and 0.1% SDS.

2.2.11 Phage Library Screening and Hybridisation

Human cDNA libraries in λ gt10 and λ gt11, and genomic EMBL3A libraries were plated on appropriate hosts (Chapter 2.1.4) on 150mm agar plates at an approximate density of 50,000 pfu per plate. Generally, 6 plates (300,000 pfu) were screened at a

time. Plaque lifts were prepared on nitrocellulose filters as described by Benton and Davis (1977). Nitrocellulose filters were hybridised to a ^{32}P -labelled probe (Chapter 2.2.10) as described by Maniatis et al (1978) except that poly (A) was omitted from the hybridisation solution and the first wash was omitted. Autoradiography was for 1 to 2 days at -70°C with an intensifying screen. Areas of the master plates corresponding to positive hybridisation signals were picked into 500 μl SM buffer (Maniatis et al, 1982) containing 10 μl CHCl_3 /IAA (49:1, v/v) and eluted at 4°C for 6 h/overnight. Suitable dilutions were re-plated at an approximate density of 200pfu per 90mm plate, which were re-screened as above. Positively-hybridising well-isolated plaques were used to prepare plate lysate stocks (Maniatis et al, 1982). Where necessary, a second re-screen was performed as above.

2.2.12 Preparation of DNA and RNA from Human Tissues

High molecular weight genomic DNA was prepared from human blood by the method of Grunebaum et al (1984) and resuspended at an approximate concentration of 1mg/ml. RNA was prepared from term placenta by the method of Chirgwin (1979), while the method of Chomczynski and Sacchi (1987) was used to prepare RNA from cultured transformed human lymphocytes and lymphocytes isolated from freshly drawn human blood by density step gradient centrifugation through Lymphoprep. Poly (A) $^{+}$ RNA was prepared by oligo(dT)-cellulose chromatography as described by Aviv and Leder (1972).

2.2.13 Nucleic Acid Capillary Transfer and Hybridisation

DNA fragments were transferred from agarose gels (Chapter 2.2.3) to nylon membranes (Genescreen Plus) by the Southern Blot technique (Southern, 1975) as modified by Reed and Mann (1985), employing 0.4M NaOH as the transfer solution. Prior to transfer, gels were depurinated in 0.25M HCl for 15 min to aid in the transfer of large fragments. Transfer proceeded from 1 to 2 h (plasmid mini-gels) to overnight (genomic digests). Hybridisation to a ^{32}P -labelled probe (Chapter 2.2.10) was as described by the manufacturer (Du Pont) in 10% (w/v) dextran sulfate; 1M NaCl; and 1% SDS at 65°C. For re-hybridisation, hybridised probe was removed from the membrane as described by Reed and Mann (1985).

RNA samples (Chapter 2.2.12) were treated with glyoxal (Maniatis *et al*, 1982) and electrophoresed (Chapter 2.2.3). RNAs were transferred to Genescreen Plus ("Northern Blot") employing 10mM NaOH as the transfer solution (Dr Ken Reed, personal communication) for 16 h. Hybridisation to a ^{32}P -labelled probe (Chapter 2.2.10) was as described by the manufacturer (Du Pont) in 10% (w/v) dextran sulfate; 50% (v/v) deionised formamide; 1M NaCl; and 1% SDS at 42°C. For rehybridisation, hybridised probe was removed as described by the manufacturer ("Preferred Method").

2.2.14 Restriction Mapping of Recombinant λ Phage Inserts

Restriction maps were determined by a combination of methods. (1) Complete digestion with a range of enzymes singly or in pairs, followed by size fractionation on agarose gels (Maniatis *et al*, 1982). (2) Southern blotting of these gels

(Chapter 2.2.13) and hybridisation with probes for specific gene regions. (3) A modification of the method of Rackwitz et al (1984) as described elsewhere (Baker and Board, 1987a) involving hybridisation of an EMBL3A right arm-specific probe to blotted partial phage digests. (4) A further modification of the method of Rackwitz et al (1984) as described elsewhere (Baker and Board, 1988) involving specific labelling of either the left or the right phage cos end, followed by partial restriction, electrophoresis and autoradiography of the dried gel. In addition, some regions were mapped in more detail as plasmid subclones.

2.2.15 DNA Sequencing and Computer-Assisted DNA Sequence

Analysis

DNA sequencing was accomplished by the "dideoxy" chain termination sequencing technique (Sanger et al, 1977) as applied to the M13 cloning vectors (Messing, 1983). Actual reaction conditions used were from the M13 cloning/sequencing manual produced by Bethesda Research Laboratories, MD, USA. With the combination of cloning sites available in the M13 vector polylinkers (Messing and Vieira, 1982; Norrander et al, 1983), most regions could be sequenced to completion by subcloning various restriction fragments. However, some regions did not contain "convenient" restriction sites, and were sequenced by the method of Lin et al (1985) employing DNase I to generate deletion subclones. Where performed, plasmid sequencing was by the method of Hattori and Sakaki (1986) employing forward and reverse sequencing primers as appropriate. Sequencing reactions of subclones of 200bp or

less were generally loaded on a 0.4/1.2mm "wedge" denaturing gel (Chapter 2.2.3), while those of longer subclones were subjected to double loading on a 0.4/0.8mm wedge gel, re-loaded when the Xylene Cyanole FF (slower) dye of the first loading had travelled the length of the gel. This procedure routinely allowed the determination of 350 to 400 bases per subclone. DNA sequences thus generated were stored and analysed using the programs developed by W. Bottomley, CSIRO Division of Plant Industry, Canberra, Australia.

2.2.16 S1 Nuclease Mapping and Primer Extension Analysis

S1 nuclease protection mapping was modified from Berk and Sharp (1977). Uniformly labelled ss probe fragment (2×10^5 cpm Cerenkov; Chapter 2.2.10) and total cellular RNA (40 to 50 μ g, Chapter 2.2.12) were co-precipitated and resuspended in 15 μ l of 80% (v/v) deionized formamide, 0.4M NaCl, 40mM PIPES pH 6.4, 1mM EDTA. Nucleic acids were denatured at 65°C for 10 min, hybridized overnight at 48°C, diluted to 500 μ l with cold S1 buffer (Maniatis et al, 1982), 300 units of S1 nuclease added, and digested at 37°C for 50 min. Digestion was terminated with EDTA to 15mM, followed by isopropanol precipitation with 2 μ g yeast tRNA carrier, and resuspended in 5 μ l TE buffer.

Primer extension was performed essentially as described by Basler et al (1986) with 10 μ g poly (A)⁺ RNA (Chapter 2.2.12) and 10^4 cpm (Cerenkov) 5'-[³²P]-labelled primer (Chapter 2.2.10), using 80 units of M-MLV reverse transcriptase and including 10 units of RNasin ribonuclease inhibitor in the extension reaction.

The products of both S1 nuclease mapping and primer extension experiments were analysed by denaturing PAGE, employing a sequencing ladder as size markers.

CHAPTER 3

THE HUMAN UBIQUITIN Ubb SUBFAMILY

CHAPTER 3: THE HUMAN UBIQUITIN Ubb SUBFAMILY3.1 Isolation and Sequence Analysis of a Human Liver Ubiquitin cDNA Clone.

A human liver cDNA library was constructed and screened in the expression plasmid pEX₂ (Chapter 2.1.4) as described in Chapter 2.2.9. An initial screen of approximately 40,000 recombinants produced 15 immunopositive colonies. Following two further rounds of screening, the resulting clone with the largest cDNA insert of 720bp was selected and termed pRBL26. The cDNA insert was characterised by restriction analysis, with BglIII, SalI and PvuII digestion producing fragments of approximately 230bp (not shown), the expected size for a 76aa ubiquitin coding unit. In addition, partial restriction revealed that the 230bp SalI fragment was repeated within the cDNA. These restriction enzyme sites and others were utilised in determining the nucleotide sequence of pRBL26 by the strategy shown in Figure 3.1.1. The sequence thus determined is presented in Figure 3.1.2.

Sequence analysis revealed one long open reading frame of 525bp followed by a termination codon TAA, a 3' non-coding region (NCR) of 142bp and a poly(A) tail. A polyadenylation signal AATAAA (Proudfoot and Brownlee, 1976) occurs 17bp upstream of the polyadenylation site. The encoded 175aa polypeptide consists of 7aa of "non-ubiquitin" sequence (cloning artefact - see below) followed by 15aa from the C-terminus of ubiquitin, two direct repeats of the 76aa ubiquitin sequence, and ends with an extra non-ubiquitin cysteine residue (Figure 3.1.2). The encoded ubiquitin proteins are identical to the

Figure 3.1.1: Restriction Map and Sequencing Strategy of the UbB cDNA Clone pRBL26.

Boxes represent the coding region as indicated. Lines represent vector and 3' non-coding sequences and the poly(A) tail [(A)]. Restriction fragments used to generate probes are indicated above the map. Arrows indicate the direction and extent of sequencing.

B:BamHI, E:EcoRI, G:BglII, S:Sal

Figure 3.1.2: Nucleotide Sequence of the pRBL26 cDNA.

The determined sequence is listed with the translation above. Methionine residues at the start of each coding unit are underlined. The extra cysteine codon preceding the stop codon (*) is boxed. The first 23bp are due to a cloning artefact - see page 48. Inverted repeats within each coding unit are underlined with arrows. The first inverted repeat has been extended by 2bp as a result of the cloning artefact (double headed arrows). The AATAAA polyadenylation signal is underlined.

sequenced human protein (Schlesinger and Goldstein, 1975) plus the Gly-Gly C-terminal dipeptide. Thus pRBL26 contains a partial ubiquitin cDNA: presumably a full-length clone would contain at least 3 ubiquitin coding units and include a 5' NCR.

The first 23bp of the cDNA appear to have resulted from a cloning artefact. Sequence analysis reveals the presence of a 10bp inverted repeat spaced by 11bp within each coding unit (Figure 3.1.2). The start of the cDNA is co-linear with the inverted repeats in the first coding unit. A model to explain this artefact is presented in Figure 3.1.3. This model involves the pairing of the inverted repeats in the first coding unit to form a loop ($\Delta G = -8.4$ kcal; Tinoco et al, 1973) following first strand cDNA synthesis. During second strand synthesis, the first strand is nicked, leading to elongation at the 3' end of the nick and opening of the loop. Therefore, the first 23bp of the cDNA are the inverted complement of the actual mRNA sequence (see Figure 3.1.3), and give rise to the "non-ubiquitin" nature of the first 7aa (Figure 3.1.2).

Similar cloning artefacts have been reported previously (eg: Fields and Winter, 1981; Volckaert et al, 1981; Basler et al, 1986).

The pRBL26 cDNA is not of the UbA, ubiquitin-tail fusion type genes, as the reading frame terminates one codon after the last ubiquitin coding unit. In addition, pRBL26 is not the product of the UbC nine coding unit ubiquitin gene (Wiborg et al, 1985), as the 3' NCRs exhibit poor homology, either in length or nucleotide sequence (not shown). Therefore, by exclusion, pRBL26 most likely represents a UbB gene

a) 1st strand cDNA synthesis

3' TGAGAAAGACTGATGTTGTAGGTCTTCCTCAGCTGGGACGT---poly(T) 5'
 (i) (ii)

b) loop formation

TAGT (i)
 G CAGAAAGAGT 3'
 T *****
 T GTCTTCCTCAGCTGGGACGT---poly(T) 5'
 GTAG (ii)

c) chain extension

TAGT (i)
 G CAGAAAGAGTCGACCCTGCA---poly(A) 3'
 T *****
 T GTCTTCCTCAGCTGGGACGT---poly(T) 5'
 GTAG (ii) ↖ nick

d) nicking and unfolding

(ii) TAGT (i)
 5' CGACTCCTTCTGGATGTTG CAGAAAGAGTCGACCCTGCA---poly(A) 3'
 ++++++++
 3' TGGGACGT---poly(T) 5'

e) repair and extension

(ii) (i)
 5' CGACTCCTTCTGGATGTTGTAGTCAGAAAGAGTCGACCCTGCA---poly(A) 3'
 ++++++++
 3' GCTGAGGAAGACCTACAACATCAGTCTTTCTCAGCTGGGACGT---poly(T) 5'

f) derived mRNA (top) versus artefact cDNA (bottom)

5' ACTCTTTCTGACTACAACATCCAGAAGGAGTCCACCCTGCA---poly(A) 3'
 - - - - -
 5' CGACTCCTTCTGGATGTTGTAGTCAGAAAGAGTCGACCCTGCA---poly(A) 3'

Figure 3.1.3: pRBL26 cDNA Sequence Rearrangement Model.

After first strand synthesis (a) the inverted repeats (IRs) (i) and (ii) fold to form a loop (b). The 3' end of the cDNA is elongated by reverse transcriptase (c) and accidental nicking occurs (arrow). (d) and (e): Elongation occurs from the nick and unfolds the loop (shown as separate events). Note that the IRs have been reversed. (f) Sequence of UbB mRNA deduced from the genomic sequence (Chapter 3.6) compared to the artefact cDNA. Asterisks represent IR base-pairing, crosses represent inter-strand base pairing, dashes represent mismatches. The mRNA boxed region (f) is in the inverted orientation in the cDNA (e).

transcript. This conclusion was later confirmed by Northern analysis (Chapter 3.6.5). This cDNA was then used to generate probes to enable the isolation of the corresponding gene.

3.2 Isolation and Characterisation of Human Ubiquitin Genomic Clones

An M13 subclone containing a 228bp SalI fragment spanning the last coding unit of pRBL26 (Figure 3.1.1) was used to generate a coding-region probe. This probe was used to screen some 600,000 phage from two human genomic libraries, "B" and "D" (Chapter 2.1.4), from which 24 positive clones were selected. These clones were then screened with a probe specific for the cDNA 3' NCR. This probe was a 235bp SalI/BamHI fragment containing 37bp of ubiquitin coding sequence, the extra cysteine and stop codons, and the entire 3' NCR and poly(A) tail (Figure 3.1.1). This 3' NCR probe hybridised to 5 of the 24 ubiquitin coding region-positive clones (see Figure 3.2). The 37bp of ubiquitin coding region present in the 3' NCR probe did not cross-hybridise to ubiquitin coding regions under the hybridisation conditions used (Figure 3.2). Clones were named by prefixing their number with EHB or EHD, for EMBL Human B or D library respectively. The five 3'-positive clones, EHB4, EHB6, EHB7, EHB8 and EHD1, contained different genomic fragments based on restriction fragment patterns, although EHB6 and EHB7 shared some fragments (not shown). Hybridisation analysis with coding and 3' NCR probes was used to identify suitable fragments for subcloning and further analysis.

Figure 3.2: Isolation of UbB-related Genomic Clones.
An example of the identification of UbB 3' NCR sequences among ubiquitin genomic clones by dot-blot analysis is shown. Approximately 1 μ g of denatured phage DNA from 12 ubiquitin-positive clones (listed at left) was dotted onto duplicate nylon membranes and hybridised with a ubiquitin coding-region probe (Panel A) or a 3' NCR probe (Panel B). Clones EHB4 and EHD1 were selected for further analysis as described in Chapters 3.3 and 3.4. Clone EHD5 is discussed further in Chapter 4.1.

A

B

EHB 1

EHB 2

EHB 3

EHB 4

EHB 5

EHB 10

EHB 11

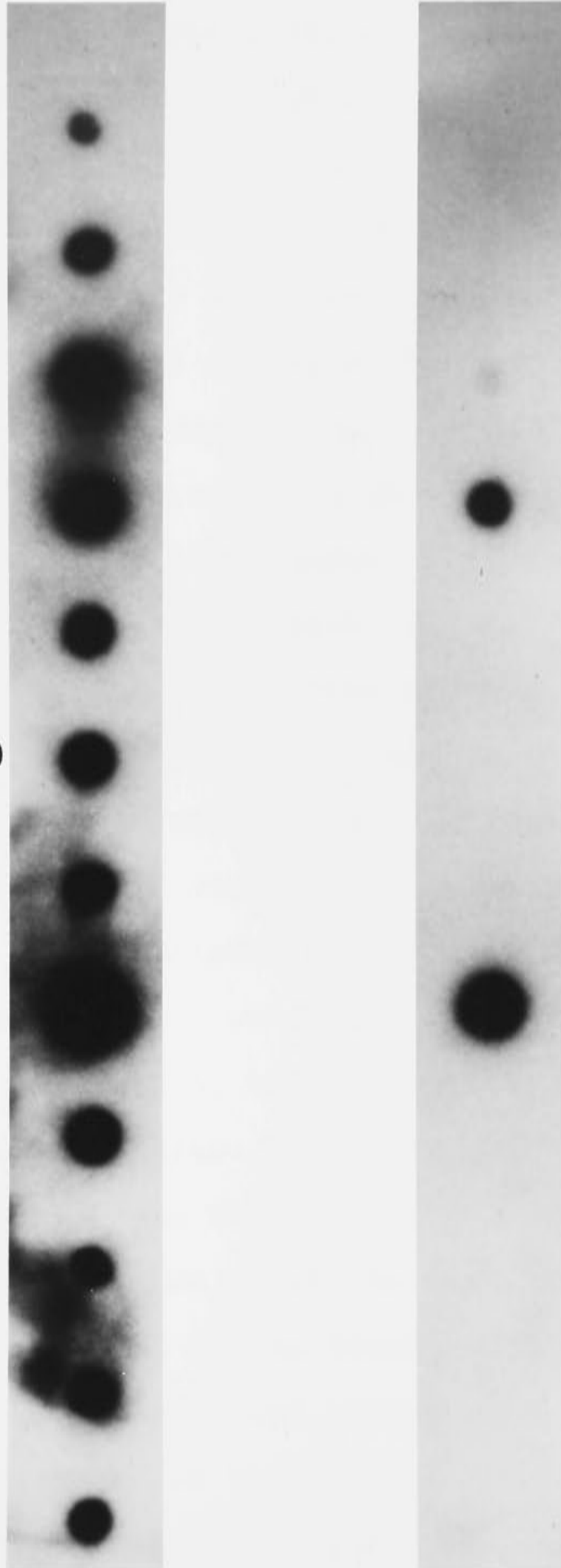
EHD 1

EHD 2

EHD 3

EHD 4

EHD 5



3.3 Sequence Analysis of a UbB Processed Pseudogene EHB4

The restriction map of genomic clone EHB4 is presented in Figure 3.3.1. Hybridisation analysis located all coding and 3' non coding sequences to a 3.2kb PstI fragment, which was subcloned into pUC18 to yield pB4.1 (Figure 3.3.1). The nucleotide sequence of the pRBL26-like region of EHB4 was determined by the strategy shown in Figure 3.3.1 and is presented in Figure 3.3.2. Analysis of the sequence reveals that it is very similar but not identical to pRBL26. The most notable difference is that EHB4 contains only two copies of a ubiquitin-like coding unit, compared to at least 3 for a full length cDNA. The similarities include highly similar coding (92 to 96%) and 3' NCRs (95%) (see Chapter 3.9) and the extra cysteine codon preceding the stop codon (Figure 3.3.2). The flanking regions of EHB4 exhibit features characteristic of processed pseudogenes. First, a poly(A) tail is encoded at the position corresponding to the cDNA polyadenylation site (Figure 3.3.2, nt 814 to 833). Second, the sequence is flanked by the direct repeat AGAAAYAGYTCAGTG (nt 84/98 and 829/843), Y is a pyrimidine), of which the first 5 bases of the 3' repeat overlap the poly(A) tail. The coding region also exhibits pseudogene features. The first coding unit contains an in-frame stop codon TAG at the 65th codon. In addition, there are 13 other DNA mutations leading to codon changes: 10 in the first and 3 in the second coding unit (Figures 3.3.2 and 3.9.1). The second coding unit is also interrupted by an 11bp insertion between the 38th and 39th codons (Figure 3.3.2, nt 541 to 551). The lack of a cDNA 5' NCR prevents analysis of the 5' NCR of EHB4. However, the

Figure 3.3.1: Restriction Map and Sequencing Strategy of Genomic Clone EHB4.

Top line: Restriction map of EHB4 SalI insert.

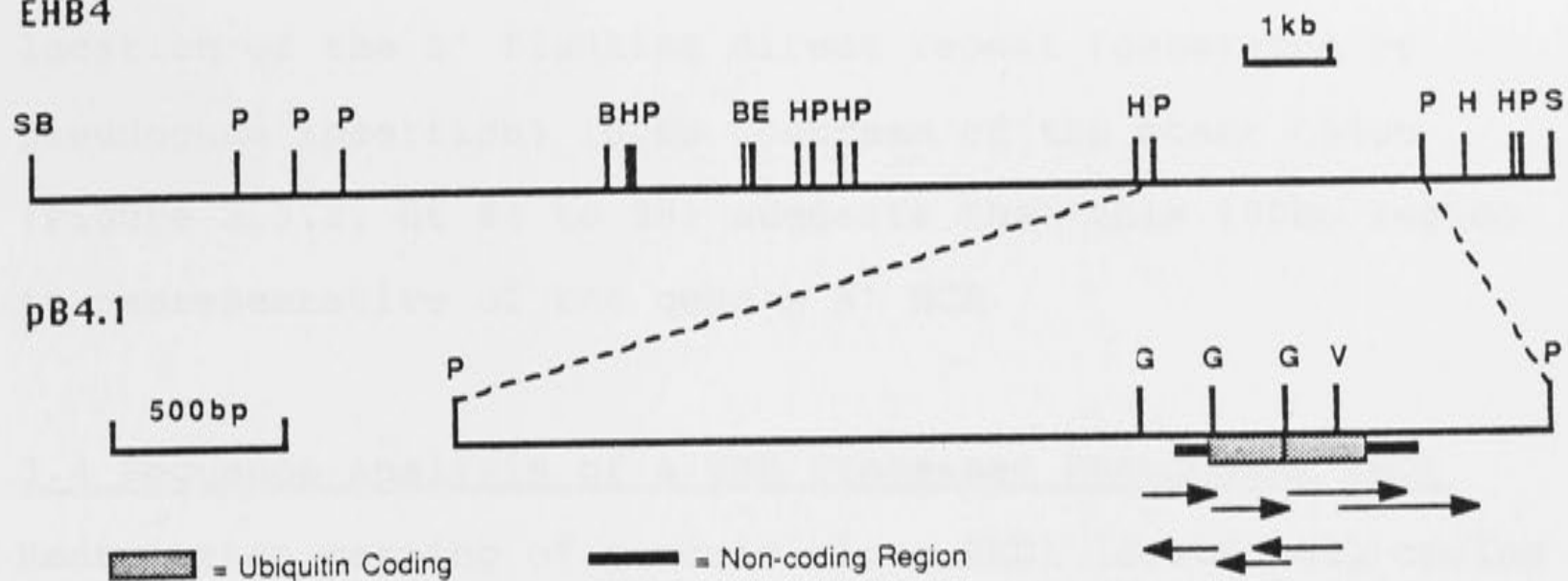
Lower line: 3.2kb PstI subclone pB4.1. Boxes represent coding and non-coding regions as indicated. Arrows indicate the direction and extent of sequencing.

B:BamHI, E:EcoRI, G:BglII, H:HindIII, P:PstI, S:SalI, V:PvuII.

Figure 3.3.2: Nucleotide Sequence of a UbB Processed Pseudogene EHB4.

The determined sequence is listed with the translation of the ubiquitin-like region given above. Encoded residues differing from ubiquitin are overlined with a bar. An 11bp insert within the second coding unit and the AATAAA polyadenylation signal are underlined. The encoded poly(A) tail-like region is indicated by a dashed underline, while the position corresponding to the cDNA polyadenylation site is marked by an arrowhead. Flanking direct repeats are overlined with arrows. Asterisks are stop codons.

EHB4



AGATCTTGTATCTAATCATGTCTCTAACTTTGCATAAAGACAGGGATTTGTCCCAGCTTAAGGGCCAAGG
 10 20 30 40 50 60 70

GAAGAACTCGCATAGAAATAGTTCAGTGGTAGGGACGACTGGCTTTGTTGGGTGAGTATGTGGTGGTGTGTC
 80 90 100 110 120 130 140

CCTGTGGCTGGACATGACTGGTGTTCGGAGCCTCGTCATATCCGCTAACAGGTCAAAATGCAGATCTTC
 150 160 170 180 190 200 210

V K T L T S̄ K T I Ā L E Ē E P R̄ D T I E N V K
 GTGAAGACCCTTACCAGCAAGACCATTGCCCTTGAAGAGGAGCCCAGGGACACCATCGAAAATGTGAAGG
 220 230 240 250 260 270 280

A K I Q D K E G I T̄ P D Q Q R L I F A V̄ K Q V̄ E
 CCAAGATCCAGGATAAGGAAGGCATTACCCCGGACCAGCAGAGGCTCATCTTTGCAGTCAAGCAGGTGGA
 290 300 310 320 330 340 350

D D̄ R T H̄ S D Y N I Q K E T̄ T L H L V L S̄ L R
 AGATGACCGCACTCATTCTGACTATAACATCCAGAAAGAGTAGACCCTGCACCTGGTCCTGAGTCTGAGA
 360 370 380 390 400 410 420

G G MET Q I F V K T L T G K T I T L E V E P S D
 GGTGGTATGCAGATCTTCGTGAAGACCCTGACTGGCAAGACCATCACCCCTGGAAGTGGAGCCCAGTGACA
 430 440 450 460 470 480 490

T MET E N V K A K I Q D K E G T̄ P P D Q Q
 CCATGGAAAATGTGAAGGCCAAGATCCAGGATAAAGAAGGCACCCCCCAGCCCCGACCAGCAG
 500 510 520 530 540 550 560

R L I F A G K Q L E D G R T L S D Y N I Q K E
 AGGCTCATCTTTGCAGGCAAGCAGCTGGAAGATGGCCGCACTCTTTCTGACTACAACATCCAGAAAGAGT
 570 580 590 600 610 620 630

S T L H L V L H̄ L R G G C *
 CAACCCTGCACCTGGTCCTGCACCTGAGGGGTGGCTGTTAATTCTCCAGTCTTGCATTTCGCAGTGCCCAG
 640 650 660 670 680 690 700

TGATGGCATTACTCTGCCGTATAGCCATTTGCCCAACTTAAGTTTAGAAATTACAAGTTTCAGTAATAG
 710 720 730 740 750 760 770

CTGAACCTGTTCAAATGCTAATAAAGTTTTGTTGCATGGTTAAAAAAGAAAAGGAAAGAAACAGCTCA
 780 790 800 810 820 830 840

GTGTTGCAGGAAAGGCCTGGATCATTATTTCAAAAACCTCAATTTCTCAGACTGCTTGTCTTGAATAAAC
 850 860 870 880 890 900 910

location of the 5' flanking direct repeat (generated by pseudogene insertion) 100bp upstream of the start codon (Figure 3.3.2, nt 84 to 98) suggests that this 100bp region is representative of the gene's 5' NCR.

3.4 Sequence Analysis of a UbB Processed Pseudogene EHD1

Restriction mapping of genomic clone EHD1 located all coding and 3' non-coding hybridisation on a 1.8kb EcoRI/HindIII fragment, which was subcloned into pUC18 to yield pD1.1 (Figure 3.4.1). The nucleotide sequence of the pRBL26-like region of EHD1 was determined by the strategy shown in Figure 3.4.1 and is presented in Figure 3.4.2. Sequence analysis reveals that while it is similar to the cDNA clone and to EHB4, it shows marked differences to both. Most notably, EHD1 contains only one ubiquitin-like coding unit. As with EHB4, EHD1 contains the extra C-terminal cysteine codon, and a 3' NCR that is respectively 94% and 93% similar to those of pRBL26 and EHB4 (see Chapter 3.9). EHD1 also appears to be a processed pseudogene. A short poly(A) tail is encoded 6bp downstream of the position corresponding to the cDNA polyadenylation site (Figure 3.4.2, 569 to 576). In addition, the sequence is flanked by the direct repeat ATTCTGGA (83/90 and 576/583). The single coding unit is 93% similar to those of the cDNA, while its translation differs by 8 residues from ubiquitin (Figure 3.4.2). Unlike EHB4, no nonsense codons or insertions disrupt the reading frame. The EHD1 coding unit terminates with an amber codon (TAG), compared to the TAA ochre codons of pRBL26 and EHB4. Most interestingly, the 98bp upstream of the ubiquitin-like initiation codons of the two

Figure 3.4.1: Restriction Map and Sequencing Strategy of Genomic Clone EHD1.

Top line: Restriction map of EHD1 SalI insert.

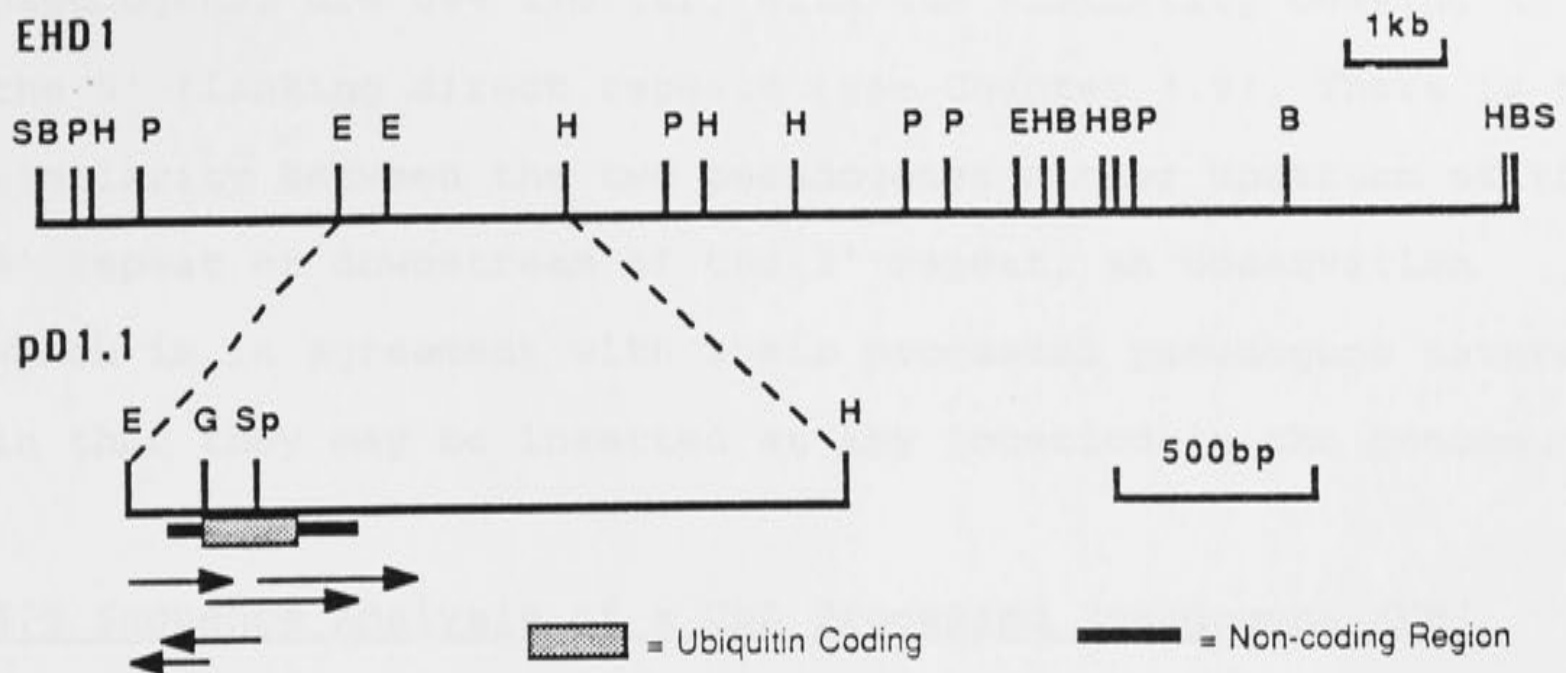
Lower line: 1.8kb EcoRI/HindIII subclone pD1.1. Boxes represent coding and non-coding regions as indicated. Arrows indicate the direction and extent of sequencing.

B:BamHI, E:EcoRI, G:BglII, H:HindIII, P:PstI, S:SalI, Sp:SphI.

Figure 3.4.2: Nucleotide Sequence of a UbB Processed Pseudogene EHD1.

The determined sequence is listed with the translation of the ubiquitin-like region given above. See Figure 3.3.2 legend for symbols.

EHD1



GAATTCAGGCTTTTGTGTTTATATCCTACATCTACCACATTCAAATATACATAAAAATAA
 10 20 30 40 50 60
 AACATATAGATAGAATAAAAATATTCTGGAGGGGTGAATGGCTTTGTCGGGTGAGCGTGT
 70 80 90 100 110 120
 GGAGGTGTTCTGTGGCTGGACATGACTGGCGATTTCGCAGCATCCTCGTATCTGCTAACA
 130 140 150 160 170 180
 MET Q I F V K T L T G K T I T L E V E
 GGTCAAAATGCAGATCTTCGTGAAGACCCTGACTGGCAAGACCATCACCCCTTGAAGTGGA
 190 200 210 220 230 240
 P S D T I E N V K A N I Q D K E G I L P
 GCCCAGTGACACCATCGAAAATGTGAAGGCCAATATCCAGGATAAGGAAGGCATCCTCCC
 250 260 270 280 290 300
 D Q Q R L I F A G MET Q L E D G C T L S D
 CGACCAGCAGAGGCTCATCTTTGCAGGCATGCAGCTAGAAGATGGCTGTA CTCTTTCTGA
 310 320 330 340 350 360
 Y N I Q K E L T L Y L V Q R L R C G C
 CTACAACATCCAGAAAGAGTTGACCCTGTACCTGGTCCAGCGTCTGAGATGTGGCTGTTA
 370 380 390 400 410 420
 GTTCTTCAGTCTTGCATTTGCAGTGCTCAGTGATGGCATTACACTGCACTATAGCCATCT
 430 440 450 460 470 480
 GCCCCCACTTAAGTTTAGAAATTACAAGTTTCAGTAATAGTTGAACCTGTTCAAATGTT
 490 500 510 520 530 540
 AATAAAGGTTTTGTTGCATGGTAGCATTAAAGAAAATTCTGGATGCCATACTTTTGGAAA
 550 560 570 580 590 600
 AACATTATTTCAAAGTCCTTCTATAAAAGTTGTACACGACTGTACTTTG
 610 620 630 640 650

pseudogenes are 88% similar, with the similarity ceasing at the 5' flanking direct repeats (see Chapter 3.9). There is no similarity between the two pseudogenes either upstream of the 5' repeat or downstream of the 3' repeat, an observation which is in agreement with their processed pseudogene nature, in that they may be inserted at any location in the genome.

3.5 Sequence Analysis of a UbB Processed Pseudogene EHB7

Restriction mapping of genomic clones EHB6 and EHB7 indicated that they contained overlapping genomic fragments, with all ubiquitin coding and 3' non-coding hybridisation present in the area of overlap (Figure 3.5.1). A 3.5kb BamHI fragment from EHB6 was subcloned in pUC18 to yield pB6.1. Similarly, pB7.2 contains a 5.7kb HindIII fragment from EHB7 (Figure 3.5.1). One of the BamHI sites used to produce the EHB6 3.5kb BamHI fragment is a cloning artefact resulting from the fusion of the genomic insert Sau3A site with the EMBL3A phage right arm BamHI site. For this reason, EHB7 was chosen as the clone best representing the in vivo genomic arrangement, although much of the sequence was determined on EHB6. Figure 3.5.2 shows the nucleotide sequence of the EHB6/EHB7 pRBL26-like region determined by the strategy shown in Figure 3.5.1. Sequence analysis reveals that EHB7 is analogous to EHD1 in that it is a single coding unit UbB processed pseudogene. However, EHB7 is present at a different genomic location than is EHD1, based on differences in their genomic maps (Figures 3.4.1 and 3.5.1) and different nucleotide sequences upstream and downstream of the pseudogenes (Figures 3.4.2 and 3.5.2). EHB7 also appears to have arisen from reverse translation of

Figure 3.5.1: Restriction Maps and Sequencing Strategies of EHB6 and EHB7.

Restriction maps of genomic clones EHB6 and EHB7 and their subclones pB6.1 and pB7.1 are shown. Boxes represent coding and non-coding regions and Alu repeats as indicated. Arrows indicate the direction and extent of sequencing. Inserts are aligned to show their overlap. The Sau3A site (Su) responsible for the BamHI site at the end of the EHB6/pB6.1 insert is shown: other Su sites are not shown. B:BamHI, E:EcoRI, H:HindIII, P:PstI, S:SalI, Su:Sau3A, T:TagI. Only the T site used for sequencing is shown.

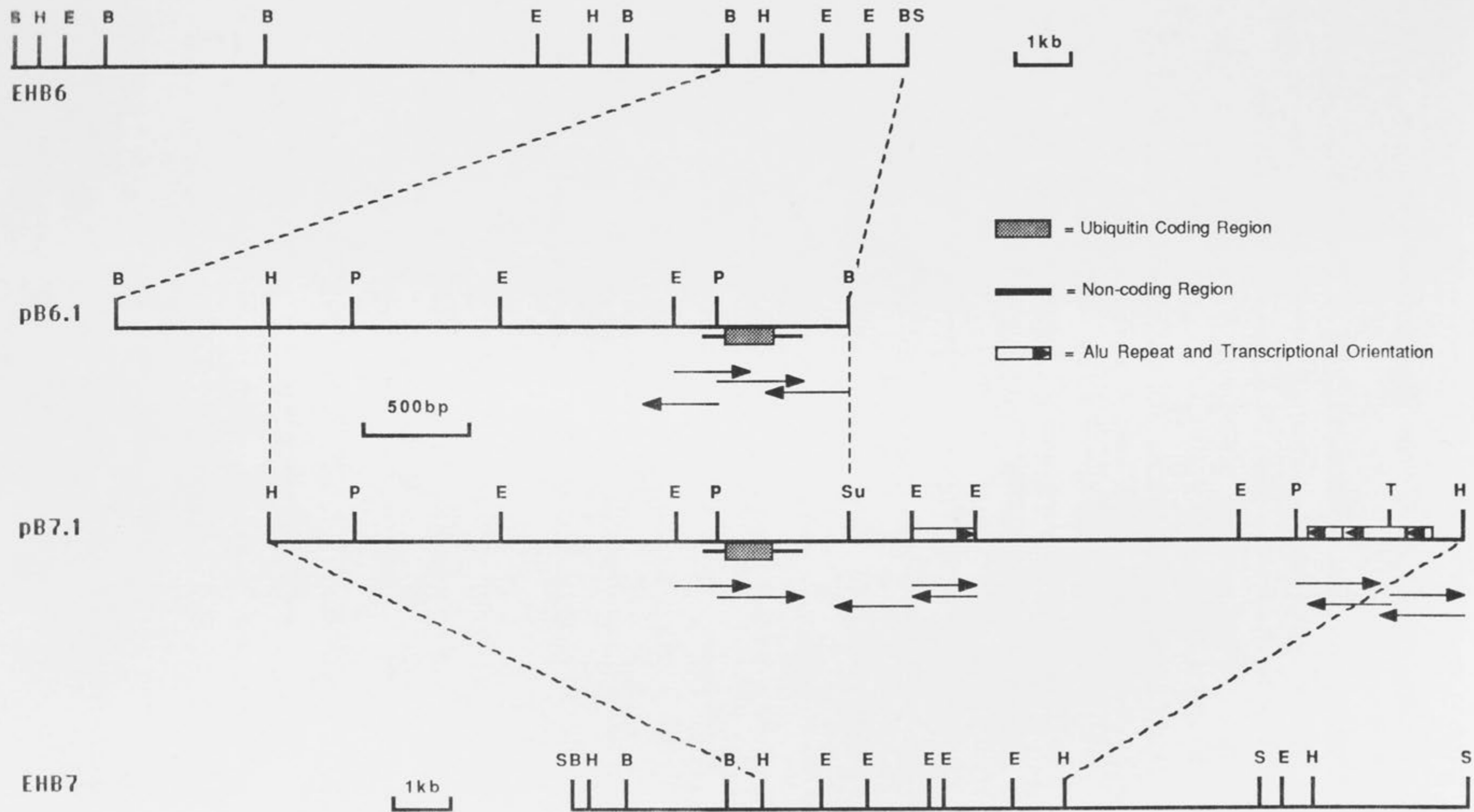


Figure 3.5.2: Nucleotide Sequence of a UbB Processed Pseudogene EHB7.

The determined sequence is listed with the translation of the ubiquitin-like region given above. The sequence is presented in two portions spaced by an ~1.5kb unsequenced region. See Figure 3.3.2 legend for symbols. Additional symbols are: Alu poly(A) tails (dashed underline) Alu flanking direct repeats (overlined with arrows), and a BamHI semisite (overlined) at which partial Sau3A cleavage occurred during genomic library construction to generate the BamHI site at the end of the genomic clone EHB6 insert (see Figure 3.5.1).

GAATTCAGTTCCTCACTGGGGGTTGGCTAGAAAGCTGCCCTCTGTTCTTCATCATGTGGGCTCACACAGCTGTAGGCATC
 10 20 30 40 50 60 70 80
 ATGAGGAACAAGTAAGAAAGAGGGAACAGAGTGGCGAATGGCCTTGTGTAATGGCGAGCGTCTGGAGGCATTCCAGTGGC
 90 100 110 120 130 140 150 160
 TGAATGTGATTGGTGTATCTGCAGCTTCTCGCATCCTCCAAGAGGTCAAAATGCAAATCTTCTGAAAACCTGACCGGC
 170 180 190 200 210 220 230 240
 MET Q I F L K T L T G
 K T I T L E V E P S D I L Q N V K A K I H V K E G I P
 AAGACCATCACCTGGAGGTGGAGCCAAGTGACATCCTCCAAAATGTGAAGGCCAAGATCCATGTTAAAGAGGGCATCCC
 250 260 270 280 290 300 310 320
 P D Q H S L I F V G K Q L E D G C T V C D Y N I Q K
 CCCTGACCAGCACAGTCTCATCTTTGTAGGCAAGCAGTTAGAAGATGGCTGCACTGTTTGTGACTACAACATTTCAGAAAG
 330 340 350 360 370 380 390 400
 E S A L H L V L H L R G G Y *
 AGTCAGCCCTGCACCTGGTCTCCATCTGAGGGGTGGCTATTAATTCTTCAGTCTTGCATTTCGTAGTGACAAGTGATGGC
 410 420 430 440 450 460 470 480
 ACTACTCTGCACTAAAGCCATTTGCCCAATTTAAGTTTATAAATTACCAGTTTCGGTAATAGCTGAACCTGCTCAAAAT
 490 500 510 520 530 540 550 560
 GTTAATAAGCGTTTTTGTTCATGGTTAAAAAAAAAAAAAGGGAATAAGACAGAAACCACAGTTCTTCATAGCTTAATCTC
 570 580 590 600 610 620 630 640
 ATCCATCATTTTTGCCATATTCCATTGATTAGAAGCAAGTCACTGGTCTAGCACACATTCAGAGGGAAATGGTCACACAA
 650 660 670 680 690 700 710 720
 GGATGTAATAATAAGAGTCATGTATCATCCAGGGCATCTTAAAAAGCTGCCTGCCACAGCAGCCAATAAGTTAATTAGT
 730 740 750 760 770 780 790 800
 TGTTCTTCTGGATCACTTACACTGAAAATAGAGCAGAGAGTAGAGTCAAAGGATACTTTGGAAACAAAATAAACCC
 810 820 830 840 850 860 870 880
 TTCCCTCAAAAAGCCTCAGTCTAGCTGGAAGATAAGCTGTAGCAAATTAAGAGAAAATAACAGCCATTCTAAGCAAAA
 890 900 910 920 930 940 950 960
 ACCCAAGATAGCAACCATCAAAGGAAAAAGCAATATATTACCCCTCCTAAAAATATAAAGTTCTATACATCAAAAACCA
 970 980 990 1000 1010 1020 1030 1040
 TACACAATAGTAAAATAAGACAATCAGTTAAGGATGGCTTCTGCAACATATGTGACAAAGAGTTGATATCTTTAATAAAG
 1050 1060 1070 1080 1090 1100 1110 1120
 AATTCTCGGCCAGGCCAGTGGCTCACATCTGTAATCCCGGCACTTTGGGAGGCCAAGGCAGGTGGATCACGAAGTCAGG
 1130 1140 1150 1160 1170 1180 1190 1200
 AGATCAAGATCATCCTGGCCAACATGTGAAACCCCGCCTCTACTAAAAATACAATTAGCTGGGCTTGGTGGCACATGCCT
 1210 1220 1230 1240 1250 1260 1270 1280
 GTAATCCCAGCTACTCCGGAGGCTGAGTCAGGAGAATCGCTTGAACCCAGGAGCGGAGCTTGCAGTGAGCTGAGATCAT
 1290 1300 1310 1320 1330 1340 1350 1360
 GCCACTCCACTCCAGCCTGGTGACAGAGAGAGACTCTGTCTCAAAAAAAAAAAAAAAAAAAGATTTC ~1.5kbv
 1370 1380 1390 1400 1410 1420
 CTGCAGATGGCTTGGAACTCTGAACATGACTTTATTTCCCTTAAATGACCCATTCCCTTATTTATTTATTTATTTATTT
 10 20 30 40 50 60 70 80
 ATTTATTTATTTATTTATTTATTTGAGACAAGGTCTCATCGCCAGGGTGGAGTGCAGTGGCACAATCACAGCTCATTGTC
 90 100 110 120 130 140 150 160
 CCGACTCATCCTCCCAGGTAGCTGGGACTACAGGCATGTGCATCCATGCCTTGCTAATTTTTTGTATCTTTTTTTTTTTT
 170 180 190 200 210 220 230 240
 TTTTGGAGACAGAGTCTCGCTCTGCTCTAGGCTGGAGTGCAGTGGCGGATCTCGGCTCACTGCAAGTCTGCCTCCCA
 250 260 270 280 290 300 310 320
 GGTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGCGCCCGCCACCAGGCCCGGCTAATTTTTTGGTATTTTTAG
 330 340 350 360 370 380 390 400
 TAGAGATGGGGTTTACCACGTTAGCCAGTATGGTCTCGAACTTCTGACCTAGTGATCTGCCACCTCAGCCTCCTAAG
 410 420 430 440 450 460 470 480
 TGCTGGGATTACAGGCATAAGCCACCAGCCAGCCTAATTTTTTGTATGTTTAGTAGAGACGGGGTTTACCATGTGCG
 490 500 510 520 530 540 550 560
 CCAGGCTGGTCTTGAACCTCTGGCCTCAAGTGATCTGCCTCCCAAAGTGCTGGGATTACAGGCATAAGCCTCTGTGCCCA
 570 580 590 600 610 620 630 640
 GCCTGACCCACTCCCTTTTAAAGAGGAGAGGGAGTAATTGAATTACCAATGAATTAATTGGATAAGACACTCTAACTCAGA
 650 660 670 680 690 700 710 720
 AAAGAGATCAATTGTTATAAATAATACTTAAAAAGCCCAATCCTTTTCATAAATAGTGAAGCTT
 730 740 750 760 770 780

a different mRNA than EHD1, as the former is polyadenylated at the cDNA position (Figure 3.5.2, nt 587), while the latter is polyadenylated 6bp 3' to this position (Figure 3.4.2, nt 569). In addition to the encoded poly(A) tail (Figure 3.5.2, nt 587 to 569), EHB7 also exhibits processed pseudogene direct repeats (nt 67/76 and 565/574). The coding region and 3' NCR are respectively 87/89% and 89% similar to the cDNA regions (see Chapter 3.9). The coding unit encodes an open reading frame differing from ubiquitin at 14 residues due to DNA mutations, while the extra C-terminal cysteine codon TGT has mutated to a tyrosine codon TAT (Figure 3.5.2). Again, the 101bp upstream of the EHB7 initiation codon is respectively 70% and 77% similar to those of EHB4 and EHD1 (see Chapter 3.9), with similarity ceasing at the direct repeat, supporting the likelihood that this region represents the gene 5' NCR.

Sequence analysis also revealed the presence of 3 members of the Alu family of repetitive DNA sequences close to the EHB7 processed pseudogene. One Alu repeat occurs 524bp downstream (Figure 3.5.2, nt 1128 to 1421) while a further 1.5kb downstream one complete and two half Alu repeats are present (Figures 3.5.1 and 3.5.2). The Alu repeats are described in detail in Chapter 7.

3.6 Sequence Analysis of a Human UbB Gene EHB8

3.6.1 Characterisation of Coding and 3' Non-coding Regions

The restriction map of genomic clone EHB8 is presented in Figure 3.6.1. Hybridisation analysis located all coding and 3' non-coding sequences to a 4.1kb PstI fragment, which was

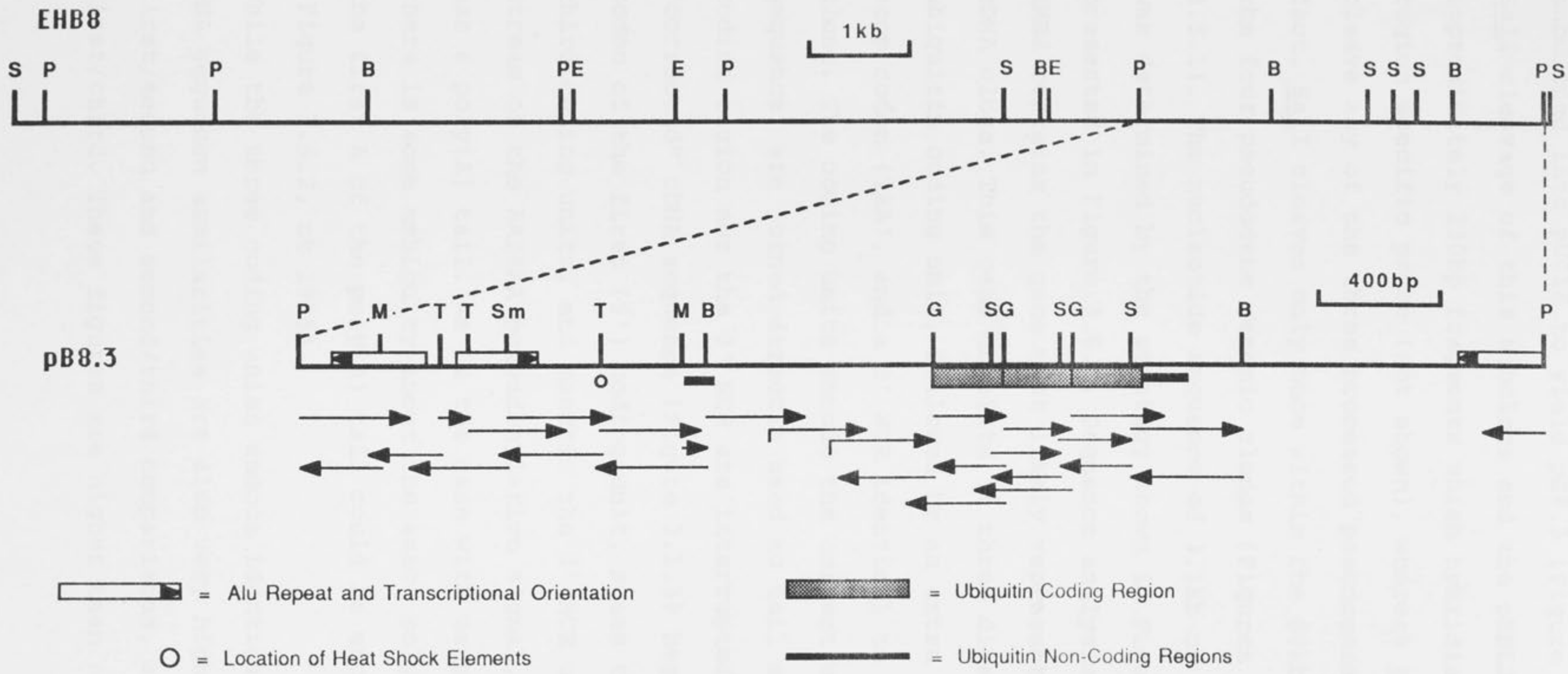


Figure 3.6.1: Restriction Map and Sequencing Strategy of Genomic Clone EHD8.

Top Line: Restriction map of EHB8 Sall insert.
 Lower line: Restriction map and sequencing strategy of the 4.1kb PstI subclone pB8.3. Boxes represent coding and non-coding regions and Alu repeats as indicated. An open circle represents the location of the heat shock elements. Arrows indicate the direction and extent of sequencing. Arrows originating from a vertical bar represent DNaseI-generated subclones. B:BamHI, E:EcoRI, G:BglII, M:Mspl, P:PstI, S:Sall, Sm:Smal, T:Tagl. Only those M, Sm and T sites used for sequencing are shown. HindIII does not cleave within the EHB8 insert.

subcloned into pUC18 to yield pB8.3 (Figure 3.6.1). Notably, SalI cleavage of this subclone and the pRBL26 cDNA produced approximately 230bp fragments which hybridised to the coding region specific probe (not shown), whereas SalI did not cleave any of the three processed pseudogene subclones. In fact, SalI cleaves only once within the 60kb represented by the four pseudogene genomic clones (Figures 3.3.1, 3.4.1 and 3.5.1). The nucleotide sequence of 3.1kb of the pB8.3 insert was determined by the strategy shown in Figure 3.6.1 and is presented in Figure 3.6.2. Sequence analysis reveals that EHB8 contains the gene most likely represented by the pRBL26 cDNA clone. This gene consists of three direct repeats of the ubiquitin coding unit, followed by an extra cysteine codon, stop codon (TAA), and a 3' NCR identical to that of the cDNA clone. The coding units encode the correct ubiquitin protein sequence, are joined directly head to tail and neither the coding region nor the 3' NCR are interrupted by introns. The "corrected" cDNA sequence (Figure 3.1.3) begins with the 55th codon of the first (5') coding unit, spans the second and third coding units, and matches the 3' NCR until 17bp downstream of the AATAAA polyadenylation signal, where the cDNA has a poly(A) tail. As is the case 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 (Figure 3.6.2, nt 2915).

While the three coding units encode identical proteins, their DNA sequence similarities are also very high: 96.5% for first/second and second/third comparisons, and 94.7% for first/third. These figures are higher than necessary, given

Figure 3.6.2: (Following Pages) Nucleotide Sequence of a Human Ubiquitin UbB Gene.

The sequence determined by the strategy outlined in Figure 3.6.1 is given with the translation shown above. The TATA box is indicated by a heavy underline. The most upstream mRNA start site is indicated by an arrowhead. The positions of the splice acceptor and donor sites of the 5' NCR intron are indicated "intron". The AATAAA polyadenylation signal is underlined and the pRBL26 cDNA polyadenylation site is shown by an arrowhead. The variant polyadenylation site of pseudogene EHD1 is shown in brackets. Alu repeats upstream of the gene are indicated and numbered, with their poly(A) tails shown by a dashed underline. Alu flanking direct repeats (where present) and a direct repeat in the 3' NCR are overlined with arrows. Sequences resembling the consensus heat shock element are numbered HSE 1, 2 and 3.

CTGCAGTGAACGGTGATCACACCACTGCACACCAGCCTGGGGACACAGCCAGACTTTGTCACAAAAAAGC
10 20 30 40 50 60 70
AAAAACAACCTGGCCAGTGTATGAGGGGCTCGTGTTTTTTGGTTTGTCTGTTTGGTTGAGACAGAGTCTCACT
80 90 100 110 120 130 140
CTGTCCGACTGGAATGCAGTGGCACATTCTCGGCCACTGCAATCTCTGCCTCCTAGGTTCAAGCAA
150 160 170 180 190 200 210
TTATCTGCCTCAGCCTCCCAAGTAGCTGGGATTACAGGCGCCCGCACCACGCCCGGCTAATTTTTTTGTA
220 230 240 250 260 270 280
TTTTTAGTAGAGACGGGGTTTCACCACCTTGGCCAGGCTGGTCTTGAACCCCTGACCTCATGATCCACCC
290 300 310 320 330 340 350
GCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCTCCCGCCCGGCCAGGGGCGCGCGTTTTTAA
360 370 380 390 400 410 420
AACATGGGAGAGGGAATTGTGCTTCACAATCACCATCAGGTGTCTCGATATCGGGTGCCACGCCGTCCCG
430 440 450 460 470 480 490
CTTCTGAGGCGCGGCGGCCACTTTGGCAGGCCGAGGCGGGTGGATTACCTGAGGTCAGGAGTTCGAGAC
500 510 520 530 540 550 560
CAGCCTGACAAACATGGTGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGCCGGACGTGGTGGCGCA
570 580 590 600 610 620 630
TGCCTGTAATCCCAGCTACTTGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGCGGAGGTTGCG
640 650 660 670 680 690 700
ATGAGCCGAGATCGCGCCATTGCACTCCAGCCTGGGAAACAAGAGCGAAATCCGTCTCAAGAAAAAAG
710 720 730 740 750 760 770
GAAAGACCCCCCTCCTTCTCCCGCCGAAATACCCTCTTTCAGGACGGCGCGCCTGTGCGGCGACGCGC
780 790 800 810 820 830 840
GCTCAGTTACTTAGCAACCTCGGCGCTAAGCCACCCAGGTGGAGCCAGCAACAACAGAGCCACCGCGT
850 860 870 880 890 900 910
CCCCACCAATCAGCGCCGACCTCGCCTTCGCAGGCCTAACCAATCAGTGCCGGCGCTGCAAGGAAGTTT
920 930 940 950 960 970 980
HSE2 HSE1
CCAGAGCTTTCGAGGAAGGTTTCTTCAACTCAAATTCATCCGCTGATAATTTTCTTATATTTTCCTAAA
990 1000 HSE3 1010 1020 1030 1040 1050
GAAGGAAGAGAAGCGCATAGAGGAGAAGGGAAATAATTTTTTAGGAGCCTTTCTTACGGCTATGAGGAAT
1060 1070 1080 1090 1100 1110 1120
TTGGGGCTCAGTTGAAAAGCCTAAACTGCCTCTCGGGAGGTTGGGCGCGGCGAACTACTTTCAGCGGCGC
1130 1140 1150 1160 1170 1180 1190
ACGGAGACGGCGTCTACGTGAGGGGTGATAAGTGACGCAACACTCGTTGCATAAATTTGCCTCCGCCAGC
1200 1210 1220 1230 1240 1250 1260
CCGGAGCATTAGGGGCGGTTGGCTTTGTTGGGTGAGCTTGTTTGTGTCCCTGTGGGTGGACGTGGTTGG
1270 1280 1290 1300 1310 1320 1330
TGATTGGCAGGATCCTGGTATCCGCTAACAGGTACTGGCCCGCAGCCGTAACGACCTTGGGGGGGTGTGA
1340 1350 1360 1370 1380 1390 1400
GAGGGGGGAATGGGTGAGGTCAAGGTGGAGGCTTCTTGGGGTTGGGTGGGCCGCTGAGGGGAGGGCGTGG
1410 1420 1430 1440 1450 1460 1470
GGGAGGGGAGGGCGAGGTGACGCGGCGCTGGGCCTTCCGGGACAGTGGGCCTTGTTGACCTGAGGGGGG
1480 1490 1500 1510 1520 1530 1540
CGAGGGCGGTTGGCGCGCGGGTTGACGGAACTAACGGACGCCTAACCGATCGGCGATTCTGTCGAGT
1550 1560 1570 1580 1590 1600 1610
TACTTCGCGGGGAAGGCGGAAAAGAGGTAGTTTGTGTGGTTTCTGGAAGCCTTTACTTTGGAATCCCAG
1620 1630 1640 1650 1660 1670 1680

Alu B8#1

Alu B8#2

Intron

TGTGAGAAAGGTGCCCTTCTTGTGTTTCAATGGGATTTTTATTTTCGCGAGTCTTGTGGGTTTGGTTTTG
 1690 1700 1710 1720 1730 1740 1750
 TTTTCAGTTTGCCTAACACCGTGCTTAGGTTTGAGGCAGATTGGAGTTCGGTCGGGGGAGTTTGAATATC
 1760 1770 1780 1790 1800 1810 1820
 CGGAACAGTTAGTGGGGAAAGCTGTGGACGCTTGGTAAGAGAGCGCTCTGGATTTTCCGCTGTTGACGTT
 1830 1840 1850 1860 1870 1880 1890
 GAAACCTTGAATGACGAATTTTCGTATTAAGTGACTTAGCCTTGTAATAATTGAGGGGAGGCTTGCGGAATA
 1900 1910 1920 1930 1940 1950 1960
 TTAACGTATTTAAGGCATTTTGAAGGAATAGTTGCTAATTTTGAAGAATATTAGGTGTAAAAGCAAGAAA
 1970 1980 1990 2000 2010 2020 2030
 TACAATGATCCTGAGGTGACACGCTTATGTTTTACTTTTAAACTAGGTCAAATGCAGATCTTCGTGAAA
 2040 2050 2060 2070 2080 2090 2100
 T L T G K T I T L E V E P S D T I E N V K A K
 ACCCTTACCGGCAAGACCATCACCTTGAGGTGGAGCCCAGTGACACCATCGAAAATGTGAAGGCCAAGA
 2110 2120 2130 2140 2150 2160 2170
 I Q D K E G I P P D Q Q R L I F A G K Q L E D G
 TCCAGGATAAGGAAGGCATTCCCCCGACCAGCAGAGGCTCATCTTTGCAGGCAAGCAGCTGGAAGATGG
 2180 2190 2200 2210 2220 2230 2240
 R T L S D Y N I Q K E S T L H L V L R L R G G
 CCGTACTCTTTCTGACTACAACATCCAGAAGGAGTCGACCCTGCACCTGGTCCTGCGTCTGAGAGGTGGT
 2250 2260 2270 2280 2290 2300 2310
 MET Q I F V K T L T G K T I T L E V E P S D T I
 ATGCAGATCTTCGTGAAGACCCTGACCGGCAAGACCATCACCTGGAAGTGGAGCCCAGTGACACCATCG
 2320 2330 2340 2350 2360 2370 2380
 E N V K A K I Q D K E G I P P D Q Q R L I F A G
 AAAATGTGAAGGCCAAGATCCAGGATAAAGAAGGCATCCCTCCCGACCAGCAGAGGCTCATCTTTGCAGG
 2390 2400 2410 2420 2430 2440 2450
 K Q L E D G R T L S D Y N I Q K E S T L H L V
 CAAGCAGCTGGAAGATGGCCGCACTCTTTCTGACTACAACATCCAGAAGGAGTCGACCCTGCACCTGGTC
 2460 2470 2480 2490 2500 2510 2520
 L R L R G G MET Q I F V K T L T G K T I T L E V
 CTGCGTCTGAGAGGTGGTATGCAGATCTTCGTGAAGACCCTGACCGGCAAGACCATCACTCTGGAGGTGG
 2530 2540 2550 2560 2570 2580 2590
 E P S D T I E N V K A K I Q D K E G I P P D Q Q
 AGCCCAGTGACACCATCGAAAATGTGAAGGCCAAGATCCAAGATAAAGAAGGCATCCCCCGACCAGCA
 2600 2610 2620 2630 2640 2650 2660
 R L I F A G K Q L E D G R T L S D Y N I Q K E
 GAGGCTCATCTTTGCAGGCAAGCAGCTGGAAGATGGCCGCACTCTTTCTGACTACAACATCCAGAAAGAG
 2670 2680 2690 2700 2710 2720 2730
 S T L H L V L R L R G G C *
 TCGACCCTGCACCTGGTCCTGCGCCTGAGGGGTGGCTGTTAATTCTTCAGTCATGGCATTTCGAGTGCCC
 2740 2750 2760 2770 2780 2790 2800
 AGTGATGGCATTACTCTGCACTATAGCCATTTGCCCAACTTAAGTTTLAGAAATTACAAGTTTTCAGTAAT
 2810 2820 2830 2840 2850 2860 2870
 AGCTGAACCTGTTCAAATGTTAATAAAGGTTTCGTTGCATGGTAGCATACTTGGTGTTTTGTGCATGAAA
 2880 2890 2900 2910 2920 2930 2940
 TTCTCTAGTGATGTGTGGGTACGCTTAAACTGGTGAAAATGTTTLAGGATTTAATTTTGGAGATTGGTAA
 2950 2960 2970 2980 2990 3000 3010
 TGTGCTCAAAGTTAAGTCACTTTGACTTTGGTATACTTGGGTGGGCTGAGGGGCAAGAGCCTTCTTTGC
 3020 3030 3040 3050 3060 3070 3080
 TGTTTAAGTCATTACAAGTTAGGATCC
 3090 3100

Intron ←

MET Q I F V K

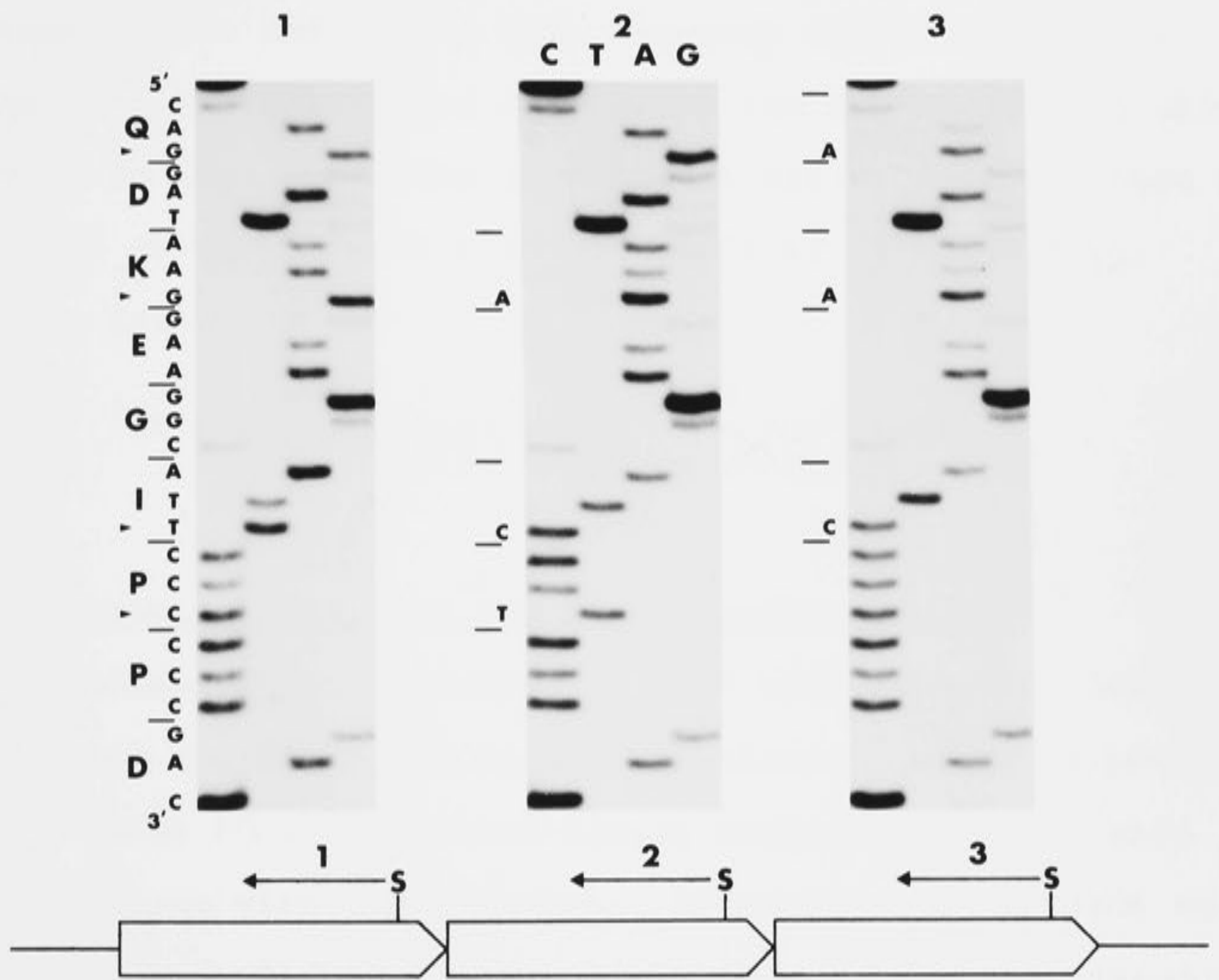
(v)

the degeneracy of the genetic code, and are a further feature of the stringent evolutionary constraints on the ubiquitin system. Some of the DNA sequence differences are shown in Figure 3.6.3.

Analysis of the 3' NCR around the polyadenylation signal and site reveals other sequences that have been implicated in 3' end formation. McLauchlan et al (1985) have identified a consensus sequence YGTGTTY present in 67% of mammalian genes examined approximately 30bp 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 31bp from the AATAAA signal (Figure 3.6.2, 2924 to 2931). Another sequence, CAYTG, is possibly involved in the selection of a polyadenylation site in conjunction with the AATAAA signal via hybridisation with the small nuclear RNA U4 (Berget, 1984). A similar sequence, CATGG (Figure 3.6.2, nt 2909 to 2913) occurs in EHB8 within the polyadenylation region AATAAA(N)₁₀CATGGNA*, where N is any nt and A* is the first residue of the poly(A) tail. This region matches the consensus for a Class I U4 RNA hybrid (Berget, 1984), suggesting 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 (Urano et al, 1986). Another sequence observed in several eukaryotic genes is TTCAA or close derivatives, which occurs 32 to 62bp downstream of AATAAA (Urano et al, 1986). EHB8 contains a recognised derivative, TTAAA, 67bp from AATAAA (Figure 3.6.2, nt 2965 to 2970), but also contains the consensus TTCAA beginn-

Figure 3.6.3: UbB Coding Unit Nucleotide Sequence Variation

Panels 1, 2 and 3 represent sequencing gel autoradiograms of the region encoding codons 31 to 39 of the first, second and third coding units of the UbB gene respectively. The nucleotide sequence and translation of the first coding unit are given at the left of Panel 1, reading 5' to 3'/N- to C-terminal from top to bottom. Positions of nucleotide sequence variation are indicated with an arrowhead on Panel 1 and the variant nucleotides are listed beside Panels 2 and 3. All changes are in the third position of the codon and do not affect the encoded aa sequence. The diagram below the figure is a schematic representation of the sequencing reactions. Boxes are ubiquitin coding units. S:SalI



ing 11bp upstream of AATAAA (nt 2882 to 2887). Involvement of this sequence in mRNA 3' end formation has yet to be confirmed (Urano et al, 1986).

The 3' NCR contains direct and inverted repeats downstream of the termination codon. The 13bp sequence CAGTC/GATGGCATT occurs at nt 2778/2790 and 2800/2812. The region 2783 to 2830 contains an imperfect 22bp inverted repeat which could form a stem-loop structure with a ΔG of -19.0kcal (Tinoco et al, 1973) as follows:

```

                CGC           T
2770 TAATTCTTCAGTCATGGCATT AGTGCCCAGTGA G
                2830 TACCGATA TCACGTCTCATT CG
                                   A

```

3.6.2 Characterisation of the 5' Non-coding Region

Initial sequence determination of EHB8 only extended 350bp upstream of the first coding unit initiation codon. As pRBL26 was a partial cDNA, no 5' non-coding sequence was available for comparison with EHB8. However, comparison can be made to the ~100bp mutually homologous 5' flanks of the three pseudogenes EHB4, EHD1 and EHB7, which are theoretically representative of the mRNA 5' NCR. The results of such a comparison are as follows. All 4 sequences are identical for the first 8bp upstream of the first coding unit, but further upstream EHB8 shows no similarity to any of the pseudogene regions. The sequences diverge 5' of the sequence AGGT, which matches the consensus AGGT/G resulting from the splicing of an intron (Breathnach and Chambon, 1981). The EHB8 sequence at this point is TTTTACTTTTAAACTAGGT (Figure 3.6.2, nt 2060 to 2078), which is a reasonable match to the splice acceptor consensus $(y)_m nyagGT/G$ (intron in lower case; Breathnach and Chambon,

1981). These observations indicate that an intron may exist in the 5' NCR of EHB8.

A pseudogene 5' NCR-specific probe was generated from a 196bp EcoRI/BglII fragment from the 5' flank of pseudogene EHD1 (Figure 3.4.1). This probe contained 9bp of ubiquitin coding sequence, 98bp of 5' flank common to all 3 pseudogenes and a further 89bp specific to EHD1 (Figure 3.4.2, nt 1 to 196). The probe hybridised to pB8.3 restriction fragments upstream of a BamHI site located ~ 740bp upstream of the first coding unit, in particular to 350bp TaqI/BamHI and 80bp MspI/BamHI fragments (Figure 3.6.1). The nucleotide sequence around this BamHI site was determined by the strategy shown in Figure 3.6.1 and is included in Figure 3.6.2. Comparison of this sequence with the pseudogene 5' flanks revealed that this BamHI site was within the gene's 5' homologous region and had been lost from each pseudogene as a result of base changes. Homology resumes at the sequence CAGGTA (Figure 3.6.2, nt 1359 to 1364), which matches the splice donor consensus A/CAGgtr (intron in lower case, r is a purine; Breathnach and Chambon, 1981). These results further support the presence of an intron within the 5' NCR of this gene.

3.6.3 Characterisation of an Intron Within the 5' Non-coding Region

Sequencing of the 740bp BamHI/BglII fragment upstream of the coding region (Figures 3.6.1 and 3.6.2) revealed that the intron was 715bp in length, with the "coding-like" strand relatively G rich (35.5%) and C poor (15.8%), producing an overall GC content of 51.3%. Many (56%) of the G residues lie

within the clusters G_nXG or GXG_n ($n = 2$ to 7 , $X = A, T$ or C) which occur 37 times within the intron. The sequence AGGT at the splice junctions is redundant: the exon/intron and intron/exon boundaries in Figure 3.6.2 have been chosen to confer with the "GT-AG" rule (Breathnach and Chambon, 1981). Two possible splice branch point sequences (Keller and Noon, 1984) are present: GTGAC (nt 2046 to 2050) and CTTAT (nt 2054 to 2058).

There are several direct repeats within the intron. The two major series are: (numbering refers to Figure 3.6.2)

1455 TGAGGGGAGGGCGTGGGG	1951 ggcTTGc-GGAATATTA
1473 GAGGGGAGGGCGaGGtG	1977 ATTTTGAAGGAATAgTt
1940 TGAGGGGAGG-CtTGcGG	1998 ATTTTGAA-GAATATTA

The intron also contains an imperfect 18bp inverted repeat spaced by 7bp (Figure 3.6.2, nt 1753/1770 and 1778/1795) which could form a stem-loop structure with a ΔG of -15.8 kcal (Tinoco et al, 1973).

3.6.4 Determination of the Transcription Initiation Site

The transcription initiation (mRNA start) site was determined by a combination of S1 nuclease mapping (Chapter 2.2.16) and comparison with the processed pseudogenes, and a subsequently isolated full-length cDNA (Chapter 3.7).

The probe employed for S1 mapping was a 350bp TaqI/BamHI fragment located upstream of the coding units (Figure 3.6.1). The BamHI site is within the region of homology with the pseudogene 5' flanks and provides a reference point to fix one end of protected fragments. Hybridisation of total human lymphocyte RNA (Chapter 2.2.12) to the probe protected a set of 5 fragments 68 to 72 bases in length from digestion by S1

nuclease (Figure 3.6.4). This result places the mRNA start sites over the 5bp range 1274 to 1278, Figure 3.6.2. As each protected fragment was of approximately equal intensity, no major site is indicated.

A second estimate of the mRNA start site was obtained by comparison with the length of the processed pseudogene 5' NCRs. Generally, processed pseudogenes represent full-length DNA copies of mature mRNAs (Sharp, 1983). The three processed pseudogenes described above have arisen from at least two separate mRNAs, based on the different position of polyadenylation of EHD1 (Chapter 3.4). Thus the fact that all three 5' limits of similarity to the gene (Chapter 3.9) cease within a 2bp region, strongly indicates that this position corresponds to the mRNA start site. For EHB4, the 5' limit is nt 1271, Figure 3.6.2, while for EHD1 and EHB7, it is 1272. Interestingly, these positions are respectively 3 and 2bp upstream of the most upstream mRNA start site determined by S1 nuclease mapping.

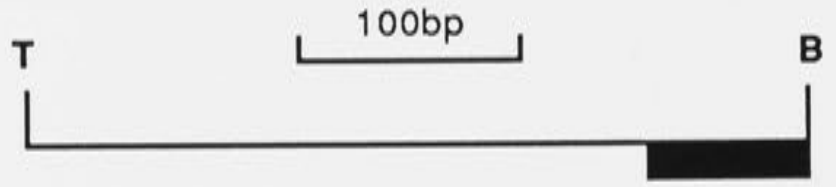
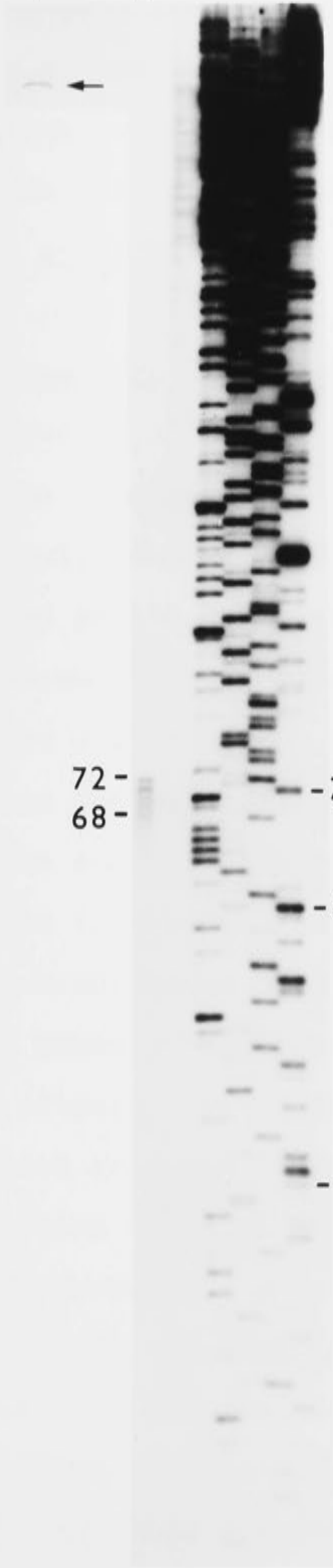
A third estimate was obtained by isolating and sequencing a full length UbB cDNA, described in detail in Chapter 3.7. This cDNA initiated at nt 1274, Figure 3.6.2, which is identical to the most upstream mRNA start site determined by S1 mapping.

These results place the most upstream mRNA start site between nt 1271 and 1274. As both S1 mapping and the full-length cDNA identify nt 1274, this position has been chosen as the mRNA start site for this gene, although it should be noted that transcription also initiates at several positions just 3' to this site. Thus the first exon is 88bp in length, and the 5'

Figure 3.6.4: S1 Nuclease Mapping of UbB mRNA.

The diagram at right indicates the origin of the ~355nt probe generated from the TagI/BamHI fragment of the UbB gene containing most of exon 1. The panel at left shows the resulting S1 mapping results. Lane 1 = undigested probe (arrowed). Lane 2 = fragments protected from S1 nuclease digestion following hybridisation with total lymphocyte RNA. Sizes are in nt and are from the sequencing ladder.

1 2



Uniformly Labelled Probe (Lane 1)

Protected Fragments (Lane 2)

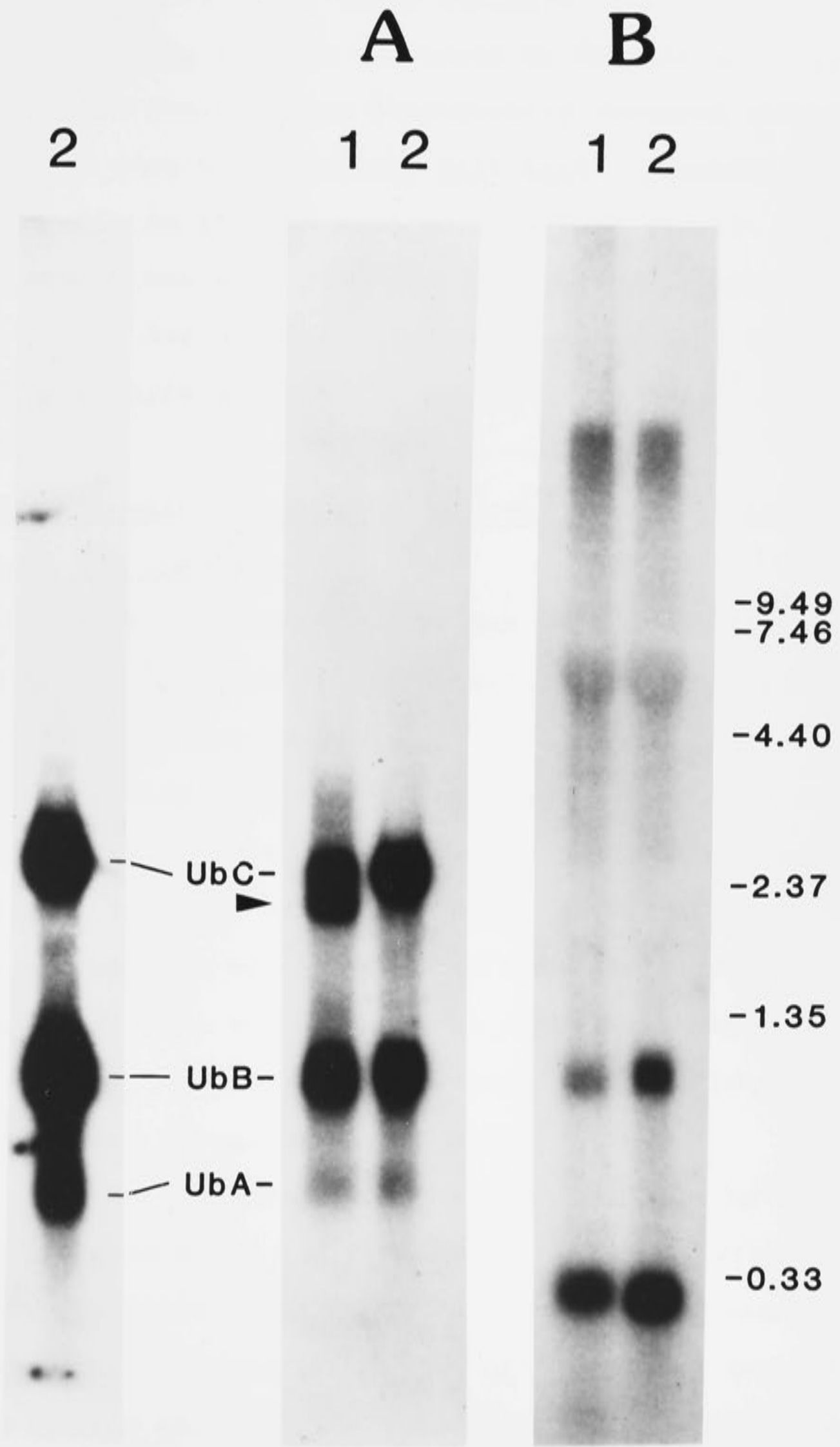
NCR is 94bp, interrupted by a 715bp intron. The 96 to 97bp pseudogene-homologous region exhibits between 69 and 80% similarity to the pseudogene 5' NCRs (Chapter 3.9). A transcript of this gene initiating at this mRNA start site with the intron spliced would be 927 nucleotides plus a poly(A) tail in length, which corresponds to the observed UbB mRNA of approximately 1000 to 1100 nucleotides (Lund et al, 1985; Wiborg et al, 1985).

3.6.5 Specific Hybridisation of the EHB8 Gene to the UbB mRNA

The identity of EHB8 as a UbB gene was confirmed by the specific hybridisation of a EHB8 5' NCR probe to the previously identified UbB mRNA species (Wiborg et al, 1985). EHB8 and pRBL26 had been previously assumed to be of the UbB class based on their dissimilarity to known UbA and UbC genes (Chapter 3.1) and from the similarity in lengths of the UbB mRNA and the calculated mature EHB8 mRNA (Chapter 3.6.4). Northern analysis (Chapter 2.2.13) was conducted on RNA isolated from human lymphocytes prepared from whole blood and from a transformed cell line (Chapter 2.2.12). A coding region probe (Chapter 3.2) was used to identify all ubiquitin mRNAs (Figure 3.6.5). The EHB8 5' NCR specific probe was generated from the 1.3kb PstI/BamHI fragment upstream of the gene, which contains 72bp of the 5' NCR exon. Hybridisation of this probe to a parallel Northern blot specifically selected the UbB mRNA, confirmation that EHB8 is a UbB gene. However, the probe also identified a non-ubiquitin RNA of about 300nt (Figure 3.6.5). Unknown at the time, it is most likely that this observed hybridisation is between the Alu

Figure 3.6.5: Northern Analysis with a UbB-Specific Probe.

Lane 1: Total lymphocyte RNA prepared from freshly drawn blood. Lane 2: Total RNA prepared from cultured human lymphocytes. Panel A represents hybridisation with a ubiquitin coding region probe (Chapter 3.2). Ubiquitin RNA species are identified UbA, UbB and UbC after Wiborg et al (1985). Lane 2 at left is a longer exposure of a similar blot to emphasise the UbA species. Panel B represents hybridisation with a UbB 5' NCR probe (Chapter 3.6.5) which hybridises only to the UbB species of the ubiquitin mRNAs. Additional hybridisation to a ~300nt species is most likely due to Alu sequences present in the probe (see text). The individual in Lane 1 contains an extra ubiquitin mRNA (arrowed) and is discussed in detail in Chapter 5.



repeats subsequently identified upstream of the gene (Figure 3.6.1) and 7S RNA (295nt), which exhibits extensive sequence homology to the Alu repeats (reviewed by Jelinek and Schmid, 1982). Northern analysis was subsequently repeated with a probe derived from the re-cloned TaqI/BamHI fragment, which hybridised only to the UbB mRNA species (not shown). This re-cloned fragment was also used as a probe for S1 mapping (Chapter 3.6.4) and genomic analysis (Chapter 3.8) to prevent further Alu interference.

3.6.6 Identification of Promoter and Enhancer Elements

Upstream of the UbB Gene

Several promoter and enhancer elements were identified upstream of the UbB gene by both position and similarity to consensus promoter sequences.

A TATA box, CATAAAT, begins 34bp upstream of the mRNA start site (Figure 3.6.2, nt 1240 to 1246). A transition from T to C is the most commonly observed deviation at the first position of the TATA box (Breathnach and Chambon, 1981). The position of the TATA box relative to the mRNA start site is in agreement with the consensus positioning of TATA promoters from other genes (Breathnach and Chambon, 1981).

A second conserved promoter often present 50 to 70bp upstream of the TATA promoter is the CCAAT box (consensus GGYCAATCT, Benoist et al, 1980). This sequence may also be present on the complementary strand (eg Graves et al, 1986). Several sequences weakly matching the consensus are present upstream of the TATA box. The most proximal is GGCGAACT, 71bp upstream (Figure 3.6.2, nt 1169 to 1176). Other possible CCAAT boxes

begin 74bp upstream (GCCCAACCT, complement of nt 1158 to 1166) and 113bp upstream (CCCCAAATTC, complement of nt 1117 to 1126), both on the complementary strand.

Most interestingly, three copies of the heat-shock element (HSE; Bienz and Pelham, 1982) are located between 266 and 301bp upstream of the mRNA start site (Figure 3.6.2, nt 971 to 1006). The HSE is a 14bp palindromic sequence exhibiting the consensus CNGAANNTTCNNG, and is necessary for the transcriptional activation of heat shock genes in response to environmental stress (recently reviewed by Bienz and Pelham, 1987). The UbB gene promoter region contains three overlapping HSEs as follows:

```

                                C--GAa--TTC--G
961 CCGGCGCTGCAAGGAAGTTTCCAGAGCTTTCGAGGAAGGTTTCTTCAACT 1010
                                c--GAA--TTC--G          c--GAA--TTC--g

```

The first two HSEs match the consensus at 7 of the 8 conserved positions, while the third matches at only 5. Overlapping HSEs occur frequently in heat-shock gene promoter regions (Bienz and Pelham, 1987). HSEs function as both TATA-proximal elements and as upstream enhancers to confer heat-inducibility on heterologous promoters (Bienz and Pelham, 1987), and the location of these HSEs upstream of the UbB gene infers that this gene is a heat-shock gene.

3.6.7 Identification of Alu Repetitive DNA Sequences Upstream of the UbB Gene

Sequence analysis revealed the presence of two members of the Alu family of repetitive DNA sequences upstream of the UbB promoter region. The most upstream of the two is a complete Alu repeat in opposite transcriptional orientation to the UbB

gene and is flanked by imperfect 16bp direct repeats (Figure 3.6.2, nt 93 to 418). The second Alu repeat is truncated by 34bp at its 5' end and is in the same transcriptional orientation as the UbB gene (nt 510 to 776). The two repeats are thus spaced by only 100bp and are 195bp upstream of the heat-shock elements (Chapter 3.6.6). These Alu repeats are described in detail in Chapter 7.

3.7 Isolation of Full-Length Human Liver UbB cDNA Clones

Full-length UbB cDNA clones were isolated from a human liver cDNA library constructed in λ gt11 (Chapter 2.1.4) to confirm the location of an intron within the 5' NCR of the UbB gene (Chapters 3.6.2 and 3.6.3). The probe used for library screening was a 355bp TaqI/BamHI fragment, which contains 72bp of the 5' non-coding exon and continues upstream until the heat shock elements (Figure 3.6.2, nt 990 to 1345). This probe contains no ubiquitin coding sequence and is therefore specific for the UbB gene. Screening (Chapter 2.2.11) of approximately 200,000 recombinants produced a large number of positives, from which 10 clones were selected for re-screening. Three clones (λ Li1, 2 and 4) were selected for sequence analysis of the 5' ends. Notably, clone λ Li1 hybridised much more strongly than the other clones. The initiation points of these clones are indicated in Figure 3.7.1. All 3 clones were sequenced from their 5' ends across the splice junction to confirm the position of the intron. Individual sequences are not presented, as all 3 clones matched exactly the determined UbB gene sequence except for the absence of intronic sequences. Clone λ Li2 initiated at


```

GCCACCGCGTCCCCCACCAATCAGCGCCGACCTCGCCTTCGCAGGCCTAACCAATCAGTG
      910           920           930           940           950           960
                                HSE2
CCGGCGCTGCAAGGAAGTTTCCAGAGCTTTCGAGGAAGGTTTCTTCAACTCAAATTCATC
      970 HSE1 980           990 HSE3 1000           1010           1020
CGCCTGATAATTTTCTTATATTTTCTAAAGAAGGAAGAGAAGCGCATAGAGGAGAAGGG
      1030 1           1040           1050           1060           1070           1080
AAATAATTTTTTAGGAGCCTTTCTTACGGCTATGAGGAATTTGGGGCTCAGTTGAAAAGC
      1090           1100           1110           1120           1130           1140
CTAAACTGCCTCTCGGGAGGTTGGGCGCGGCGAACTACTTTCAGCGGCGCACGGAGACGG
      1150           1160           1170           1180           1190           1200
CGTCTACGTGAGGGGTGATAAGTGACGCAACACTCGTTGCATAAATTTGCCTCCGCCAGC
      1210           1220           1230           1240            1250           1260
      ▽ ▽ ▽ ▽ ▽
CCGGAGCATTTAGGGGCGGTTGGCTTTGTTGGGTGAGCTTGTTTGTGTCCCTGTGGGTGG
      1270 2 1280 4           1290           1300           1310           1320
ACGTGGTTGGTGATTGGCAGGATCCTGGTATCCGCTAACAGGTACTGGCCCGCAGCCGTA
      1330           1340           1350           1360            1370           1380

```

Figure 3.7.1: Full-length UbB cDNA Clones.

This Figure presents the region of Figure 3.6.2 from nt 901 to 1380 in detail. The initiation points of the UbB cDNA clones λ L11, 2 and 4 are shown by horizontal arrowheads numbered 1, 2 and 4 respectively. Breakpoints of homology with the processed pseudogenes are shown by open triangles, while mRNA start sites identified by S1 nuclease mapping (Figure 3.6.4) are shown by small arrowheads. Heat shock element (HSE)-like sequences are underlined/overlined and numbered. The TATA promoter and the splice donor GT dinucleotide are heavily underlined. Numbering is the same as in Figure 3.6.2.

the most upstream mRNA start site (Figure 3.7.1, nt 1274) as determined by S1 nuclease mapping (Chapter 3.6.4), and may therefore be considered a full length clone. Clone λ Li4 initiated 8bp downstream of this point (nt 1282). Clone λ Li1 appears to have arisen from an aberrant mRNA transcript and initiates 242bp upstream of the mRNA start site (nt 1030). This result explains the strong hybridisation signal of λ Li1, which shares 315bp with the probe, compared to only 72 and 64bp for λ Li2 and λ Li4 respectively. The λ Li1 cDNA is discussed further in Chapter 3.11.5.

Clones λ Li2 and λ Li4 were sequenced completely through the 5' NCR and the first ubiquitin coding unit, and were identical to the gene over this region (not shown), thus confirming the cloning artefact nature of the first 23bp of pRBL26 (Chapter 3.1). Clone λ Li1 was also sequenced from the 3' end and found to be polyadenylated at the same position as pRBL26, EHB4 and EHB7 (Chapters 3.1, 3.3 and 3.5), rather than the alternate polyadenylation site of EHD1 (Chapter 3.4). The 3' NCR of λ Li1 differs from that of EHB8 (and pRBL26) at one position: nt 2806 of EHB8 is a T, while λ Li1 has a C (Figure 3.6.2). This difference could be an error of reverse transcription during cDNA cloning, or it may represent allelic variation.

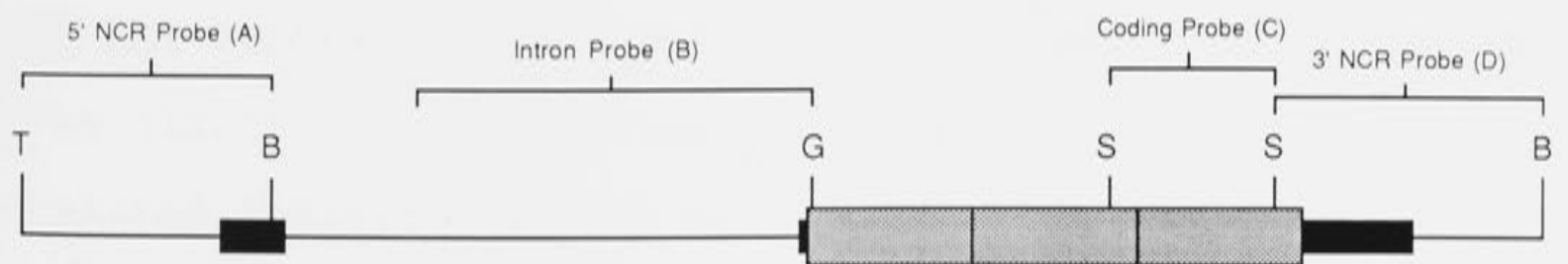
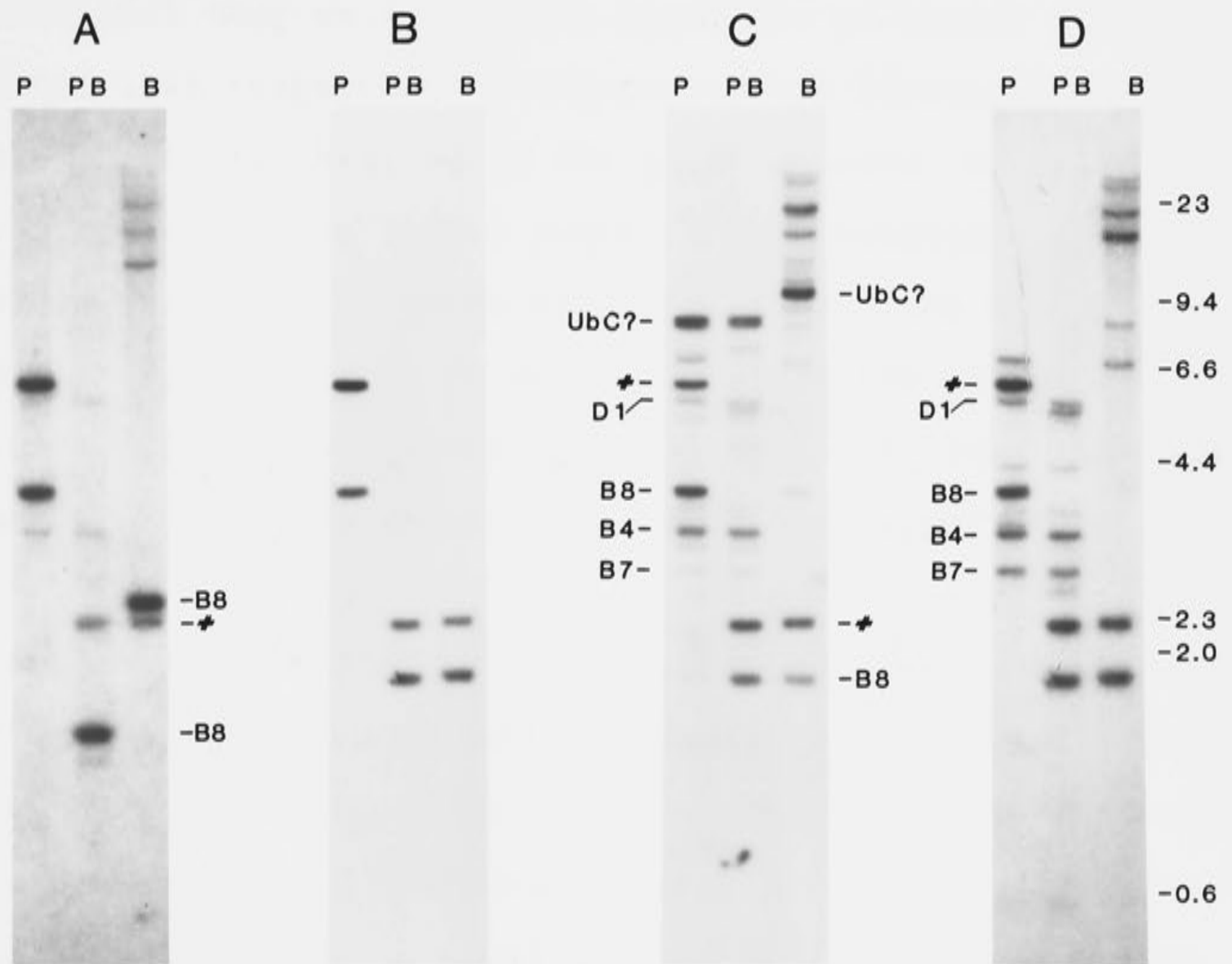
3.8 Hybridisation Analysis of Total Human Genomic DNA

Human genomic DNA (Chapter 2.2.12) was digested with PstI, PstI/BamHI, or BamHI, electrophoresed (Chapter 2.2.3), transferred to a nylon membrane (Chapter 2.2.13) and hybridised with probes derived from different regions of the UbB gene. Neither enzyme cuts within known ubiquitin coding

sequences, while BamHI cleaves within the UbB 5' non-coding exon (Figure 3.6.1) and PstI cleaves within the EHB7 processed pseudogene 5' NCR (Figure 3.5.1). Probes used were (A) A 5' flanking probe, being the same 355bp TagI/BamHI fragment described in Chapter 3.7; (B) An intron probe, being a DNase I deletion subclone (Chapter 2.2.15) spanning three quarters of the intron (Figure 3.6.2, nt 1545 to 2091); (C) A coding region probe as described in Chapter 3.2, and (D) A 3' non-coding probe, consisting of a 375bp SalI/BamHI fragment from the 3' end of the gene (Figure 3.6.2, nt 2730 to 3107). The results of these hybridisations are shown in Figure 3.8.1.

The coding region probe hybridises to a large number of fragments, consistent with previous observations (Wiborg *et al*, 1985) and the large number of positive clones observed during genomic library screening (Chapter 3.2). The other probes produce fewer bands, consistent with their UbB-specific nature. Comparison of the hybridisation patterns with known restriction maps allows the assignment of some ubiquitin genes and pseudogenes to observed hybridising fragments (HFs). A 3.7kb PstI fragment which hybridises to all 4 probes corresponds to the UbB gene EHB8 and is marked B8 (Panel C, Figure 3.8.1). This gene also produces 1.7kb BamHI HFs with the intron, coding and 3' NCR probes, and 1.3kb PstI/BamHI and 2.5kb BamHI HFs with the 5' NCR probe, consistent with its restriction map (Figure 3.6.1). The processed pseudogenes can be similarly assigned. EHB4 (B4) is assigned to a 3.2kb PstI HF with the coding and NCR probes, EHD1 (D1) is assigned to a 5.5kb PstI HF, and EHB7 (B7) to a

Figure 3.8.1: UbB Genomic Hybridisation Analysis.
Genomic DNA ($\sim 5\mu\text{g}/\text{lane}$) was digested with PstI (P),
PstI/BamHI (PB) or BamHI (B). Panels A, B, C and D
represent hybridisation of the resulting Southern blots
with probes derived from different regions of the UbB
gene shown schematically below the figure. Probe A is a
5' NCR probe, Probe B is a DNase I subclone from the
intron, Probe C is from the coding region, and Probe D
is from the 3' NCR. Shaded and black boxes represent
coding and non-coding regions respectively. See text for
details of band assignments.



2.7kb PstI HF. The 5' NCR probe hybridises weakly to the pseudogene HFs, as the 355bp probe shares only ~70bp with the pseudogene 5' flanks at 71 to 80% similarity (Chapter 3.9.2). Restriction maps of the EHB4, EHD1 and EHB7 pseudogenes suggest that they would produce BamHI HFs of larger than 9.5, 10.5 and 14kb respectively, and presumably correspond to some of the large HFs observed in the BamHI digests. Notably, no hybridisation of the intron probe to the pseudogene HFs is observed. The hybridisation signal strength with the coding region probe appears to be proportional to the number of coding units present in the HF. Thus D1 and B7 represent one-repeat HFs, B4 a two-repeat HF, and B8 a 3-repeat HF.

A further HF identified by all 4 probes has been asterisked in Figure 3.8.1. This HF exhibits approximately equal signal intensity with all 4 probes when compared to B8. This result suggests firstly a 3 coding-unit gene; secondly that it has 5' and 3' flanks highly homologous to EHB8, not only in the transcribed regions, but also further up - and downstream; and thirdly that both contain a very similar intron. This HF does not represent restriction fragment length polymorphism, as an identical BamHI/coding probe pattern was observed in 37 unrelated individuals (not shown). These results argue strongly that a duplicated UbB gene exists in the human genome. This duplicated gene produces a 2.2kb BamHI HF with all 4 probes, whereas with EHB8, 5' NCR hybridisation lies on a 2.5kb BamHI fragment adjoining the 1.7kb BamHI fragment that hybridises with the other 3 probes. This observation suggests that the BamHI site within the EHB8 5' NCR is not

present in the duplicated gene (nor is it present in any of the processed pseudogenes - see Chapter 3.9) while the 5' NCR has not been significantly disturbed. Thus, the duplicated gene must contain a BamHI site approximately 500bp upstream from the EHB8 5' NCR BamHI site. Interestingly, the EHB8 sequence contains a BamHI semisite 460bp upstream of the 5' non-coding site, the sequence GGAGCC (Figure 3.6.2, nt 882 to 887), which could revert to a BamHI site with a single transversion from G to T at the fourth position. This observation is mere speculation, as the exact nature of the proposed gene duplication has not been determined. However, the duplication event must have occurred relatively recently in evolutionary history, as hybridisation results suggest that both intronic sequences and sequences outside the transcribed region are very similar in the two genes.

These studies do not reveal the proximity of the two UbB genes, as they are on separate BamHI and PstI fragments. Hybridisation analysis of the EHB8 genomic insert reveals no other ubiquitin-hybridising regions in the clone (not shown), and thus the duplicated gene must be more than ~12.5kb upstream or ~1.2kb downstream of the EHB8 UbB gene (Figure 3.6.1).

An attempt was made to isolate the duplicated UbB gene by screening a human genomic library with the UbB intron-specific probe described above. However, restriction and hybridisation analysis of the 20 clones of varying hybridisation intensity thus obtained revealed that all were identical to the EHB8 genomic clone (not shown), suggesting that

this clone may have been over-represented in the genomic library.

Another HF identified only with the coding region probe (Figure 3.8.1, "UbC?") may represent the 9 coding unit UbC gene. The ~8.5kb PstI and ~10kb BamHI HFs produce a much stronger signal than the 3 coding unit gene, are consistent with the published restriction map (Wiborg et al, 1985), and do not hybridise with any of the UbB-specific probes.

The coding probe identifies several other HFs which presumably represent as yet uncharacterised ubiquitin genes and/or pseudogenes. Some of these HFs correspond to genes and pseudogenes of the UbA ubiquitin-tail fusion protein type, which are discussed in Chapter 4. The 3' non-coding probe also produces other weak HFs which may represent further examples of UbB pseudogenes.

3.9 UbB Gene and Pseudogene Comparisons

3.9.1 Comparison of Coding Regions

The DNA sequences of each of the gene and pseudogene coding units are compared in Figure 3.9.1, both by direct sequence alignment, and as a percentage similarity of each pairwise comparison of coding units. Figure 3.9.1 also presents a comparison of the proteins encoded by each coding unit.

3.9.2 Comparison of Non-Coding Regions

The UbB gene 5' and 3' NCRs are compared with the corresponding pseudogene regions in Figure 3.9.2. Comparison is made in three ways: direct sequence alignment, overall

A

10	20	30	40	50	60
ATGCAGATCTTCGTGAAAACCCTTACCGGCAAGACCATCACCCCTTGAGGTGGAGCCCCAGT					
.....G.....G.....G..A.....					
.....G.....G.....T..G.....					
.....G.....A.....TG.....A.A.....G					
.....G.....G..T.....G..A.....					
.....G.....G..T.....A.....					
.....A.....C.....G.....G.....A...					

70	80	90	100	110	120
GACACCATCGAAAATGTGAAGGCCAAGATCCAGGATAAGGAAGGCATTCCCCCGACCAG					
.....A.....C..T.....					
.....A.....A.....C.....					
.....A.....G.....					
.....G.....A.....CC.....A.....					
.....T.....C..T.....					
...T.C..C.....T.T..A..G....C....T.....					

130	140	150	160	170	180
CAGAGGCTCATCTTTGCAGGCAAGCAGCTGGAAGATGGCCGTACTCTTTCTGACTACAAC					
.....C.....					
.....C.....					
.....T.....G.....A..C...A.....T...					
.....C.....					
.....T.....A.....T.....					
..C..T.....T.....T.A.....T.C...G...G.....					

190	200	210	220	230	
ATCCAGAAGGAGTCGACCCTGCACCTGGTCCTGCGTCTGAGAGGTGGT					EHB8/1
.....					EHB8/2
.....A.....C....G....CTGT					EHB8/3
.....A...A.....A.....					EHB4/1
.....A...A.....AC....G....CTGT					EHB4/2
.....A...T.....T.....A.....T....CTGT					EHD1
..T...A...AG.....C.A.....G....CTAT					EHB7

EHB8/1 vs:	EHB8/2	EHB8/3	EHB4/1	EHB4/2	EHD1	EHB7
% Match :	96.5	94.7	92.1	92.5	93.0	86.4

B

10	20	30	40	50	60
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQORLIFAGKQLEDGRTLSDYN					
.....S...A..E..R.....T.....V..V..D..H....					
.....M.....T.....					
.....N.....L.....M.....C.....					
...L.....ILQ.....HV.....HS...V.....C.VC...					

70		
IQKESTLHLVLRG	EHB8	%MATCH
.....*.....S.....	EHB41	85.5
.....H....C	EHB42	96.0
...L..Y..Q...C.C	EHD1	89.5
.....A.....H....Y	EHB7	81.6

Figure 3.9.1: UbB Gene and Pseudogene Coding Unit Comparisons.

A: Coding unit comparisons. Dots indicate sequence identity with the first EHB8 coding unit (EHB8/1) arbitrarily chosen as a reference. Extra Cys (-like) codons are shown where appropriate.

B: Encoded protein comparisons. Percent similarity to the gene-encoded ubiquitin is given at lower right.

percentage similarity, and a "mutation score". The latter is a sum of the number of positions where one sequence differs from all of the other 3, and gives an indication of new mutations arising in one sequence since its separation from the others. Some of these mutations include insertion and deletion events. EHB7's 5' flank contains a 3bp insertion relative to the other 3 sequences, while the gene EHB8 has suffered a 2bp deletion compared to the 3 pseudogenes (Figure 3.9.2). Likewise, the gene's 3' flank contains a single base insertion compared to the 3 pseudogenes.

3.9.3 Results of UbB Gene and Psuedogene Comparisons

Several conclusions can be drawn from these sequence comparisons. First, EHB7 appears to represent a much older pseudogene than EHB4 or EHD1, as it has the most variant coding unit, and 5' and 3' flanks, with respect to both percentage similarity and mutation score. It also encodes the most variant protein (Figures 3.9.1 and 3.9.2).

Second, the 5' NCRs are more variant than the 3' NCRs. The 5' similarities range from 70.8 to 80.2%, compared to 86.5 to 95.1% for the 3' values (Figure 3.9.2). Likewise, the 5' NCRs exhibit an average mutation score of 8.3 per 100 bp, while that of the 3' regions is only 4.6 per 100 bp.

Third, the gene's 5' NCR appears to be evolving at a faster rate than its 3' NCR, assuming that the pseudogenes are accumulating mutations at a neutral rate. The gene 3' NCR appears to be subject to a slightly negative rate, as may be expected for the conservation of the various 3' processing signals. The gene 3' NCR has the lowest mutation score (3),

5' 10 20 30 40 50 60
 GGAGCATTTAGGGGCGGTTGGCTTTGTTG---GGTGAGCTTGT--TTGTGTCCTGTGGG
 TTCAGTGG.....A..AC.....TA...GG.G.....C
 AT.TTC.GG.....T.AA.....C.....G...GGAG....T.....C
 A.G.A.CAG..T...AA...C...AAT..C...G.C.GGAG.CA.T..A...C
 4 23 1 4 3 4 4 2 4 1 1 44 4 1

70 80 90 100 110
 TGGACGTGGTTGGTGATTGGCAGGATCCTGGTATCCGCTAACAGGTCAAAATG EHB8
A..AC.....T..C.G..CC..G.CA..... EHB4
A..AC...C...C...C...C...T..... EHD1
 ..A.T...A.....CT...CT...C.C...T.C..G..... EHB7
 4 4 1 3 2 4 2 1 2 124 34 4 4

% SIMILARITIES

	EHB4	EHD1	EHB7
EHB8	80.2	81.3	70.8
EHB4	-	84.7	70.4
EHD1	-	-	76.5

Mutation Score

	EHB4	EHD1	EHB7
EHB8	7	6	15

3' 10 20 30 40 50 60
 TAATTCTTCAGTCATGGCATTTCGCAGTGCCCAGTGATGGCATTACTCTGCACTATAGCCA
C.....T..-.....CG.....
 ..G.....T..-...T.....T.....A.....
T..-...T...A.A.....C.....A.....
 2 1 1 3 4 434 4 3 22 4

70 80 90 100 110 120
 TTTGCCCAACTTAAGTTTAGAAATTACAAGTTTCAGTAATAGCTGAACCTGTTCAAAAT
C.....C.....T.....
 ..T.....T.....C.....G.....C.....
 3 3 4 4 4 4 3 4

130 140 150(v) 160
 GTTAATAAAGGTTTCGTTGCATGGTAGCATACTTGGTGTTT EHB8
 .C.....T.....TAA.A.AGAAAA.GAA EHB4
T.....TAAA.AAAA.. EHD1
GC...T.....TAA.A.AAAAAAAGG EHB7
 2 44 1

% SIMILARITIES

	EHB4	EHD1	EHB7
EHB8	95.1	93.9	89.4
EHB4	-	92.9	88.7
EHD1	-	-	87.2

Mutation Score

	EHB4	EHD1	EHB7
EHB8	3	4	12

Figure 3.9.2: UbB Gene and Pseudogene NCR Comparisons.

Comparison of 5' NCRs (top) and 3' NCRs (bottom). Dots signify identity with the first sequence, while dashes represent gaps introduced to maximise alignment. Sequences are named at lower right. Percentage similarities and mutation scores (see text) are given below the figures. Numbers below the EHB7 sequence indicate the number of the sequence at that point differing uniquely to the other three (EHB8=1, EHB4=2, EHD1=3, EHB7=4). The BamHI site in the EHB8 5' NCR is underlined. Start and stop codons are boxed. The mRNA start and polyadenylation sites are arrowed. The variant EHD1 polyadenylation site is in brackets.

compared to EHB4 (4), EHD1 (7) and EHB7 (12). This is also reflected in the percentage similarities, with the gene vs pseudogene value higher than the pseudogene vs other pseudogene values (Figure 3.9.2). Conversely, the gene 5' NCR has the second highest mutation score (7) compared to EHB4 (6), EHD1 (4) and EHB7 (15). Similarly, a calculation of relative mutation scores suggests a faster evolving gene 5' flank. For example, the gene's 5' NCR mutation score is 48.6% of the EHB7 5' score per 100bp (7 vs 15/96bp) compared to only 17.6% for the 3' NCRs (3 vs 12/142bp). This trend is also reflected in the percentage similarities. For example, the 5' regions of pseudogenes EHB4 and EHD1 are more similar to each other (85%) than either is to the gene 5' NCR (79 to 80%). These results suggest that the gene 5' NCR has a positive mutation rate and is actually evolving faster than the neutral (pseudogene) rate.

The above analysis assumes that all 3 pseudogenes have arisen from the UbB gene EHB8. However, it is likely that a duplicated UbB gene is present in the genome (Chapter 3.8). If this gene is transcriptionally active, some or all of the processed pseudogenes may have arisen from it. However, the conclusions reached above should still be valid for the following reasons. First, hybridisation analysis suggests strong sequence conservation between the duplicated genes; most notable is their approximately equal hybridisation to the intron probe (Figure 3.8.1). Therefore, the genes would be expected to be more similar to each other than to any of the pseudogenes. Second, the limited available evidence suggests that only the EHB8 UbB gene is transcriptionally

active, at least in liver tissue. Each of the 3 full-length liver UbB cDNA clones (Chapter 3.7) have 5' NCRs identical to the EHB8 gene, including an intact BamHI site, whereas it appears that the duplicated gene lacks this BamHI site (Chapter 3.8). In addition, the 3' NCRs of pRBL26 and λ Lil1 (with the exception of one base, Chapter 3.7) are identical to that of EHB8, indicating their origin from this gene. However, the possibility exists that the duplicated gene may be transcribed at relatively low levels, thus being represented at a lower frequency in the cDNA libraries, and/or it may be transcriptionally active in a tissue(s) other than liver.

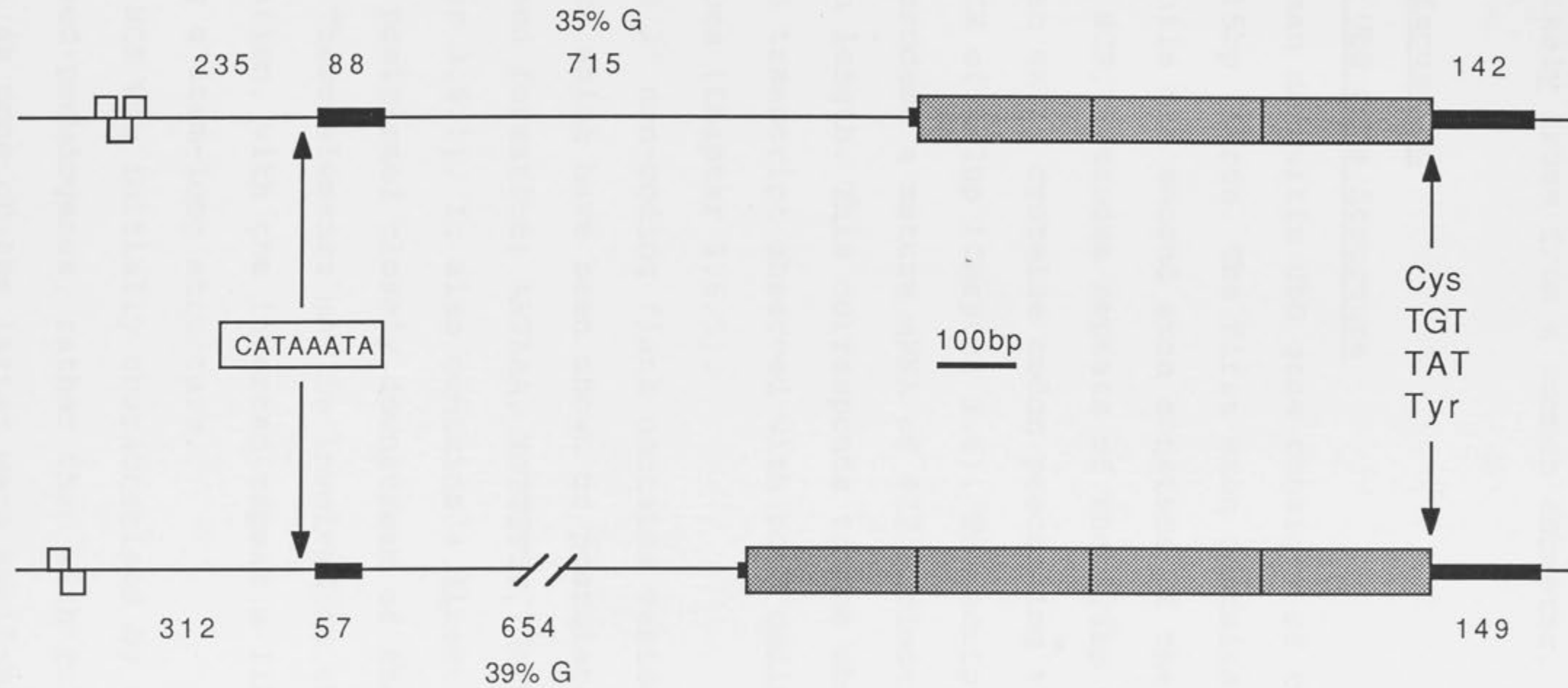
An alternate possibility is that the duplicated gene is not transcriptionally active and is a pseudogene arising from inactivation of a recently duplicated gene, still possessing high sequence similarity to the active gene, and hence, able to hybridise with the intron and non-coding probes. Analysis of the 5' NCR sequence comparison (Figure 3.9.2) suggests that EHB8 has acquired the BamHI site since the generation of the pseudogenes: all pseudogenes have a C at the second position of the BamHI site (Figure 3.9.2). Thus a possible scenario involves the generation of the pseudogenes from a single ancestral gene, followed by duplication of the gene relatively recently, inactivation of one gene of the gene pair, and acquisition of a BamHI site by the other gene (the last two events could occur in either order). Confirmation (or otherwise) of this scenario requires a comparison of the times of pseudogene generation and gene duplication/inactiv-

ation, which must await the sequencing of the duplicated gene/pseudogene.

3.10 Structural Similarities of the Human UbB and Chicken UbI Ubiquitin Genes

At approximately the same time as the UbB gene sequence was reported (Baker and Board, 1987a), Bond and Schlesinger (1986) described the characterisation of a chicken 4 coding unit polyubiquitin gene. With the exception of the difference in coding unit number, the two genes have several features in common which suggest their origin from a common ancestor. The two genes are compared schematically in Figure 3.10.1. Both genes consist of two exons, interrupted by an intron in exactly the same position, between the 6th and 7th bp upstream of the first initiation codon. While the sequences of the NCRs are not conserved (with the exception of the 3' processing signals), their lengths are quite similar, at 142 (human) and 149 (chicken) bp for the 3' NCR, and 94 (human) and 63 (chicken) bp for the 5' NCR. Both genes contain the variant TATA box CATAAAT, and both have heat shock elements in their upstream regions. The human gene has 3 HSEs 235bp upstream of the TATA box, compared to two HSEs 312bp upstream of the chicken TATA box. The introns vary in length (human 715bp, chicken 654bp) but both have a G-rich coding-like strand (human 35.5%, chicken 39.1%) and show obvious clustering of the G residues (see Chapter 3.6.3). Finally, the human gene has Cys (codon TGT) as the extra C-terminal residue, while the chicken gene has Tyr, codon TAT. Thus one codon differs from the other only by a single transition at the central

Human UbB Gene



Chicken Ub I Gene

Figure 3.10.1: Schematic Comparison of the Human UbB and Chicken UbI Genes.

The human UbB gene (top) described in this Chapter is compared to the chicken four coding unit polyubiquitin gene (bottom; Bond and Schlesinger, 1986). See text for a description of features. Open boxes represent heat-shock elements. Other boxes are as in Figure 3.6.1. Numbers represent distances between, or lengths of, features, and are in bp. The % G figures refer to the percentage G composition of the intron "coding" strand.

position (Figure 3.10.1). These similarities suggest that the human UbB and chicken UbI genes are species counterparts and most likely arose from a common ancestor.

3.11 Discussion

3.11.1 UbB Gene Structure

The human ubiquitin UbB gene consists of two exons separated by a 715bp intron. The first exon contains 88bp of the 5' NCR, while the second exon consists of the remaining 6bp of the 5' NCR, 3 tandem repeats of the 228bp ubiquitin coding unit, an extra cysteine codon preceding the stop codon, and a 3' NCR of 142bp (Chapter 3.6). Transcription of the gene would produce a mature mRNA of 927 nucleotides plus a poly(A) tail in length. This corresponds to the observed length of the UbB transcript observed with both coding and NCR specific probes (Chapter 3.6.5).

The UbB 3' non-coding flank contains various sequence elements which have been shown or postulated to be involved in 3' end formation: AATAAA, YGTGTTY, CAYTG and TTCAA (Chapter 3.6.1). It also contains a direct and an inverted repeat positioned closely downstream of the termination codon. These elements may be involved in the termination of translation, with the inverted repeat a likely candidate for forming a stem-loop structure.

The 5' NCR was initially characterised by comparison with UbB processed pseudogenes, rather than with full-length cDNA clones, as none of the latter were available at the time. The common failure of ubiquitin cDNAs to extend into the 5' NCR is discussed below. Processed pseudogenes should be valid

replacements for cDNA clones as they represent in vivo-generated cDNAs, and are generally full-length copies of mature mRNA (Sharp, 1983). For example, two human apoferritin H processed pseudogenes described by Costanzo et al (1985) are full-length reverse transcripts. However, processed pseudogenes may not always be full-length. For example, Giebel et al (1987) report a human heat shock cognate 70 processed pseudogene which lacks the untranslated exon 1 and begins with the first nucleotide of exon 2. The 5' NCR length and sequence suggested by the processed pseudogenes was confirmed by S1 nuclease mapping and isolation of a full-length UbB cDNA clone (Chapter 3.6.4). The full-length cDNAs also confirmed the location of the 5' NCR intron (Chapters 3.6.3 and 3.7). S1 mapping experiments indicated that transcription initiated over a 5bp range, and that the most upstream initiation site was 2 to 3bp downstream of the processed pseudogene initiation sites. This small discrepancy may be explained in one of several ways. First, the S1 result may be due to overdigestion, or to prevention of complete hybrid formation due to interference from the CAP structure. Second, the processed pseudogenes may have arisen from a different, but very similar gene, with a different mRNA start site(s). Hybridisation analysis suggests that the UbB gene is duplicated (Chapter 3.8) and the duplicate may or may not be transcriptionally active (Chapter 3.9.3). Third, the mRNA start site may have altered since the generation of the pseudogenes. This possibility stems from the observation that the site of transcriptional initiation is more dependent on its distance from the TATA promoter than a specific sequence

at the site (Breathnach and Chambon, 1981). Thus, if the region between the TATA box and mRNA start site had suffered a short (2-3bp) deletion, this would appear to shift the mRNA site an approximately similar distance downstream. If such a deletion occurred after the processed pseudogenes had been formed, this would explain the discrepancy in observed mRNA start sites. Indeed, the gene's 5' NCR appears to have suffered a 2bp deletion since pseudogene generation (Chapter 3.9.2), forming a precedent for such a deletion between the TATA box and mRNA start site. Any evidence for or against such an event must await the characterisation of the duplicated gene.

Transcription of the UbB gene appears to be promoted from the variant TATA sequence CATAAAT (Chapter 3.6.6). In the review compiled by Breathnach and Chambon (1981), a transition from T to C was the most commonly observed deviation at the first position of the TATA box. Notably, the chicken UbI poly-ubiquitin gene contains an identical promoter sequence (Bond and Schlesinger, 1986), one of the several similarities between these two genes (Chapter 3.10). A strict adherence to the TATA consensus is not an absolute requirement and several obviously functional variations have been observed. A notable case is the mammalian α -like and β -like globin gene families, where many members contain a "CATA" promoter: for example, CATAAAA in the human β -like globin genes (see Efstratiadis et al, 1980). Of course the TATA box does not function in isolation but in concert with other promoter and enhancer elements, and thus any variation between an observed promoter

and its consensus cannot be automatically translated into a measure of its efficiency.

As described in Chapter 3.6.6, no strong matches to the CCAAT promoter consensus could be found in the expected location upstream of the UbB TATA box, while several weakly matching sequences occur between 70 and 120bp upstream of the TATA box, two of these being located on the complementary strand. The complement of the CCAAT box has been shown to be functional, for example in the herpes simplex virus thymidine kinase gene (Graves *et al*, 1986). Notably, the chicken UbI gene also lacks any strong CCAAT-homologous sequences until a reasonable match CTCCAATCC 110bp upstream from the TATA box on the complementary strand (Bond and Schlesinger, 1986). The exact function and influence of the CCAAT promoter remains to be fully characterised, and the effect of its variance or absence on a gene's promoter region is not fully known.

A most interesting discovery was the identification of heat shock elements upstream of the UbB gene (Chapter 3.6.5). These elements and their implications for UbB gene expression are discussed below (Chapter 3.11.6).

3.11.2 Polyubiquitin Gene Structure

The human UbB gene is structurally similar to other known polyubiquitin cDNAs and genes. The most striking feature is the presence of directly repeated coding units which are not separated by introns or spacer peptides, a structure unique to the ubiquitin genes. The most variable feature within this structure is the number of coding units. Known ubiquitin

genes in man contain 9 and 3 coding units (Wiborg et al, 1985; Baker and Board, 1987a/this Chapter); chicken genes have 4 and 3 (Bond and Schlesinger, 1986); slime mould genes have 5 and 3 (Giorda and Ennis, 1987), while yeast has a single polyubiquitin locus containing 5 or 6 coding units (Ozkaynak et al, 1984, 1987). In addition, cDNA sequences from Xenopus, barley and mouse suggest polyubiquitin genes of 3 or more coding units in these species (Dworkin-Rastl et al, 1984; Gausing and Barkardottir, 1986; St John et al, 1986), while a porcine polyubiquitin cDNA (Einspanier et al, 1987a) originated from a gene of 4 or more coding units. Dworkin-Rastl et al (1984) also identified at least one Xenopus genomic locus containing at least 12 ubiquitin coding units in tandem. Despite this considerable variation in coding unit number, each encodes an identical protein to its fellows, with the single exception of the slime mould 5 coding unit gene, where the second, third and fifth units encode an amino acid substitution at position 28 compared to the other two coding units, and to the slime mould 3 coding unit gene (Giorda and Ennis, 1987). This gene is transcriptionally active as its cDNA has been isolated, and if translated would produce the only known intraspecies ubiquitin protein sequence variation.

A second feature common to polyubiquitin genes is the encoding of a single extra amino acid at the end of the last ubiquitin coding unit. This residue varies both between species, and between different polyubiquitin genes of the same species (see Chapters 1.8 and 3.12.3). For example, the human genes encode Cys (UbB gene, Chapter 3.6.1) and Val (9

coding unit gene, Wiborg et al, 1985). The single reported exception to this feature is a Xenopus polyubiquitin cDNA which encoded no extra amino acid (Dworkin-Rastl et al, 1984). The function of this extra residue is discussed in Chapter 3.11.3 below.

Introns are a less well characterised component of polyubiquitin gene structure. The coding regions of all known polyubiquitin genes lack introns but characterisation of 5' NCRs is generally incomplete due to lack of available cDNA sequences (Wiborg et al, 1985; Bond and Schlesinger, 1986; Baker and Board, 1987a; Giorda and Ennis, 1986; Ozkaynak et al, 1987). Of the three fully characterised polyubiquitin genes, the yeast gene is devoid of introns (Ozkaynak et al, 1987) while the human UbB and chicken UbI (Bond and Schlesinger, 1986) genes contain an intron in their 5' NCR (see Chapter 3.10). The precise conservation of the position of intron in these two genes suggests that some polyubiquitin genes of other higher eukaryotes may also be expected to contain a similarly positioned intron.

The significance of the 5' non-coding intron is presently unclear. While many genes contain introns within their 5' flanks, the location of a gene's only intron in its 5' NCR occurs much less frequently. Such genes include the avian feather keratin and associated keratinisation genes (Molloy et al, 1982; Koltunow et al, 1986), the hamster scrapie prion protein gene (Basler et al, 1986) and the human involucrin gene, one of the human keratinisation genes (Eckert and Green, 1986). One of the two rat preproinsulin genes also has a single 5' NCR exon (Lomedico et al, 1979), but this gene is

a special case, being a functional retroposon (Soares et al, 1985), and is discussed further in Chapter 3.11.6. Interestingly, Koltunow et al (1986) found that sequences within the 5' NCR intron affect the efficiency of the accurate initiation of transcription of chicken feather keratin genes in Xenopus oocytes. Whether or not the sequences within the UbB polyubiquitin gene's 5' intron have any regulatory effect on UbB gene expression is not known. As noted in Chapter 3.6.3, the UbB intron contains several directly repeated sequences, the significance of which is presently unknown, although they could represent a recognition sequence for a regulatory protein.

3.11.3 UbB Translation and Post-translational Processing

The UbB gene does not encode a leader or signal peptide, as there is an in-frame stop codon TAA 4 codons upstream from the first ubiquitin coding unit and more importantly, there are no initiation codons in any reading frame upstream of the initiation codon of the first coding unit (Figure 3.6.2). As was noted for the rat preproinsulin and other genes with 5' NCR introns (Lomedico et al, 1979), the 5' exon has the capability to form a stable stem-loop structure ($\Delta G = -8.8$ kcal, Tinoco et al, 1973) 6 nt before the initiation codon due to a 12bp inverted repeat (Figure 3.11.3). In addition, the open loop can base pair with the 3' end of human 18S rRNA (McCallum and Maden, 1985) forming another stable association ($\Delta = -13.4$ kcal). This structure and its 18S rRNA pairing capability may function in ribosome binding and subsequent translation of the mRNA (Lomedico et al, 1979). Notably, the

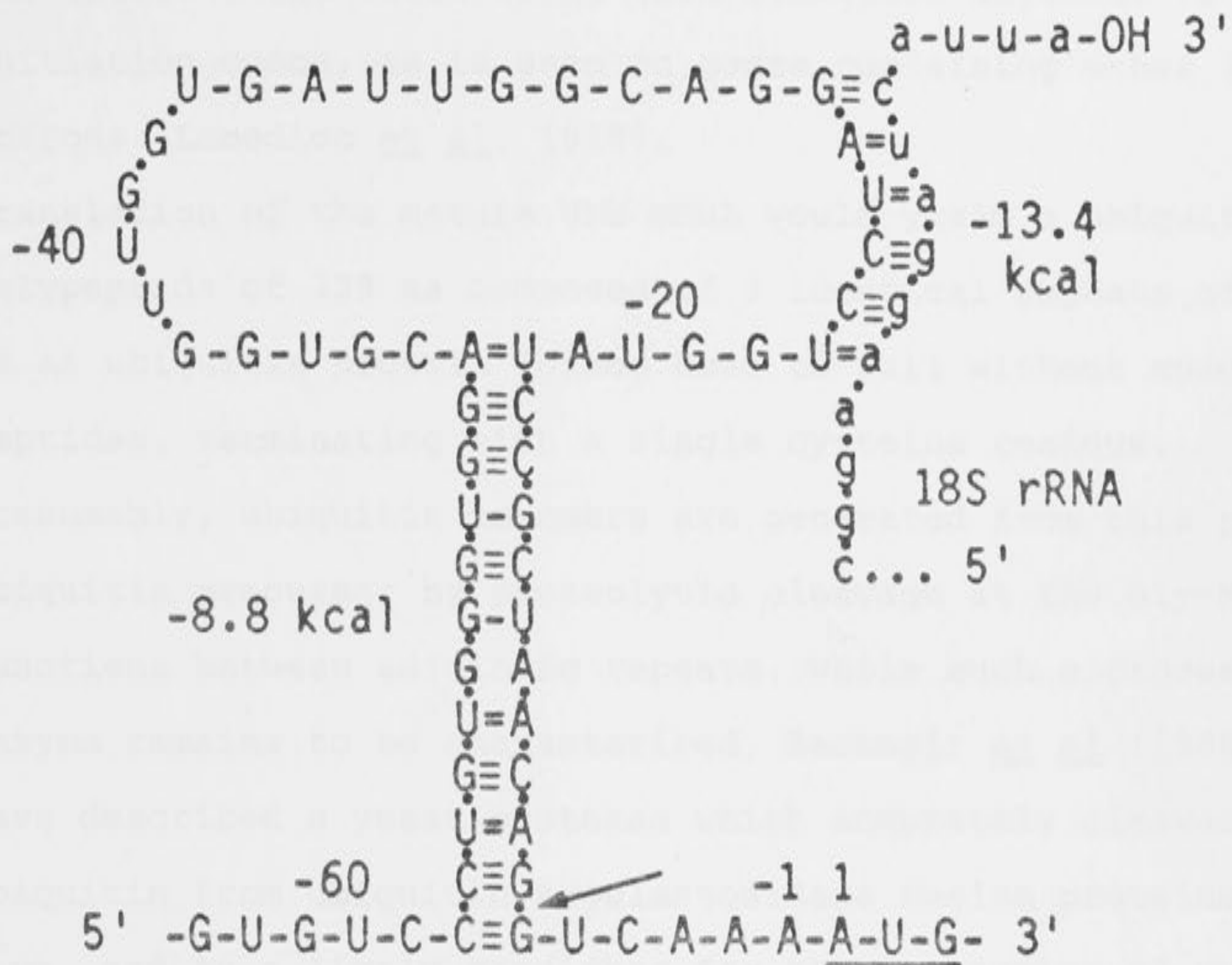


Figure 3.11.3: A Stem-Loop Structure in the UbB 5' NCR. An inverted repeat is present within the UbB 5' NCR exon and could potentially form the stem-loop structure shown above with a calculated ΔG of -8.8 kcal (Tinoco *et al*, 1973). In addition, part of the loop is complementary to the human 18S rRNA as shown, which could form a complex with $\Delta G = -13.4$ kcal. The splice junction is arrowed. The a of the initiation codon (underlined) is numbered 1, with negative numbering used upstream. See text for details and references.

splice junction lies at the base of the stem (Figure 3.11.3): the splice event would bring this structure adjacent to the initiation codon, as is seen in genes containing other 5' NCR introns (Lomedico et al, 1979).

Translation of the mature UbB mRNA would yield a ubiquitin polypeptide of 229 aa composed of 3 identical repeats of the 76 aa ubiquitin protein joined head to tail without spacer peptides, terminating with a single cysteine residue.

Presumably, ubiquitin monomers are generated from this polyubiquitin precursor by proteolytic cleavage at the Gly-Met junctions between adjoining repeats. While such a processing enzyme remains to be characterised, Bachmair et al (1986) have described a yeast protease which accurately cleaves ubiquitin from ubiquitin- β -galactosidase fusion proteins in vivo, and is a likely candidate for the processing of polyubiquitin precursors.

A consistent feature of polyubiquitin genes is the encoding of an extra amino acid at the end of the C-terminal ubiquitin repeat, which varies considerably both within and between species (see Chapter 1.8). This residue is cysteine in the UbB gene. As all of ubiquitin's (known) functions are mediated through the covalent attachment of its C-terminal glycine residue to free amino groups of other proteins, this extra residue effectively blocks any activity of the unprocessed polyprotein. This feature may allow the storage of polyubiquitin in an inactive form, although the half-life of polyubiquitin precursors is not known. A ubiquitin-specific hydrolase that cleaves small adducts from the C-terminus of ubiquitin identified in mammalian cells (Pickart and Rose,

1985b) may be responsible for the removal of these C-terminal residues. Whether the extra residue may function other than in blocking the active ubiquitin C-terminus is not known. The high level of variability in this residue suggests that much less constraint has been placed on it compared to ubiquitin itself, arguing that its location is more important than its identity. It is noteworthy that there is both inter- and intra-species variability in this residue. Three species have been observed to contain two different polyubiquitin genes: man (Wiborg et al, 1985; Baker and Board, 1987a), chicken (Bond and Schlesinger, 1986) and slime mould (Giorda and Ennis, 1987). The human genes terminate with Val (9 coding unit UbC) and Cys (3 coding unit UbB), chicken with Tyr (4 coding unit UbI) and Asn (3 coding unit UBII), and slime mould with Asn (5 coding unit) and Leu (3 coding unit) residues. While there is no direct evidence for the expression of the smaller gene in each of the latter two species, a cDNA encoding Asn at its C-terminus has been isolated which may correspond to the chicken UbII gene (Mezquita et al, 1987), while the 3 coding unit slime mould gene encodes the correct aa sequence (Giorda and Ennis, 1987). Expression of each of these gene pairs would produce polyubiquitins differing only in the number of ubiquitin monomers present, and in the C-terminal residue. If the ubiquitin specific hydrolase (Pickart and Rose, 1985b) presumably responsible for the removal of this residue can distinguish between different aa, it could then selectively post-translationally activate the different polyubiquitin precursors. However, our present knowledge of the speci-

ficities of the potential polyubiquitin processing proteases indicates that they are generally insensitive to sequences downstream of the ubiquitin monomer (Bachmair et al, 1986; Pickart and Rose, 1985b).

3.11.4 UbB Processed Pseudogene Structure and Creation

Three ubiquitin processed pseudogenes were isolated during the characterisation of the UbB gene (Chapters 3.3, 3.4 and 3.5). These pseudogenes are clearly of the UbB gene type, based on the homology of their NCRs to the corresponding UbB gene regions. They also exhibit the hallmark features of processed pseudogenes: lack of introns, an encoded poly(A) tail, and flanking direct repeats generated by pseudogene insertion which delineate the sequence similarity between the expressed gene and the pseudogene (Sharp, 1983). Processed pseudogenes appear to be most prevalent in the genomes of higher vertebrates, with some gene families almost entirely composed of processed pseudogene members. For example, the chromosomal protein HMG-17 gene family is the largest known human processed pseudogene family with at least 29 processed members (Srikantha et al, 1987) while the mouse glyceraldehyde-3-phosphate dehydrogenase multigene family contains one active gene and up to 200 processed pseudogenes (reviewed by Weiner et al, 1986).

The time since the divergence of each processed pseudogene from the UbB gene can be estimated from a calculation of the synonymous rate of nt substitution between the coding (like) regions of each gene/pseudogene pair. This rate was calculated by the method of Li et al (1985), and varies slightly

depending on which gene coding unit is used for comparison. Using a value of 5×10^{-9} nt substitutions per site per year as the synonymous mutation rate (eg Hayashida and Miyata, 1983), EHB4, EHD1 and EHB7 arose a minimum of ~ 14.5 , 18 and 28 million years (Myr) ago respectively. These estimates have large relative errors ($\pm 6, 5$ and 8 Myr respectively) due to the low number of synonymous sites (~ 50) available for comparison, and are within the vicinity of estimates of generation times of other mammalian processed pseudogenes (Weiner et al, 1986). Interestingly, these times predate current estimates of human/higher primate divergence times, and thus other higher primate species would be predicted to contain these pseudogenes. Notably, these estimates assume a constant evolutionary rate of nt substitution.

The three observed UbB processed pseudogenes appear to represent reverse transcription of at least two mRNA species. The mRNA giving rise to EHD1 (Chapter 3.4) is polyadenylated at a different site than the other two processed pseudogenes EHB4 and EHB7 (Chapters 3.3 and 3.5) and the two UbB cDNAs pRBL26 and λ Lil1 (Chapters 3.1 and 3.7). This difference could represent normal use of alternate UbB polyadenylation sites, an erroneous polyadenylation event, or a transcript of a different, but very similar gene which employs a different polyadenylation site.

The processed pseudogene EHB4 has a reading frame severely disrupted by a stop codon and a frame-shift insertion and could not produce a functional protein. The other two pseudogenes EHD1 and EHB7 both have open reading frames and potentially encode variant ubiquitins. However, as processed

pseudogenes can be inserted at any genomic location and generally only contain mature RNA sequences, they would be devoid of their promoter elements, control regions and processing signals. Thus, EHD1 and EHB7 are unlikely to be expressed. In fact, transcription of these pseudogenes is undetectable by Northern analysis with a UbB-specific probe (Chapter 3.6.5). In addition, heterogeneity in human ubiquitin protein sequence has not been observed to date.

The most striking feature of the pseudogenes is the precise deletion of one (EHB4) or two (EHD1 and EHB7) coding units compared to the UbB gene. Most importantly, the 5' and 3' NCRs have not been significantly affected by these coding unit deletions. There are several ways of explaining these structural differences. First, EHB4, D1 and B7 may represent processed pseudogenes of single and double coding unit versions of the 3 coding unit UbB gene, which may have been precursors in the evolution of ubiquitin genes (see Chapter 8). If this were true, then precursor genes themselves, or non-processed pseudogenes of them, should be present in the genome. However, the fact that to date only processed pseudogenes have been isolated tends to discount this explanation. However, such a possibility cannot be completely excluded without the characterisation of the remaining UbB-like loci identified by the UbB 3' NCR probe. Second, the processed pseudogenes may have originally been "full-length" 3 coding unit pseudogenes, but have lost one or two coding units through unequal crossover events, the most likely way for such a precise deletion to occur. However, 4 and 5 coding

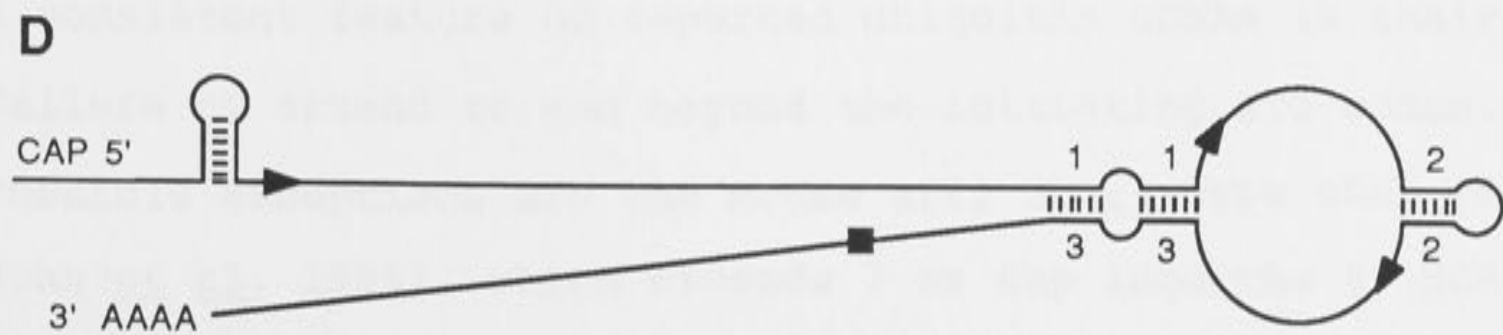
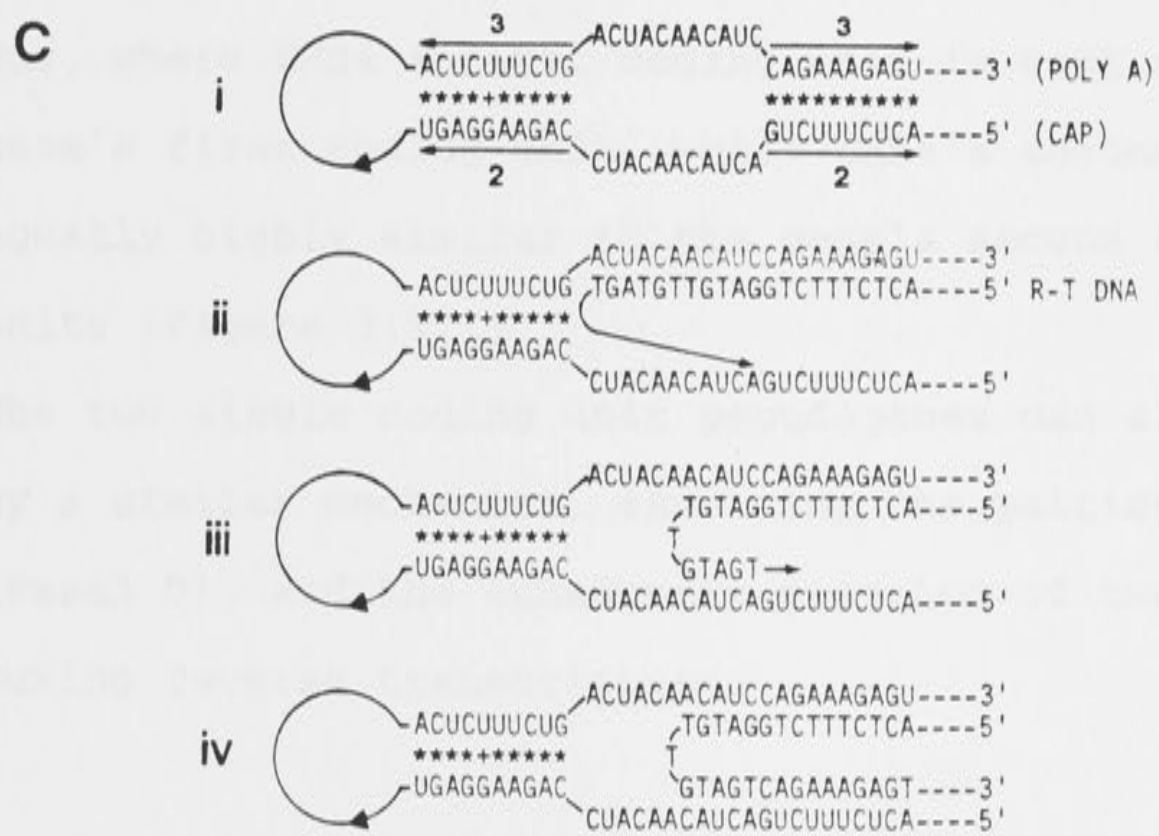
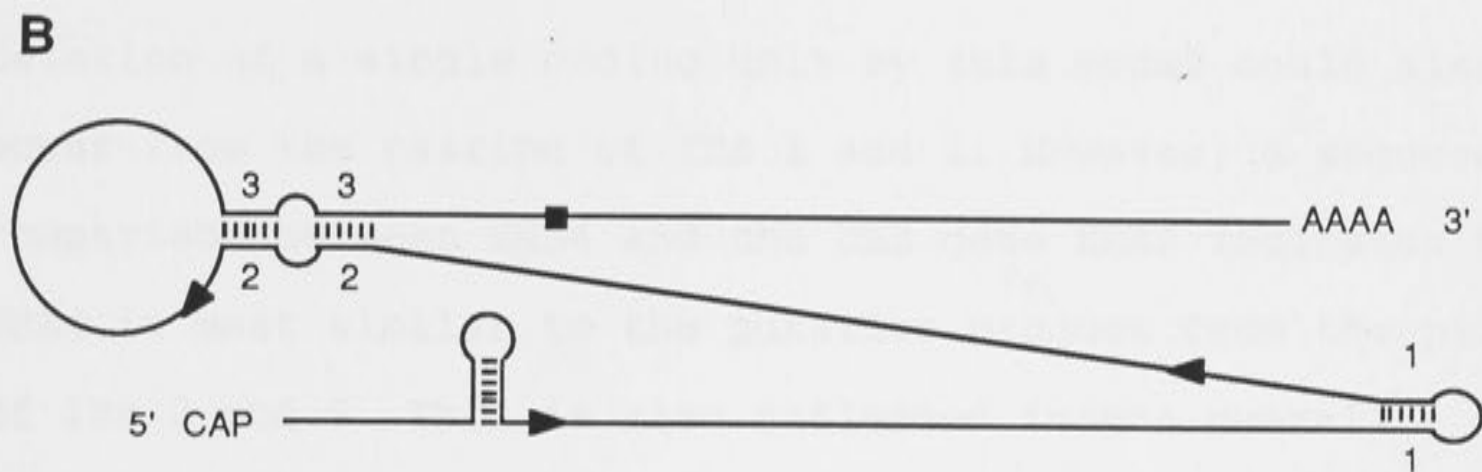
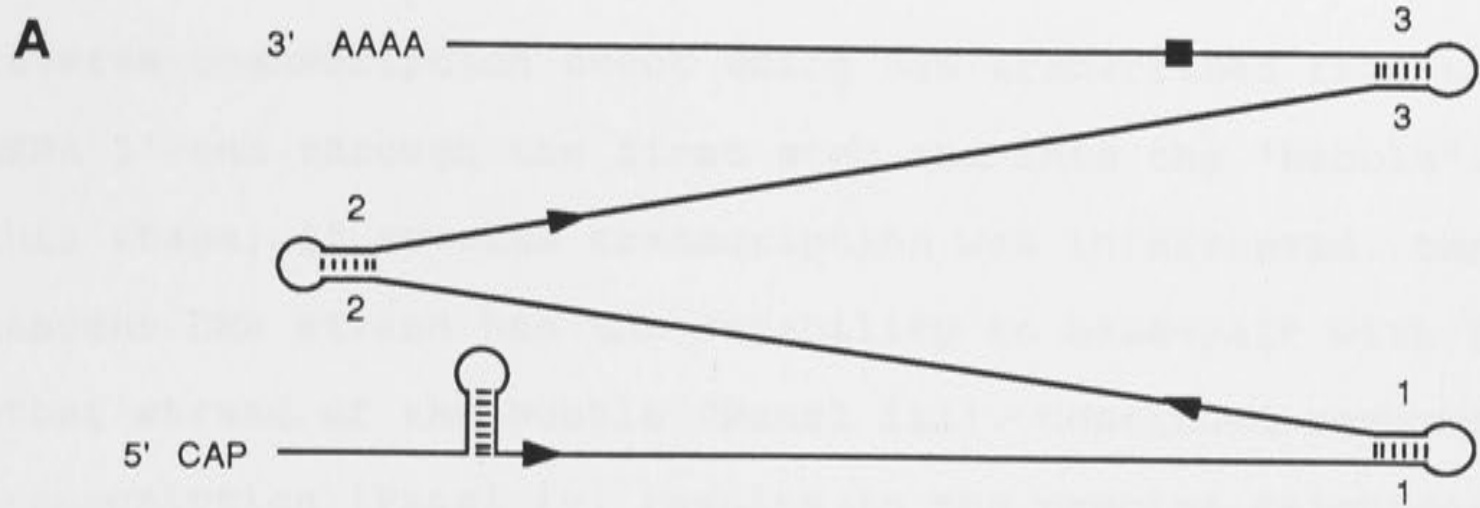
unit processed pseudogenes would also be produced from such crossovers, and as mentioned above, there is no evidence at present for the existence of such loci in the human genome. Also, variable coding unit number pseudogenes resulting from such events would be flanked by similar direct repeats, which is not observed.

The third explanation is that pseudogene creation and coding unit deletion occur simultaneously rather than as separate events. The fact that no "full-length" 3 coding unit processed pseudogenes have been observed or appear to exist (Chapter 3.8) suggests that the deletion of coding units occurs as a consequence of the reverse transcription events leading to pseudogene creation. A model developed to explain such a phenomenon is presented in Figure 3.11.4. This model involves the formation of UbB mRNA secondary structures due to the inverted repeats (IRs) within each coding unit (see Chapter 3.1) and the interference of these structures in reverse transcription. Presumably the IRs within each coding unit would pair to form stable stem-loop structures with $\Delta G = -8.4$ kcal (IRs 1 and 2) and -11.8 kcal for IR3 (Tinoco et al, 1973) as shown in Panel A. However, it is possible that the IRs in one coding unit could pair with the IRs in another unit. Such an event is shown in Panel B, with IRs 2 paired with IRs 3 to form a loop-stem-bubble-stem structure. Almost one complete coding unit is looped out (197 bases) and the structure is quite stable ($\Delta G = -25.2$ kcal). Panel C shows a putative sequence of events leading to the precise deletion of one coding unit during reverse transcription. The IR2/IR3 structure is shown in detail in (i). Panel (ii) shows a

Figure 3.11.4: UbB mRNA Secondary Structure and Processed Pseudogene Generation.

Panels A, B and D: possible UbB mRNA secondary structures formed by base-pairing (dashes) between inverted repeats (IRs) numbered by their coding unit of origin (1, 2 or 3). The 5' NCR stem-loop is also shown (Figure 3.11.3). Arrowheads indicate AUG start codons, filled boxes represent the stop codon. The mRNA 5' CAP structure and 3' poly(A) tail (AAAA) are indicated.

Panel C: detailed structure of the IR2/IR3 stem-loop structure from Panel B and putative reverse transcription events. See text for details. R-T DNA = reverse transcribed DNA. Base-pairing is shown by *, with a G-U pair by +.



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 DNA strand has the capability to base-pair with the other strand of the bubble (Panel iii). Continued reverse transcription (Panel iv) results in the precise deletion of one complete coding unit: the 228 nt between IR 2 and IR 3. Deletion of a single coding unit by this model could also occur from the pairing of IRs 1 and 2. However, a sequence comparison between EHB4 and the UbB gene EHB8 indicates that EHB4 is most similar to the putative product from the pairing of IRs 2 and 3. This is also reflected in the overall percentage similarities of the pairwise coding unit comparisons, where EHB4's first coding unit is most similar to the gene's first coding unit, while EHB4's second coding unit is equally highly similar to the gene's second and third coding units (Figure 3.9.1)

The two single coding unit pseudogenes can also be explained by a similar mechanism, involving the pairing of IRs 1 and 3 (Panel D), and the subsequent deletion of two coding units during reverse transcription.

3.11.5 Ubiquitin cDNA Clone Structure

A consistent feature of reported ubiquitin cDNAs is their failure to extend to and beyond the initiating ATG codon. Possible exceptions are the mouse arf2 UbA₅₂-type cDNA (St John *et al*, 1986), which extends 7 or 8bp into the 5' NCR, and the slime mould pLK229 cDNA clone, containing 5bp of 5' NCR (Giorda and Ennis, 1987). This phenomenon has prevented

the characterisation of gene 5' NCRs by the direct method of comparison with cDNA sequences, and has required the use of more indirect methods: the sequencing of primer extended products for the chicken UbI gene (Bond and Schlesinger, 1986) or comparison with processed pseudogenes for the human UbB gene (this Chapter). Characterisation of the transcriptionally active human UbC 9 coding unit gene (Wiborg et al, 1985) and slime mould 5 coding unit gene (Giorda and Ennis, 1987) remains incomplete for this reason.

The human liver UbB cDNA clone pRBL26 is no exception and initiates within the first coding unit at the position of the inverted repeat (Chapter 3.1). Not surprisingly, two other reported cDNA clones initiate in a similar position. The Xenopus polyubiquitin cDNA (Dworkin-Rastl et al, 1984) initiates co-linearly with the downstream inverted repeat, suggesting that the inverted repeat has formed a snap-back loop structure and primed second strand synthesis, with the loop being removed by S1 nuclease digestion (see below). The polyubiquitin cDNA isolated from barley (Gausung and Barkardottir, 1986) initiates 9 bp upstream of the Xenopus position, within the loop formed by the inverted repeats. Presumably, these three cDNA libraries were constructed by a method which encourages snap-back loop formation as a means of priming second strand synthesis, followed by blunt-ending of the loop by limited S1 nuclease digestion (eg Efstratiadis et al, 1976).

The full-length liver UbB cDNA clones described in Chapter 3.7 thus represent the first reported full-length ubiquitin cDNA clones, and enabled confirmation of the location of the

5' NCR intron. The cDNA library (Chapter 2.1.4) containing these clones was constructed by the method of Gubler and Hoffman (1983) (Dr G. Howlett, personal communication), which does not rely on snap-back self priming and is thus more likely to produce full-length clones. One of the UbB cDNA clones, λ Lil, appears to have arisen from an aberrant mRNA which has initiated 242 bp upstream of the normal mRNA start site (Chapter 3.7). This initiation site is obviously also well upstream of the UbB gene TATA box and raises the question of how it was promoted. No TATA-like sequences are evident, with the possible exception of CAAATT, 19 bp upstream (see Figure 3.7.1). One possibility is that the heat shock elements have functioned as the transcriptional promoter, as they lie 25 bp upstream of the aberrant start site. A second possibility is that the λ Lil cDNA represents part of a longer transcript initiating within the Alu repeat upstream of the heat shock elements. Although this Alu is truncated by 34 bp at its 5' end, it still contains the RNA polymerase III-like internal promoter sequence presumed responsible for promoting in vitro transcription of Alu members by RNA polymerase III (see Jelinek and Schmid, 1982). The λ Lil mRNA has been correctly spliced and polyadenylated (Chapter 3.7), but most interestingly, contains its own promoter region in the 5' elongated transcript. Should such a mRNA become a template for processed pseudogene creation, the "pseudogene" thus generated would contain its own promoter and could therefore be transcriptionally active. There is evidence that such events can occur and the genes created have been termed functional retroposons or retrogenes (Weiner

et al, 1986). One example is the rat and mouse preproinsulin I gene, which is a retrogene arising from a transcript initiating ~0.5kb upstream of the usual mRNA start site of the normal 2-intron preproinsulin gene (Soares et al, 1985). The transcript was incompletely processed (one intron unspliced) before the insertion event, leaving the preproinsulin I gene with the remnant of a poly(A) tail and flanked by a 41 bp direct repeat. This event occurred in the common rat/mouse ancestor after divergence from the rest of the mammals, as the two species have the same direct repeat (Soares et al, 1985). Similarly, the human prointerleukin 1 β gene encodes a poly(A) tail and is flanked by a 17bp direct repeat (although no introns have been removed), and is thought to be a retrogene of the prointerleukin 1 α gene (Clark et al, 1986). The intronless human and hamster β 2-Adrenergic receptor genes (Kobilka et al, 1987) and chicken calmodulin I gene (Gruskin et al, 1987) are also candidate retrogenes.

3.11.6 The UbB gene and Heat Shock

The UbB gene contains three overlapping copies of the heat shock element (HSE) positioned 266bp upstream of the mRNA start site (Chapter 3.6.6). While there is no direct evidence that transcription of the UbB gene is induced by heat shock, the identification of these HSEs is suggestive of a role for the UbB gene in the stress response, as HSEs are found in multiple copies in the promoter regions of all known heat shock genes (reviewed by Bienz and Pelham, 1987). Ubiquitin has been identified as a heat shock protein in chicken (Bond and Schlesinger, 1985) and in yeast (Finley et al, 1987). A

polyubiquitin heat shock gene has been identified in each of these organisms. In yeast, the polyubiquitin gene UBI4 has its expression elevated by stress (Ozkaynak et al, 1987), is the only gene that can provide ubiquitin for the stress response (Finley et al, 1987) and contains two overlapping HSEs 365bp upstream of the first codon, one a perfect match to HSE consensus (Pelham and Bienz, 1982) and the other matching at 7 of the 8 conserved positions (Ozkaynak et al, 1987). Similarly, expression of the chicken polyubiquitin Ubi gene is elevated approximately 5-fold during heat shock, and the unspliced mRNA also accumulates (Bond and Schlesinger, 1985, 1986). The chicken Ubi gene also contains two overlapping HSEs 345bp upstream of the mRNA start site, which match the consensus at 8 and 6 positions.

The similarities between the chicken Ubi and the human UbB genes have been described in Chapter 3.10, and it appears that these genes may be species homologues. Given that the UbB gene contains three HSEs, albeit none of which match all 8 consensus positions (7,7 and 5), it is most likely that it is a human ubiquitin heat shock gene, by analogy with the yeast and chicken polyubiquitin genes. As noted by Ozkaynak et al (1987), the repetitive structure of the polyubiquitin gene would be presumably advantageous in reducing the metabolic cost of ubiquitin synthesis during stress. Unfortunately the 5' NCR of the human 9 coding unit UbC gene has not been fully characterised (Wiborg et al, 1985), and thus there is no information on its possible role as a heat shock gene.

3.11.7 The UbB Subfamily

Ubiquitin genes constitute a gene family in all organisms examined to date, which is comprised of both polyubiquitin genes, and ubiquitin/tail fusion genes. While yeast contains a single polyubiquitin locus (Ozkaynak et al, 1987), the three other examined species (man, chicken and slime mould) each contain at least two polyubiquitin loci which can be differentiated on the basis of coding unit number and non-coding sequence variation (Wiborg et al, 1985; Baker and Board, 1987a/this Chapter; Bond and Schlesinger, 1986; Giorda and Ennis, 1987). In humans, these constitute the UbB and UbC subfamilies. The results described in this chapter delineate the UbB subfamily, which consists of one active gene, at least three processed pseudogenes, and also a duplicated gene structurally similar to the UbB gene which may or may not be transcriptionally active. Hybridisation analysis suggests that other as yet unidentified UbB-related sequences are present in the genome which are less conserved than the identified pseudogenes. These most likely represent older and/or more divergent UbB pseudogenes, either processed or non-processed. Thus the UbB subfamily joins the growing list of gene families comprised mainly of processed pseudogenes, which occur most commonly in mammals (reviewed by Weiner et al, 1986). As these gene families are stably inherited, it is generally assumed that the processed pseudogenes were created in germ line cells, even though some processed pseudogenes are of genes expressed specifically in somatic tissue(s). Pseudogenes of the latter presumably arise from aberrant germline transcripts (see Weiner et al, 1986). However,

recent evidence suggests that retroviruses may function in processed pseudogene creation. Linial (1987) has found that retroviruses can package cellular mRNAs, transport them to a new cell, reverse-transcribe them and integrate them into the genome of the new host cell. Such a process would allow the transfer of mRNAs from somatic to germline cells, and even from organism to organism, and may account for the large number of "non-germline-expressed" processed pseudogenes. However, ubiquitin is most likely to be expressed in all tissues including germline cells, and the UbB processed pseudogenes may not have originated by "retrofection".

CHAPTER 4

THE HUMAN UBIQUITIN Uba₅₂ SUBFAMILY

CHAPTER 4: THE HUMAN UBIQUITIN UbA₅₂ SUBFAMILY4.1 Isolation and Sequence Analysis of a Human UbA₅₂Pseudogene, EHD5

Chapter 3 describes the isolation and characterisation of several genomic clones containing members of the UbB subfamily. During the isolation of these genomic clones, several other clones were isolated which hybridised with a ubiquitin coding sequence probe, but failed to hybridise with a probe specific for the 3' non-coding region of a UbB cDNA clone (see Figure 3.2.1). One such clone, termed EHD5, was further characterised by sequence analysis to reveal a single ubiquitin coding unit. Figure 4.1.1 presents the restriction map of EHD5, while the nucleotide sequence of its ubiquitin-like region is presented in Figure 4.1.2. However, the encoded protein differed from ubiquitin at 12 residues, and also included a nonsense codon, indicating that it was a pseudogene (see Figures 4.1.2 and 4.8.1). The sequence 3' to the coding unit continued in an open reading frame of 36 codons, the first 18 of which were 83% similar to a C-terminal extension of a partial mouse ubiquitin cDNA (St. John *et al*, 1986), and the first 25 of which were 68% similar to the tail extensions of the yeast UBI1 and UBI2 genes (Ozkaynak *et al*, 1987) (see Chapter 4.8). It thus appeared that EHD5 was a pseudogene of a human homologue of these genes. Later results indicated that this homologue was a UbA-type gene (Chapter 4.4) and encoded a tail protein of 52aa (Chapter 4.2), and was thus named UbA₅₂, to distinguish it from the previously identified 80aa tail fusion UbA cDNA

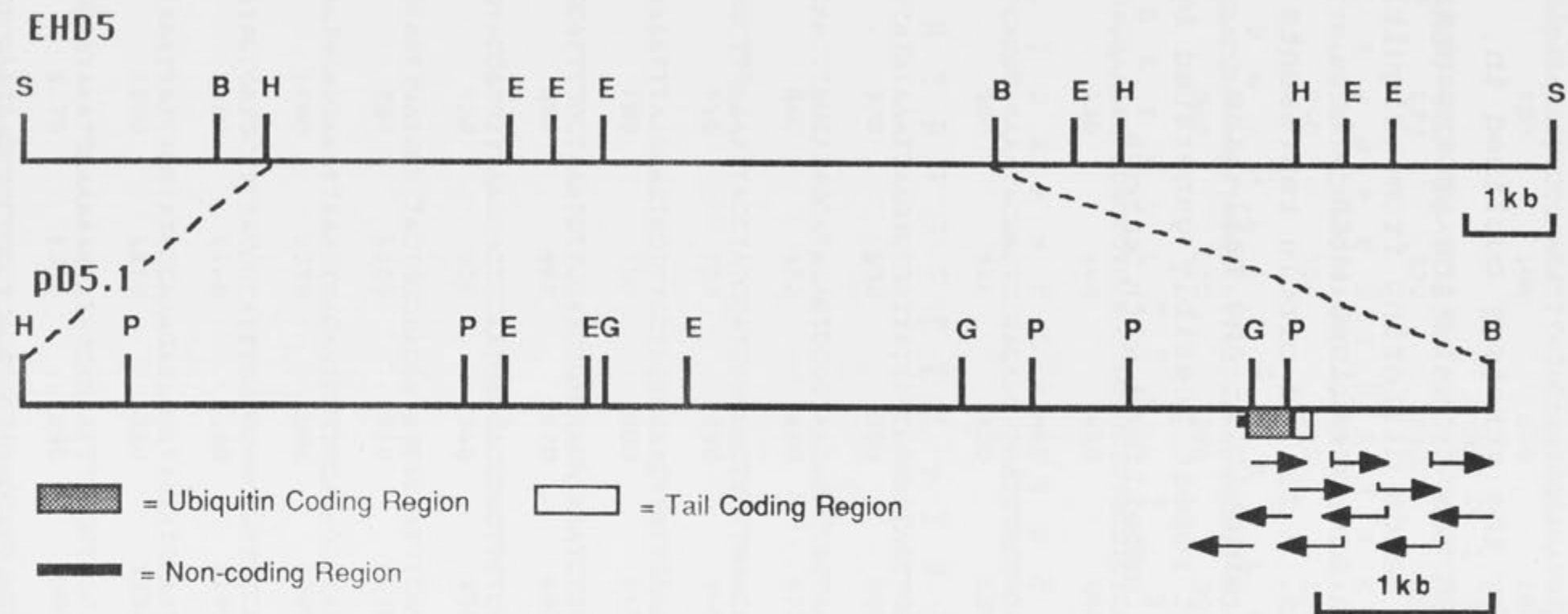


Figure 4.1.1: Restriction Map and Sequencing Strategy of UbA₅₂ Genomic Clone EHD5.

Top Line: Sall insert of EHD5.

Lower line: 8.5kb HindIII/BamHI subclone pD5.1. Boxes represent coding and non-coding regions as indicated. Arrows represent the direction and extent of sequencing. Arrows originating from a vertical bar represent DNaseI-generated deletion subclones. B:BamHI, E:EcoRI, G:BglIII, H:HindIII, P:PstI, S:Sall.

Figure 4.1.2: Nucleotide Sequence of the UbA₅₂-like Region of EHD5.

The sequence determined by the strategy outlined in Figure 4.1.1 is given with the translation of the UbA₅₂-like region above. Amino acids differing from ubiquitin and the 52aa tail protein are overlined with a bar. Asterisks are stop codons. A filled circle represents the junction between the ubiquitin- and tail-like sequences. a short direct repeat possibly generated by pseudogene insertion are underlined with arrows.

CCTGGGATTTGGCATTAGGCAAACTGCTATAAGAGTAAGGAAATATTCTTACCCCCACTGCATTTTTGT
 10 20 30 40 50 60 70
 TTTGGCTAGTCAGTCTCCCTTCTTATACTCTACTTTTATTTCACTCAAAAAGCTGCTGCCTCATACTATA
 80 90 100 110 120 130 140
 TCAGAGAGACTGCTGAACATTTAGACAAATGGGACATGTAGAAAAACAGAAGATTTTCAGCTCTTTCTCT
 150 160 170 180 190 200 210
 MET Q I S V K T L T G K T I T
 TTAGTGAGGCAGCCAAGCTGACATAGAGATGCAGATCTCTGTAAAGACCCTCACAGGCAAAACCATCACC
 220 230 240 250 260 270 280
 L E V Q P S D T S E N V K A K I H D K E G I P
 CTTGAGGTCCAGCCGAGTGACACCAGTGAAAATGTCAAAGCCAAGATCCACGACAAGGAGGGCATCCCAC
 290 300 310 320 330 340 350
 P D Q Q H L I F V N K Q L E D G R T L S D C N I
 CTGACCAGCAGCATCTGATATTTGTGAACAAACAGTTGGAGGATGGCCGTACTCTTTTCAGACTGTAACAT
 360 370 380 390 400 410 420
 Q K E S T L Q V V L C L G D I I E P S L C L
 CCAGAAAGAGTCCACCCTGCAGGTGGTGCTTTGCCTCTGAGGTGACATTATTGAGCCTTCCCTCTGCCTA
 430 440 450 460 470 480 490
 L T Q K Y N C D K MET I C C N C Y A S A H P E E
 CTCACCCAGAAATACAACCTGCGACAAGATGATCTGCTGCAATTGTTATGCTAGTGCACATCCAGAGGAAA
 500 510 520 530 540 550 560
 S H I N V L L L T I V G V Y Y L P F *
 GTCATATAAATTAAGTGTGATTATTATTAACAATAGTTGGGGTATATTACCTGCCCTGATTCTAGGGACC
 570 580 590 600 610 620 630
 CTAACTTACCATAGCATAACAGTGGGTAACACACTGCTAACAGTGATTAGAGAACAGATGGGAGGCTCCC
 640 650 660 670 680 690 700
 TACCTTGAGATTTACCTATCCATCTGGCCCTTCTTGAACATCTATTTATAGGAACTAACTGCTCTGCTCT
 710 720 730 740 750 760 770
 GGATATTTAAGAACTACTTTCTACTCAAATTTATTGAAGATCAATCAGATAACTGAGTTATTTCTCAGG
 780 790 800 810 820 830 840
 AGCATTTTGCTAAATGTTCCATCCGAGAACATAAATCTTCAGTCGGAAAGGAGTAGAATAGGAGAAGATA
 850 860 870 880 890 900 910
 CGTACCATGTTTAAAGACTCCAATGAGCAGCGACTTGTGAGCCTCACTGTCCCACCAAGTTGTTGGAACC
 920 930 940 950 960 970 980
 TCCAGTTGATCCACTACTGGGAACCATATGTCAATCTCACTAAAGGTATAAGAGGGCAGAGTCAGCACAG
 990 1000 1010 1020 1030 1040 1050
 AAATGCAGGAAATTAAGTGAGAATCCTTGCAAAGGTTATTATAGGTATCTTCTGTATCCATTAATTTA
 1060 1070 1080 1090 1100 1110 1120
 AATACTCATGCATCTACCCTATTTCTCCACTAGTGTCAAGTATAAGATTAAAGACCTGATATATAGAT
 1130 1140 1150 1160 1170 1180 1190
 TTAAATTATCAGGTATCCCAAAGATAGCATAAAATGAAATAAGCAAACCAATCTTCCCTCAAATAAGGTA
 1200 1210 1220 1230 1240 1250 1260
 GGCATAAATCAAAAAAATATAGCAGAATTTTACCTGGCAATCGCACCCCCTAACACCTTTTCTGCTTGG
 1270 1280 1290 1300 1310 1320 1330
 TCTCCAATAACCTCCTGCCTCAGGTAGTATAGCATTCTGACCCGCAACAGTACCCTAGAAAGGACAGAGG
 1340 1350 1360 1370 1380 1390 1400
 AAGAGCTGGCTCAGTAACACTGTAGAAGTTTGAGGTGGAAAATGTCTGCTAATTCTATGCTTACTTGTTA
 1410 1420 1430 1440 1450 1460 1470
 CACTGATGTTTCAAGTGCTTCTTATAACTTTTCATCTTGGAAACAAAGTGTGAGGGTTATATTTCCGGATCC
 1480 1490 1500 1510 1520 1530 1540

(Lund *et al*, 1985), which would then become UbA₈₀. EHD5 was then used to generate a tail-specific probe to isolate UbA₅₂ cDNA clones from a human placental cDNA library to characterise the human ubiquitin-tail protein.

4.2 Isolation and Sequence Analysis of a Human Placental UbA₅₂ cDNA Clone.

A 182bp PstI/HinfI fragment containing 30bp of ubiquitin-like coding sequence and 152bp 3' sequence was derived from EHD5 (Figure 4.1.2, nt 437 to 622) and used to probe a human placental cDNA library (Chapter 2.1.4). Of approximately 250,000 plaques screened, only one repeatedly positive clone was obtained, λ P15. This clone was found to have an insert of 1300bp, much longer than expected from the UbA mRNA size of 600 to 650nt (Lund *et al*, 1985; Wiborg *et al*, 1985). Sequence analysis revealed that the λ P15 insert consisted of two cDNA clones fused head to head, with a poly(A) tail at each end of the insert (Figure 4.2.1). One cDNA of 800bp was found by computer-assisted analysis of the GenBank database to be a placental lactogen hormone (chorionic somatomammotropin) cDNA (Seeburg, 1982). The other cDNA had an open reading frame of 128aa, coding for one 76aa ubiquitin protein followed by a 52aa tail (Figure 4.2.2) which was 81% similar to the yeast UBI1/UBI2 tail proteins (Ozkaynak *et al*, 1987 and Figure 4.8.2). There were 18bp between the lactogen hormone cDNA and the UbA₅₂ initiation codon which are presumably UbA₅₂ 5' NCR. The stop codon was followed by a 90bp 3' NCR and a 15bp poly(A) tail, 27bp downstream of the AATAAA polyadenylation signal (Proudfoot

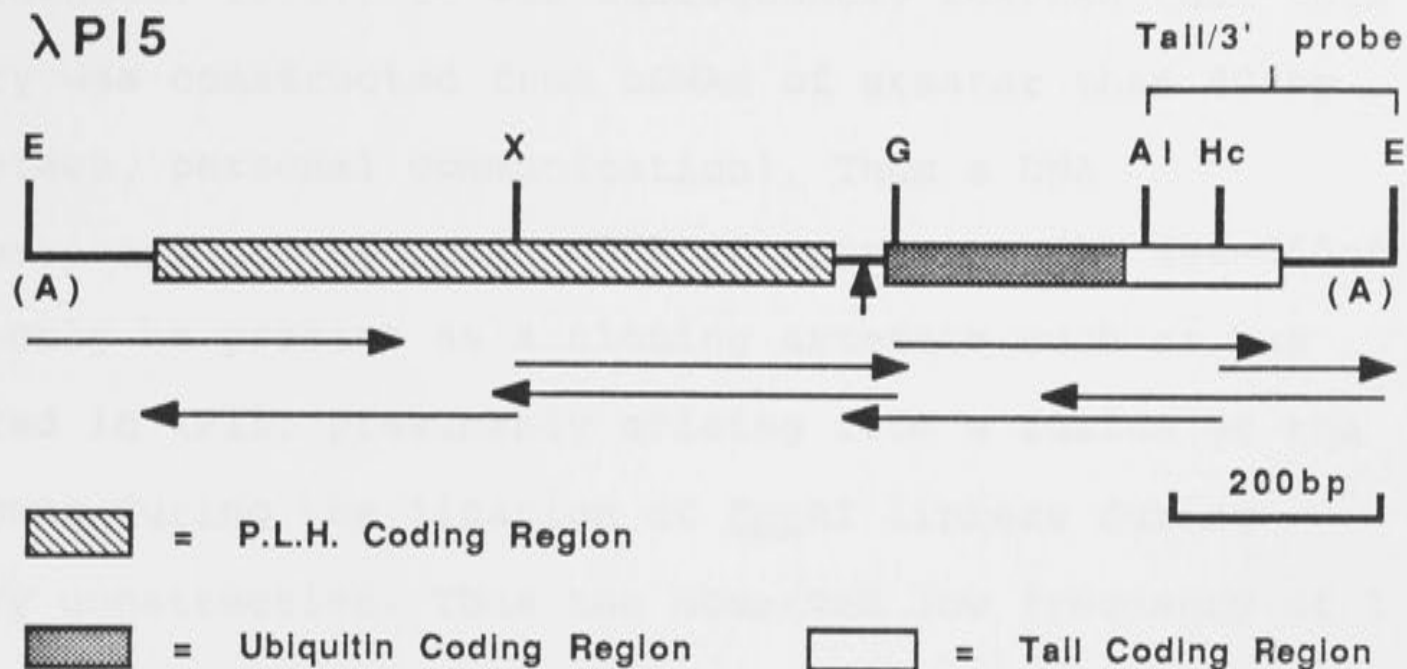
Figure 4.2.1: Restriction Map and Sequencing Strategy of λ P15.

Coding regions are boxed as indicated. Non-coding regions and poly(A) tails (A) are indicated by lines. Arrows indicate the direction and extent of sequencing. a vertical arrow points to the junction of the placental lactogen hormone (P.L.H.) and UbA₅₂ cDNA5. The AluI site (Al) used to generate the tail/3' NCR probe is indicated but is not unique. E:EcoRI; Hc:HincII; G:BglII; X:XbaI.

Figure 4.2.2: Nucleotide Sequence of the UbA₅₂ Portion of λ P15.

The junction of placental lactogen hormone (plh) and ubiquitin (ub) sequence is indicated (nt 51/52). The complement of the plh initiation codon is indicated with an arrowhead. The UbA₅₂ translation is given above the sequence. A filled circle represents the junction of ubiquitin and tail regions. Semicircles indicate tail cysteine residues forming the putative DNA binding domain. The putative nuclear translocation signal is underlined. The asterisk is the stop codon.

λPI5



plh ← → ub

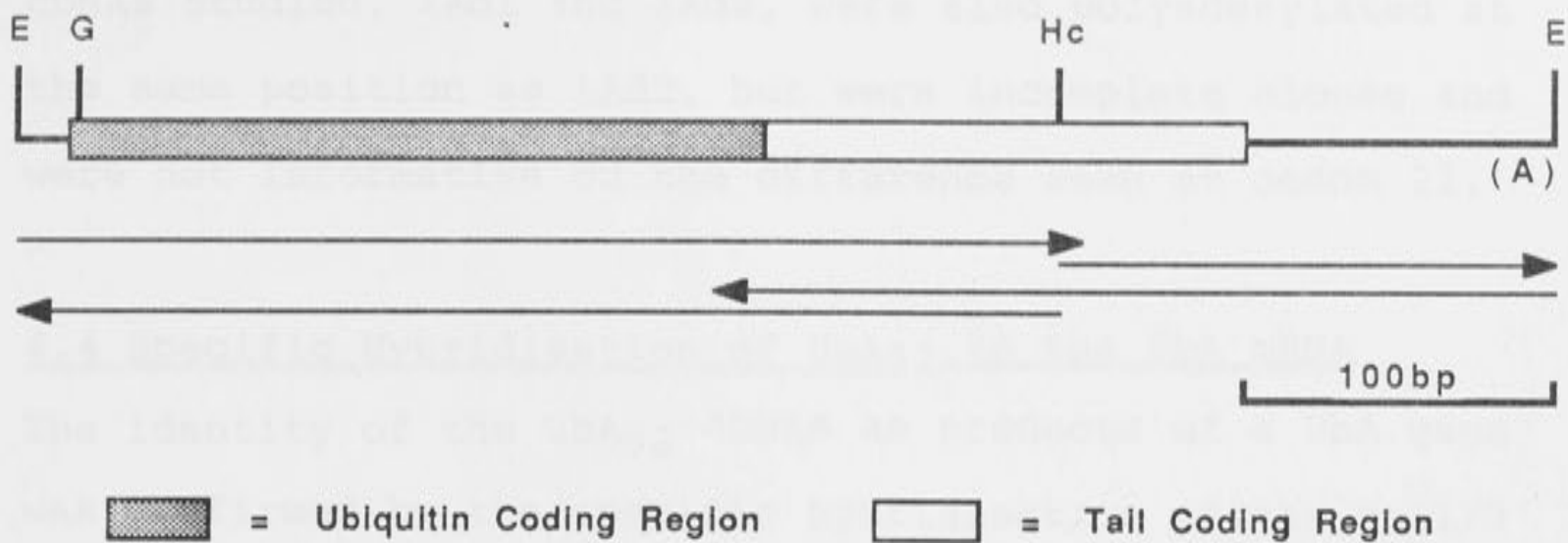
GTCCGGGAGCCTGGAGCCATTGCCACTAGGTGAGCTGTCCACAGGACCCTGGGCCGAGCT
 10 20 30 40 50 60
 MET Q I F V K T L T G K T I T L E V
 GACGCAAACATGCAGATCTTTGTGAAGACCCTCACTGGCAAACCATCACCCCTTGAGGTC
 70 80 90 100 110 120
 E P S D T I E N V K A K I Q D K E G I P
 GAGCCCAGTGACACTATTGAGAATGTCAAAGCCAAAATTCAAGACAAGGAGGGTATCCCA
 130 140 150 160 170 180
 P D Q Q R L I F A G K Q L E D G R T L S
 CCTGACCAGCAGCGTCTGATATTTGCCGGCAAACAGCTGGAGGATGGCCGCACTCTCTCA
 190 200 210 220 230 240
 D Y N I Q K E S T L H L V L R L R G G I
 GACTACAACATCCAGAAAGAGTCCACCCTGCACCTGGTGTGCGCCTGCGAGGTGGCATT
 250 260 270 280 290 300
 I E P S L R Q L A Q K Y N C D K MET I C R
 ATTGAGCCTTCTCTCCGCCAGCTTGCCCAGAAATACAACCTGCGACAAGATGATCTGCCGC
 310 320 330 340 350 360
 K C Y A R L H P R A V N C R K K K C G H
 AAGTGCTATGCTCGCCTTCACCCTCGTGCTGTCAACTGCCGCAAGAAGAAGTGTGGTCAC
 370 380 390 400 410 420
 T N N L R P K K K V K *
 ACCAACAACCTGCGTCCCAAGAAGAAGGTCAAATAAGGTTGTTCTTTCCTTGAAGGGCAG
 430 440 450 460 470 480
 CCTCCTGCCAGGCCCGTGGCCCTGGAGCCTCAATAAAGTGTCCCTTTCATTGACTGGA
 490 500 510 520 530 540
 GCAGCAAAAAAAAAAAAAAAAAA
 550 560

and Brownlee, 1976). It was subsequently learned that this library was constructed from cDNAs of greater than 800bp (Clontech, personal communication). Thus a DNA complementary to the UbA₅₂ mRNA of approximately 600-650nt could only be present as a cloning artefact such as has occurred in λ P15, presumably arising from a fusion of the two cDNAs during the ligation of EcoRI linkers during library construction. Thus the observed low frequency of 1 in 250,000 clones may not be representative of the relative abundance of UbA₅₂ mRNA in the placenta. A 240bp AluI/EcoRI tail-specific fragment coding for aa 9 through 52 of the tail protein, plus the 3' NCR and poly(A) tail (Figure 4.2.1) was used to screen an adrenal gland cDNA library in an attempt to isolate fuller-length UbA₅₂ cDNAs.

4.3 Isolation and Sequence Analysis of Human Adrenal Gland UbA₅₂ cDNA Clones.

The UbA₅₂ tail-specific probe derived from λ P15 (Figure 4.2.1) hybridised to approximately 500 clones of 250,000 plaques from a human adrenal gland cDNA library. Of 10 selected rescreened clones, three were chosen for sequence analysis. The clone with the longest insert, λ Ad2, contained a UbA₅₂ cDNA containing only 9bp of 5' NCR and differed from λ P15 at two positions (Figure 4.3.2). The first difference was a silent change in the 22nd codon of the ubiquitin coding region: threonine was encoded by ACT in λ P15, and by ACC in λ Ad2. This difference could reflect either allelic variation or an error of cDNA synthesis and cloning. The second difference was the site of polyadenylation: λ Ad2 was

λAd2



MET Q I F V K T L T G K T I T L E V
 GACGCAAACATGCAGATCTTTGTGAAGACCCTCACTGGCAAACCATCACCCCTTGAGGTC
 10 20 30 40 50 60

E P S D T I E N V K A K I Q D K E G I P
 GAGCCCAGTGACACCATTGAGAATGTCAAAGCCAAAATTCAAGACAAGGAGGGTATCCCA
 70 80 90 100 110 120

P D Q Q R L I F A G K Q L E D G R T L S
 CCTGACCAGCAGCGTCTGATATTTGCCGGCAAACAGCTGGAGGATGGCCGCACTCTCTCA
 130 140 150 160 170 180

D Y N I Q K E S T L H L V L R L R G G I
 GACTACAACATCCAGAAAGAGTCCACCCTGCACCTGGTGTGCGCCTGCGAGGTGGCATT
 190 200 210 220 230 240

I E P S L R Q L A Q K Y N C D K MET I C R
 ATTGAGCCTTCTCTCCGCCAGCTTGCCCAGAAATACAACACTGCGACAAGATGATCTGCCGC
 250 260 270 280 290 300

K C Y A R L H P R A V N C R K K K C G H
 AAGTGCTATGCTCGCCTTCACCCTCGTGCTGTCAACTGCCGCAAGAAGAAGTGTGGTCAC
 310 320 330 340 350 360

T N N L R P K K K V K *
 ACCAACAACCTGCGTCCCAAGAAGAAGGTCAAATAAGGTTGTTCTTTCCTTGAAGGGCAG
 370 380 390 400 410 420

CCTCCTGCCCAGGCCCGTGGCCCTGGAGCCTCAATAAAGTGTCCCTTTCATTGACTGGA
 430 440 450 460 470 480

AAAAAAAAAAAAAA
 490

Figure 4.3.1: (TOP) Restriction Map and Sequencing Strategy of λAd2.

Figure 4.3.2: (BOTTOM) Nucleotide Sequence of the λAd2 EcoRI Insert.

Initiation points of Ad1 and Ad8 are nt 239 and 166 respectively. See Figure 4.2.1. and 4.2.2 legends for details (previous Figure).

polyadenylated 6bp earlier than λ P15. The other two adrenal cDNAs studied, λ Ad1 and λ Ad8, were also polyadenylated at the same position as λ Ad2, but were incomplete clones and were not informative on the difference seen at codon 22.

4.4 Specific Hybridisation of UbA₅₂ to the UbA mRNA

The identity of the UbA₅₂ cDNAs as products of a UbA gene was confirmed by the specific hybridisation of the tail/3' NCR probe (Chapter 4.2) to the previously identified UbA mRNA species (Wiborg *et al*, 1985). Northern analysis (Chapter 2.2.13) was performed on RNA isolated from human placenta, lymphocytes prepared from whole blood and from a transformed lymphocyte cell line (Chapter 2.2.12). A coding region probe derived from the UbB pRBL26 cDNA clone (Chapter 3.2) was used to identify all ubiquitin mRNAs (Figure 4.4.1). A parallel northern blot was hybridised with the tail/3' NCR probe from λ P15 (Chapter 4.2), which specifically identified the UbA mRNA species of ~650nt (Figure 4.4.1). However, Lund *et al* (1985) had previously observed specific hybridisation of the ubiquitin/80aa tail fusion cDNA to the UbA mRNA. It thus appears that the UbA species represents two different co-migrating mRNAs. Hence, it is proposed that the ubiquitin/52aa tail fusion species represented by λ P15 and λ Ad2 be termed UbA₅₂, and the 80aa tail fusion species described by Lund *et al* (1985) would become UbA₈₀.

4.5 Isolation of UbA₅₂-Homologous Genomic Clones.

The λ P15 UbA₅₂ tail-specific probe (Figure 4.2.1) was also used to screen a human genomic library to isolate homologous

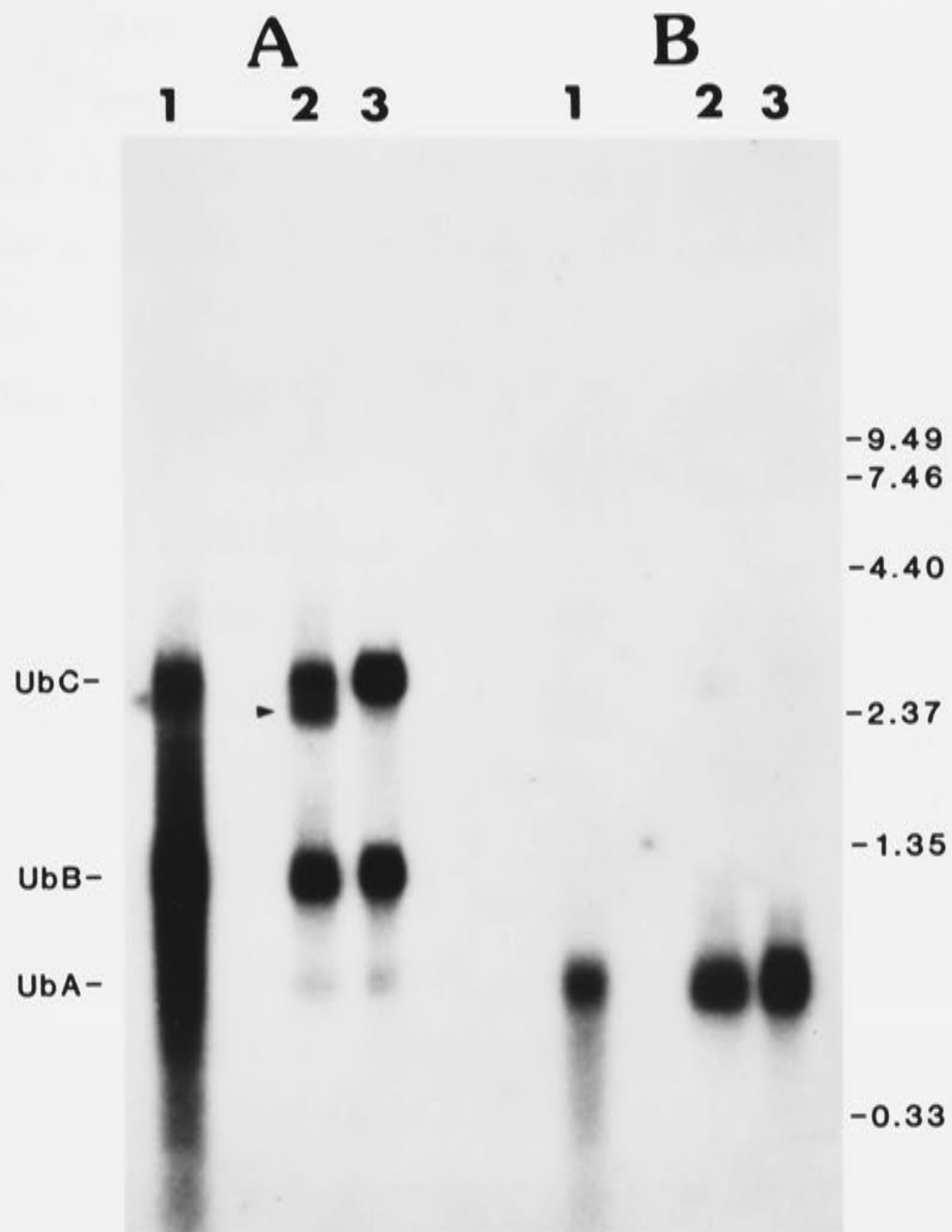


Figure 4.4.1: Northern Analysis with a UbA₅₂ Specific Probe.

Total RNA from human placenta (lane 1), freshly prepared lymphocytes (lane 2) and cultured lymphocytes (lane 3) was glyoxylated, electrophoresed, transferred to a nylon membrane and hybridised with a ubiquitin coding unit probe (Panel A) or a UbA₅₂ tail/3' NCR probe (Panel B). See Chapter 4.4 for details. Hybridising species are identified UbA, UbB and UbC after Wiborg *et al* (1985). The placental sample is partially degraded but obviously exhibits the UbA₅₂ species. Size markers at the right are in thousands of nucleotides. The individual in lane 2 is polymorphic at the UbC locus (arrowed - see also Figure 3.6.5, lane 1) and is discussed further in Chapter 5.

genomic sequences. Screening of 300,000 clones resulted in 4 repeatedly positive phage. However, restriction analysis reduced these to two unique clones, λ UA1 and λ UA4. Restriction maps of these clones are shown in Figures 4.6.1 and 4.7.1. Hybridisation analysis of the restricted clones localised all ubiquitin-coding and tail-specific hybridisation to a 5.5kb SalI/HindIII fragment of λ UA1, and to a 1.7kb HindIII fragment of λ UA4 (not shown). These fragments were subcloned into pUC18 to yield pUA1.1 and pUA4.3 respectively, and were characterised by sequence analysis as described below.

4.6 Sequence Analysis of a UbA₅₂ Processed Pseudogene, λ UA4.

Approximately 950bp of the pUA4.3 insert was sequenced by the strategy outlined in Figure 4.6.1 to reveal a UbA₅₂ processed pseudogene (Figure 4.6.2). A 504bp sequence homologous to the UbA₅₂ cDNA clones, including an 11bp poly(A) tail, was flanked by a 13bp direct repeat, features characteristic of processed pseudogenes. The position of the poly(A) tail corresponds to the site of polyadenylation of the adrenal cDNA, λ Ad2 (Chapter 4.3). The protein encoded by this pseudogene exhibits 84% similarity to the cDNA-encoded protein, but is severely disrupted by two nonsense codons and deletions of 4 and 3bp within the tail-like coding region (Figures 4.6.2 and 4.8.1). There are 30bp between the upstream direct repeat and the initiation codon, which are presumably representative of the 5' NCR. However, this region contains a 3bp insertion when compared to the 5' flanks of λ P15 and EHD5 (Figure 4.6.2). Thus λ UA4 extends

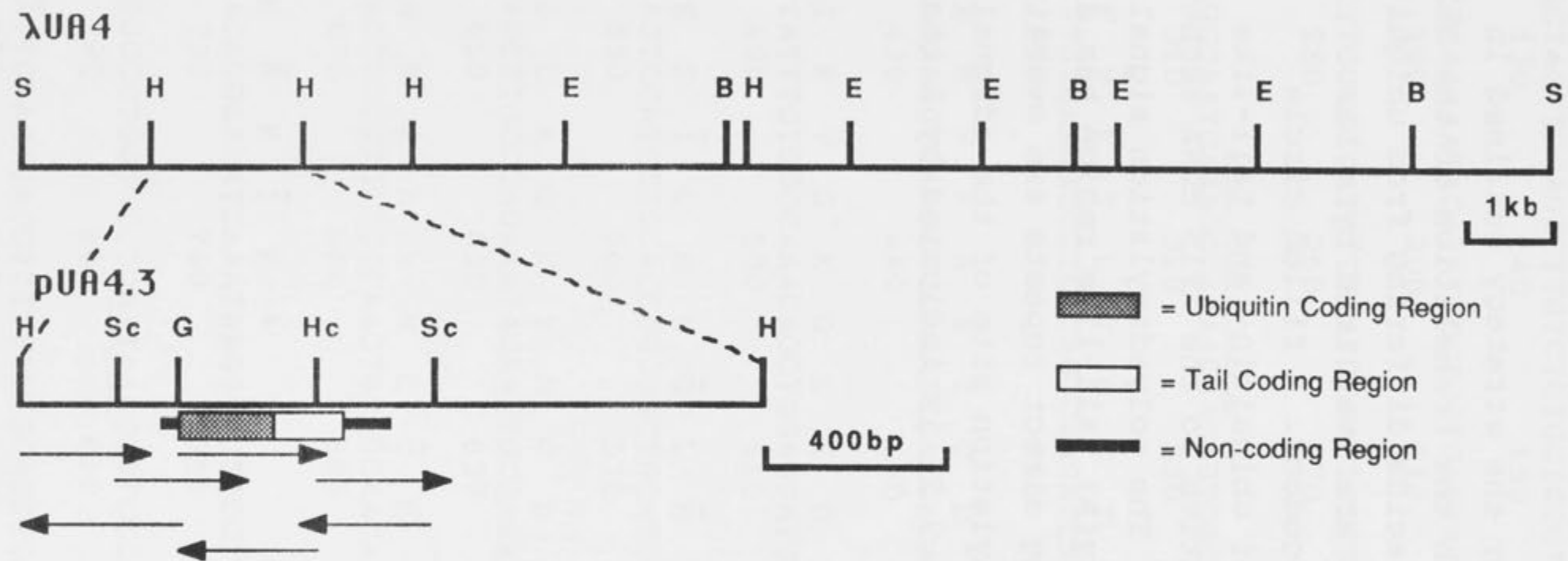


Figure 4.6.1: Restriction Map and Sequencing Strategy of UbA₅₂ Genomic Clone Lambda UA4.

Top Line: Sall insert of Lambda UA4.

Lower line: 1.7kb HindIII subclone pUA4.3. Boxes represent coding and non-coding regions as indicated. Arrows represent the direction and extent of sequencing. B:BamHI, E:EcoRI, G:BglII, H:HindIII, Hc:HincII, S:Sall, Sc:Sacl.

Figure 4.6.2: Nucleotide Sequence of the λ UA4 UbA₅₂-like Region.

The sequence determined by the strategy outlined in Figure 4.6.1 is given with the translation of the UbA₅₂-like region above. Amino acids differing from ubiquitin and the 52aa tail protein are overlined by a bar. Asterisks represent stop codons. A filled circle represents the junction of ubiquitin- and tail-like sequences. Deletions relative to the λ P15 cDNA sequence (Figure 4.2.2) are boxed. The polyadenylation signal is underlined, while the poly(A) tail-like region has a dashed underline. Flanking direct repeats are overlined with arrows. The polyadenylation site of the adrenal gland cDNA λ Ad2 (Figure 4.3.2) is indicated by A beside a vertical arrowhead.

AAGCTTTTGATCTTCTGCATACTTGGATCTGGTTTCCTGCAAAGCCATGTTGCCACATT
10 20 30 40 50 60
GGTGCCAATGACTAAAGCTGAGTTTTTTTAAAGGACAGCACTGGCTTAAATTCTTGTGCA
70 80 90 100 110 120
TTCAGACTAGTTGCAGTCTTGTCTCTGCTCCCTTCCCCTATGAAGAGTTTGGGGAAGAAA
130 140 150 160 170 180
CATCATGAAAAATACATATTCCTCCTGAGTACCTGATCCAGAGCTCAGACATTTTTGGA
190 200 210 220 230 240
AGTGACTTGGCCTGATCTTTTATCTAACTGTCTAATTTAGAAAACCACAGATTGTCTGAA
250 260 270 280 290 300
CTTAAAGGGAATCAAAGGCAAACAAGAGTGAGGCGGCTGAGCTACAGACGCAGAGATGC MET
310 320 330 340 350 360
Q I F V K T L T G K T I T L E V E P S D
AGATCTTTGTGAAGACCCCTCACAGGCAAGACCATCACCCCTCGAGGTGAGCCAGTGACA
370 380 390 400 410 420
T T̄ E N V K A Ē I Q D K E G F̄ P L̄ D Q Q
CCACTGAGAATGTCAAGGCCGAAATTCAAGACAAGGAGGGCTTTCCACTTGACCAGCAAT
430 440 450 460 470 480
C̄ L I F V̄ G K Q L E D G R T L S D S̄ N I
GTCTGATATTTGTGGGCAAACAGCTGGAGGATGGCCGTACTCTCTCAGACTCCAACATCC
490 500 510 520 530 540
Q K E S Ī L H L Ḡ L H̄ L F̄ G G I I E P S
AGAAAGAGTCCATCCTGCACCTGGGGCTGCACCTGTGAGGTGGCATTATTGAGCCCTCTE
550 560 570 580 590 600
? ? Q L A Q K Y N C D K MET I C R K C C̄ A
---GCCAGCTTGCCCAGAAATACTGCGACAAGATGATCTGCCGCAAGTGCTGTGCTT
610 620 630 640 650 660
C̄ L H P R A V N C C̄ K K ? F̄ G H T N N L
GCCTGCACCCCGTGTCTGCAACTGCTGCAAGAAG---TGAGGCCACACCAACAACCTGT
670 680 690 700 710 720
C̄ P K R̄ K Ī K *
GCCCCAAGAGGAAGATCAAATAAGGCTCTTCCTTCCCCCAAGGGCAGCCTCCTGCCAGG
730 740 750 760 770 780
CCCCATGGCCCTGGGGCCTCAATAAAGTGTCCCTTTTCATTGACTGGACAAAAAAGAC
790 800 810 820 830 840
AAAACAATTCAACACACCTTTTTTTTTTGCAGTTGATGAAACAGAACTAGAGAGGCTAAGTG
850 860 870 880 890 900
GCTCACCCAGTCAATACTGAATTGGTGACAGAACAACGGCTAAGCCCAGAGCTC
910 920 930 940 950

information on the UbA₅₂ 5' non-coding region a further 11bp upstream than the placental cDNA λ P15 (Chapter 4.2).

4.7 Characterisation of a Human UbA₅₂ Gene

4.7.1 Sequence Analysis of a UbA₅₂ Gene, λ UA1

Hybridisation analysis of restricted λ UA1 and its subclone pUA1.1 indicated that the UbA₅₂ cDNA-like regions were spread over more than 2kb of DNA (not shown), suggesting the presence of discontinuities (introns) within the gene. Figure 4.7.2 shows the sequence of 4.6kb of the pUA1.1 insert, determined by the strategy shown in Figure 4.7.1. The first 48bp of the pU1.1 insert appear to have arisen as a cloning artefact, as their sequence (determined by plasmid sequencing) contains restriction enzyme recognition sites in the order: SalI/SalI/BglII/SalI/PstI/HincII/BamHI (Figure 4.7.2, nt 1 to 48). This is most likely artefactual for the following reasons: First, the pU1.1 insert was cloned as a SalI/HindIII fragment and thus should contain only one SalI site. Second, such a high concentration of sites occurring naturally is statistically improbable. Third, most of the sites and their relative location is reminiscent of the pUC polylinker. Thus, the first base of the genomic insert is hard to define. The SalI site of the pU1.1 fragment derives from the EMBL3 phage right arm (see Figure 4.7.1); therefore the fragment should begin with a SalI site followed by a fused BamHI/Sau3A site, which may regenerate the BamHI site (Frischauf *et al*, 1983). The most likely candidate sequence is GTCAACGGATCC (nt 37 to 48), with the SalI site mutated to

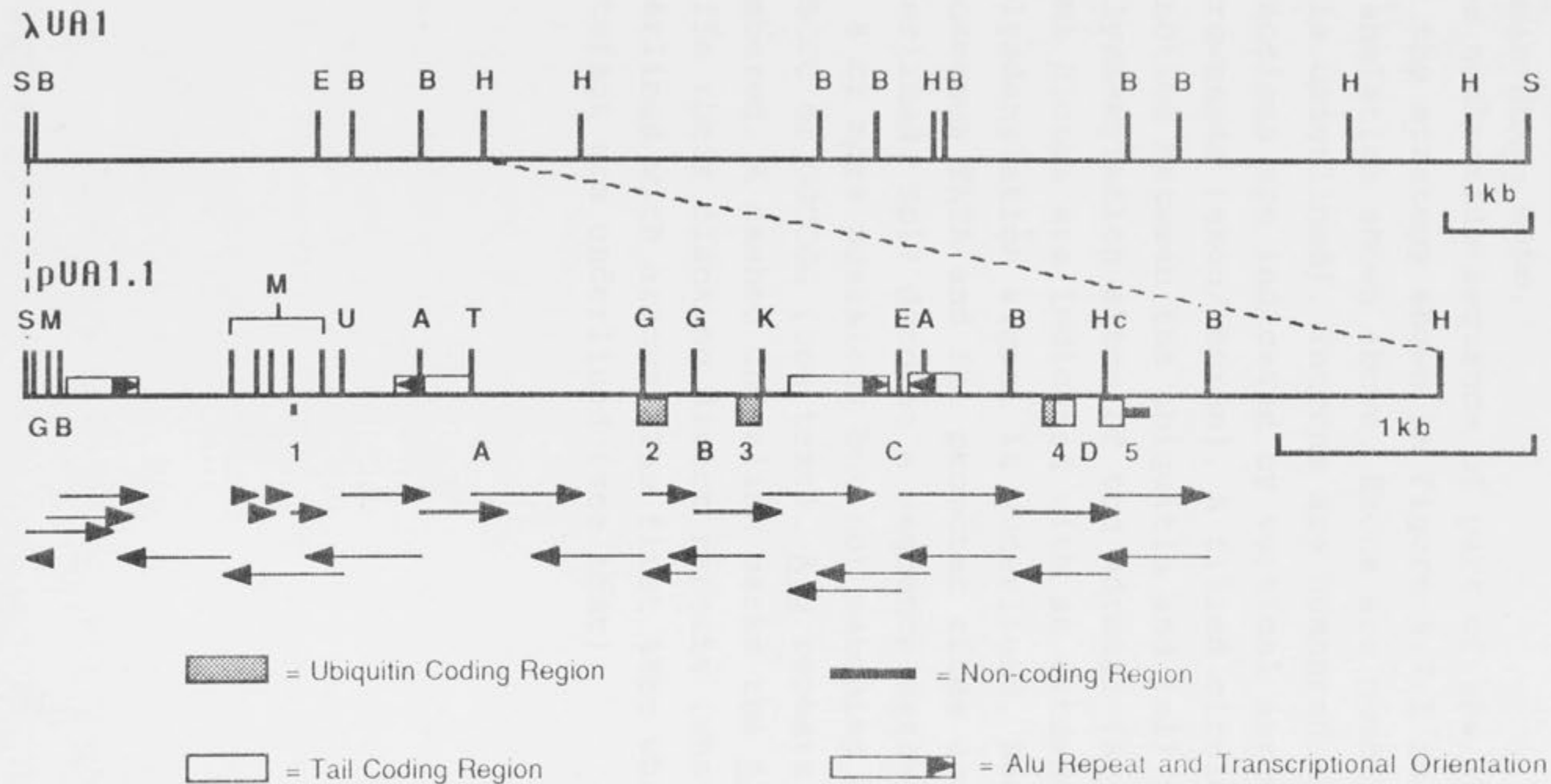


Figure 4.7.1: Restriction Map and Sequencing Strategy of UbA₅₂ Genomic Clone Lambda UA1.

Top Line: Sall insert of Lambda UA1.

Lower line: 4.5kb Sall/HindIII subclone pUA1.3. Boxes represent coding and non-coding regions and Alu repeats as indicated. Arrows represent the direction and extent of sequencing. Exons are numbered 1 to 5, introns are lettered A to D. A: SmaI, B: BamHI, E: EcoRI, G: BglII, H: HindIII, Hc: HincII, K: KpnI, M: MspI, S: Sall, T: TaqI, U: StuI.

Figure 4.7.2: (Following Pages) Nucleotide Sequence of a Human UbA₅₂ Gene.

The nucleotide sequence of part of the λ UAl insert determined by the strategy shown in Figure 4.7.1 is given with the translation shown above. Exons are numbered 1 through 5 (exon 1 is underlined). Introns are numbered A through D. Splice junctions are indicated by vertical arrows (intron/exon) and arrowheads (exon/intron). A filled circle represents the junction between the ubiquitin and tail protein regions. The polyadenylation sites of the adrenal (A) and placental (P) cDNA clones are indicated with an arrowhead. The polyadenylation signal is underlined. Sequences matching the consensus TATA and Sp1 promoter sites upstream of exon 1 are overlined. Sp1? denotes a sequence matching the Sp1 consensus at 8 or more positions but not matching the core sequence CCGCCC or GGGCGG (see text). Alu repeats are indicated and numbered. A dashed underline marks the Alu poly(A) tails, while their flanking direct repeats (where present) are overlined with arrows. The first 48bp which may be a cloning artefact are underlined (see text).

GTCGACCGGTGACCCAGATCTGGGTGACCTGCAGGTCAACGGATCCGCACATCTCGGCCTCCCAAAGTGCAGGCGTGAGTCACCGGAC
 10 20 30 40 50 60 70 80 90
 Sp 1 → Alu #1
 CCAGGTCCCGCCCTGGCACTTTTTAACCACCCACAAATCTGGATCTTACTGAAAAGAGACACTGCAGTGGCTCACGTCTGTAATCCCA
 100 110 120 130 140 150 160 170 180
 GCACTTTGGGAGGCCAAGGCGGGCGGATCACCTGAGGTGCGGAGTTTGGACCAGCCTGACCAACATGGAGAAACCCCGTCTCTACTAAA
 190 200 210 220 230 240 250 260 270
 AATACAAAAGTGGCCAGGCATGGTGTGCGCACACTTGTAAATCGCAGCTACTCGGGAGGCTGAGGCAGGAGAATTGCTTGAACCCAGGAGGC
 280 290 300 310 320 330 340 350 360
 GGAGGTTGCGGTGAGCCGAGATCGCGCCATTGCACTACAGCCTGGGCAACGAGAGCGAAACTCCGTCTCAAAAAAAAAAAAAAAAAAATC
 370 380 390 400 410 420 430 440 450
 CTGAGTCCCGCTTGACACCTTTTGTGAGGCACCACCACCTTTCTGGGCGAATGCGGTAGTACCGTCTGCTCTCCCTGCTGCTGCTCCTGAA
 460 470 480 490 500 510 520 530 540
 ATCCATTGAGGCACAGCGGCCGAGAGCTTTATAATAACCGATTCCAGGTGTTAGGTGCTTTCCAGCCCCGACTCCTGCGTCTGGACCC
 550 560 570 580 590 600 610 620 630
 TATA? Sp 1? Sp 1?
 GCAGTCTCTGCTTAATACCTTTGCTTTATTAGAAAACATTCTCCTTACTCCGTTTCCAGTATTGCTGAGGGCCCCCAACCGCCAGCG
 640 650 660 670 680 690 700 710 720
 Sp 1
 GTTGTCAATGGCCTAGAGGCAGCGGACGCAAAACACGGGGAGAGGTGCAATCGTCTCAAGTGACTCGGGCGGGGGCCACAACCGGAAG
 730 740 750 760 770 780 790 800 810
 Sp 1
 CGGGTGGGCGACCTTACCCACGTGCGCTGCGGCTTCGTTCCGAGCATCCAAGATGGCGGCAGGGCGGGGCCAAGGCGGGCGCGGAAT
 820 830 840 850 860 870 880 890 900
 TGTGACGCAGGCGTCCGGCGTCTCCGTCGCAAGCGCTTTCCGGCGGCGATTAGGTGGTTTCCGGTCCGCTATCTTTCTTTCTTCAGCG
 910 920 930 940 950 960 970 980 990
 exon 1 →
 AGGCGGCCGAGCTGGTTGGTGGCGGGCGGTGCGGGTTCCGCGCCGGGCCGAGAGCGGGTTGGGGCTGCGGGAGGCTGCAGGGGCCTGG
 1000 1010 1020 1030 1040 1050 1060 1070 1080
 intron A →
 GCGGCAGAAGAGGCGGCCCTGAGCTGGCTCATGCGGGCCAGTCTCGGCAGGGTGGCTGGGCAGGGCTCGCGAGGCCACGGCTCGGAGCCC
 1090 1100 1110 1120 1130 1140 1150 1160 1170
 AGACCGGGGCCAGGAGCGAACGCCGTTTTGGAGAGGAGCCTGCCTGCTCTGCCTGCCAGCGTGACCCACGAGGCCTCGGGCGGGAAGA
 1180 1190 1200 1210 1220 1230 1240 1250 1260
 GGTCTCGGGGCAGATCCGAGTTAATGAGAGAGGGGTATTGAGCGTGTAGCGTTAACTCTGCCAGTCACTGCGTCAGTCGCTTTGAAAT
 1270 1280 1290 1300 1310 1320 1330 1340 1350
 ACTAAATTTCTCGAGCTGAGTCTTCATACCTGGCTCCTAATCTACGTCTGTAAGGAGGAGCTGGTGGTAGTGTCTGCTTTTTAGACTTTT
 1360 1370 1380 1390 1400 1410 1420 1430 1440
 CTTTAGACTATTTGTAATTTTTTTCAGATGGAGTCTTGCTCTGTCGCTAAGCTGGAGTTTCACTGGTGGTCTCGGCTCACTGCAATCTC
 1450 1460 1470 1480 1490 1500 1510 1520 1530
 CACCTCCCGGGCTCGAGCGATTCTTCTGCCTGAGCCTCCCGAGTAGCTGGGATTATAGGCGCCTGGCACCACGCCAGTTGATTTTTGTA
 1540 1550 1560 1570 1580 1590 1600 1610 1620
 GTTTTAGTAGAGACGGAGTTTACCATGTTAGCCAGGCTCATCTTGAACCTCTGACCTCAAATGATCCGTCTGCCTCGGCCTTCCAAAGT
 1630 1640 1650 1660 1670 1680 1690 1700 1710
 Alu #2 →
 GCTGGGATTACAGGCATGAGCCCCTGCGCCCGGTCGATTCTTTGCTTTTTAAGTCAACTTTTATATGTGAACAATGCTTGGCAGGTGGT
 1720 1730 1740 1750 1760 1770 1780 1790 1800
 TGGTAGATACTAAGTGATGTTCTGTTGGTGGTGGGTCAGGCAAGAAGTGGGGTCTGGAGAGTTTTGGTGTAAATTGAGAAGGAAGCTAAGAG
 1810 1820 1830 1840 1850 1860 1870 1880 1890
 TGTTGGGTGCTCCAGCTTGGAGTTAGAGAGGAGAGAGGCTGCCACAGGAAGACATGTGTGTTGTAGGGGATGGCTTCCCATCCAGGCTGG
 1900 1910 1920 1930 1940 1950 1960 1970 1980
 CAGCAGGAGCAGCCTGTGCAGATCAGGACCTTGCTCCCTGGAAGAGGGTGGACCGCCTTCAGGGAAGATGGATCTAGCAAGATGATGCCA
 1990 2000 2010 2020 2030 2040 2050 2060 2070
 AAGGGTACTTATTCCATCAGGAGATACTGACGAGTCTTCCGCCGCTAAACCTAAGGTGAATAACCACAGTCTGTGTTCTGAAGAGCAC
 2080 2090 2100 2110 2120 2130 2140 2150 2160
 CCGTCCGGTCCAGGAGGTTGAGGACATGTGATCTTAGTTCCAGGACATGTTTAGACTACAGGCCAGGGTGTGTGAGAAGCCTAGCAGGGC
 2170 2180 2190 2200 2210 2220 2230 2240 2250
 CAGGCTTGGAGGAGTGAAGGAAGACAGGTAAGTGGGGCAGGACAGTTGGACTTGGTGCAGGCAAAGGGATAGCAACTGTGGTGTAGGCA
 2260 2270 2280 2290 2300 2310 2320 2330 2340
 exon 2 → MET Q I F V K
 CCTGAGCTTGTGCTACTCAGGCATGCATTGCTCACCAGTCTATCCTGCCGCCATCCTCCTCAGACGCAAACATGCAGATCTTTGTGAAG
 2350 2360 2370 2380 2390 2400 2410 2420 2430

T L T G K T I T L E V E P S D T I E N V K A K I Q D K E
 ACCCTCACTGGCAAACCATCACCCCTTGAGGTCGAGCCAGTGACACCATTTGAGAATGTCAAAGCCAAAATTCAGACAAGGAGGGTGTAG
 2440 2450 2460 2470 2480 2490 2500 2510 2520
 intron B→
 TAGGGCTGGGTGTGGGGCTCTGGCTGTGAAGTGGGAGTCCCTCTCTCGCCAGGGGAGTCTCAGTCTGTGGGTTGTGCTGACTTTA
 2530 2540 2550 2560 2570 2580 2590 2600 2610
 GATCTGTTTTGCCCTTGCTTCTCCATGTGATCTGAAGAACGTTTGTATCTTCTACCTCAGTTGGCCTTTTGTAGAAACTGGGGTAGTGC
 2620 2630 2640 2650 2660 2670 2680 2690 2700
 TGGAGCTCCCCTGCAGAGGACACTGCCAGTAATATGGTCCGCAGAGCCTCTAACTGAGCCTCCCCTCCCCCTCAGGTATCCCACCTGACCA
 2710 2720 2730 2740 2750 2760 2770 2780 2790
 G I P P D Q
 Q R L I F A G K Q L E D G R T L S D Y N I Q K
 GCAGCGTCTGATATTTGCCGGCAAACAGCTGGAGGATGGCCGCACTCTCTCAGACTACAACATCCAGAAAGGTACCGGGTTGGGGTTGC
 2800 2810 2820 2830 2840 2850 2860 2870 2880
 Intron C→
 TGGGCAGGGACCCAAGATCCCCAGGTCTTAGGAAAGGAGCATTGATGGCCTCAGGGGTTGGGGAGCAGTTCAAATGACTTGTGTTTTGTT
 2890 2900 2910 2920 2930 2940 2950 2960 2970
 TAAATAATGGGACTGGGCACAGTGGCTCATGCCTGTAATCCCGGCACCTTTGGGAGGCTTAGGCGGGTGGATCACCTGAGGTGAGGAGTTC
 2980 2990 3000 3010 3020 3030 3040 3050 3060
 AAGACCAGCCTGGACAACGTGGTGAATCCCGTTTCTATTAAAAATACAAAAATCAGCTGGGTGCAGTGGCTCAGGCCTGTAATCCCAGC
 3070 3080 3090 3100 3110 3120 3130 3140 3150
 Alu #3 2nd unit
 ACTTCGGGAGGCTGAGGCGGGCAGATCACAAAGTCAAGAGATTGAGATCATCATGACCAACATGGTGAATCCCATCTCTACTAAAAATA
 3160 3170 3180 3190 3200 3210 3220 3230 3240
 CAAAAATTAGCTAGGCATGGTGGTGCCTGCTAGTCCCAGCTACTCAGGAGGCTGAGGAAGGAGAATTGCTTGAAGTCCGGGAGACAAA
 3250 3260 3270 3280 3290 3300 3310 3320 3330
 AAAAAAAGTCATAATGTGAATTTTTTATCACTGCAATAAGGAAATTAGTGTCACTTGTGGGAGCGACAAGAATTCAGTGTCTTTTTT
 3340 3350 3360 3370 3380 3390 3400 3410 3420
 TGTGAGACAGAGTCTTACTCTGTACCCAGGCTGGAGTGCAGTGCAGCGATCTCACTGTGACCTCCGTCTCCCGGGTTCAAGCGATTCCC
 3430 3440 3450 3460 3470 3480 3490 3500 3510
 CTGCCTCAGCCTCCCGAGTAGCTGGGATTACAGGCACCCGCCACCAGCCAGCTAATTTTTTTTGTATTTTTTAGTAGAGACAGGGTTTC
 3520 3530 3540 3550 3560 3570 3580 3590 3600
 ACTACGTTGGCCAGGCTGGTCTCTTAAAGTGCTAGGATTACAGGCGTGAGCCATGGTCCCCCGCCTAGACTTCAGTGTCTGACCTTGCCT
 3610 3620 3630 3640 3650 3660 3670 3680 3690
 Alu #4
 GAACCACTTAGAGGTCGGCTTCCATGTTAGAAACCCAGATGGATGCCTCAGTTGGCATGTGTGAGTCTCAGACTCCCCCAGGGCTCGTG
 3700 3710 3720 3730 3740 3750 3760 3770 3780
 GTCAGTGTGAGATGGAGATTTCTGGGGCAGGCTGGCTGGGACAGTGTATCATCCACAGTAGAACGACGGCGGGGGATCCCGACTTGG
 3790 3800 3810 3820 3830 3840 3850 3860 3870
 TGTCCCATCACACTTGAGAAAGCAGCAGACTATAGGCCCTGGAGGGTCTGCCCTGTGACTGAGGAGCCAGGGCTGGGCTCAGTCGCC
 3880 3890 3900 3910 3920 3930 3940 3950 3960
 E S T L H L V L R L R G G I I E P S L R Q L A
 GTCCTTCTGGCTGTCTCCTGCAGAGTCCACCCTGCACCTGGTGTGGCCTGCGAGGTGGCATTATTGAGCCTTCTCTCCGCCAGCTTGC
 3970 3980 3990 4000 4010 4020 4030 4040 4050
 Intron D→
 Q K Y N C D K M E T I C R K
 CCAGAAATACAACCTGCGACAAGATGATCTGCCGCAAGTATGTGTGCTCCGATGCTTGGGGGGCTGTGGGGGGCTGCCGGAGTCCGGGTATG
 4060 4070 4080 4090 4100 4110 4120 4130 4140
 C Y A R L H P R A V N C R K K K C G H T
 CCCTCACCCACCCCTCCTGTCTCTGTGCAGGTGCTATGCTCGCCTTACCCTCGTGTCTCAACTGCCGCAAGAAGAAGTGTGGTACAC
 4150 4160 4170 4180 4190 4200 4210 4220 4230
 N N L R P K K K V K *
 CAACAACCTGCGTCCCAAGAAGAAGGTCAAATAAGGTGGTTCTTTCTTGAAGGGCAGCCTCCTGCCAGGCCCCGTGGCCCTGGAGCCT
 4240 4250 4260 4270 4280 4290 4300 4310 4320
 CAATAAAGTGTCCCTTTTCAATTGACTGGAGCAGCAATTGGTGTCTCATGGCTGATCTGTCCAGGGAGGTGGCTGAAGAGTGGGCATCTCC
 4330 4340 4350 4360 4370 4380 4390 4400 4410
 CTTAGGGACTCTACTCAGCACTCCATTCTGTGCCACCTGTGGGGTCTTCTGTCTAGATTCTGTACATCGGCATTGGTCCCTGCCCTAT
 4420 4430 4440 4450 4460 4470 4480 4490 4500
 GCCCCTGACTCTGGATTTGTATCTGTAAAACCTGGAGTAAAAACCTCAGTCGTGTAATTGGTGGGACTGAGGATCAGTTTTGTATTGCT
 4510 4520 4530 4540 4550 4560 4570 4580 4590
 GGGATCC

a HincII site as a consequence of the cloning artefact (underlined base).

Comparison of the λ UA1 sequence with the λ P15 cDNA, λ UA4 processed pseudogene and the EHD5 pseudogene revealed firstly that λ UA1 represents a bona fide UbA₅₂ gene, and secondly that the transcribed region is spread over at least 5 exons (Figure 4.7.2). Exon 1 most likely contains 16bp of the 5' NCR (Figure 4.7.2, nt 989 to 1004) and is described in more detail in Chapter 4.7.2. Exon 2 is 111bp in length, consisting of 8bp of 5' NCR and 103bp coding for aa 1 to 34.33 of ubiquitin. Exon 3 is 87bp long and codes for aa 34.33 to 63.33 of ubiquitin. Exon 4 is 103 bp long encoding aa 63.33 to 76 of ubiquitin plus aa 1 to 21.67 of the tail protein. Finally, Exon 5 is 178 or 184bp long, encoding aa 21.67 to 52 of the tail protein, the termination codon, and the 84 (adrenal) or 90 (placental) bp 3' NCR. Thus the coding region is spread over 4 exons of roughly equal coding capacity: 34.33, 29, 34.33, and 30.33 codons for exons 2 to 5 respectively (Figure 4.7.2). The protein encoded by λ UA1 is identical to those encoded by the placental and adrenal cDNA clones. The gene matches the adrenal cDNA sequence at ubiquitin codon 22 (ACC) rather than the placental ACT (Figure 4.7.2, nt 2460 to 2462). However, the gene differs from both cDNAs 4bp downstream of the stop codon (nt 4252): the gene has a G, while both cDNAs have a T. As mentioned previously, such differences may represent allelic variation or cDNA cloning errors.

The introns vary considerably in length. Intron A is 1400bp long, intron B is 259bp, intron C is 1122bp, and intron D is

84bp long. The exon/intron and intron/exon boundaries are shown in Table 4.7.1. All comply with the "GT-AG" rule and are reasonable matches to the consensus junction sequences (Breathnach and Chambon, 1981). In addition, a sequence matching the putative branch-point consensus can be found in each intron near the splice acceptor site (Keller and Noon, 1984). All 3 coding-region introns occur within, rather than between, codons: introns B and C between the first and second bases of a codon, and intron D between the second and third.

The 3' NCR region contains sequences in addition to the AATAAA polyadenylation signal (Proudfoot and Brownlee, 1976) that have been implicated in 3' end formation. The sequence CATTG (Figure 4.7.2, nt 4338 to 4342) matches the consensus element CAYTG, which may be involved in polyadenylation site selection (Berget, 1984). A second sequence, TGGTGTCCT (nt 4357 to 4365) is a reasonable match both in sequence and position to a consensus element YGTGTTY observed in most mammalian genes examined (McLauchlan *et al*, 1985). Similar sequences occur in the corresponding region of the UbB polyubiquitin gene (Chapter 3.6.1). Another consensus element of unknown significance, TTCAA (Urano *et al*, 1986), is absent from the UbA₅₂ polyadenylation region.

4.7.2 Characterisation of the UbA₅₂ Gene 5' Non-Coding Region

The placental UbA₅₂ cDNA clone λP15 contains 18bp between the placental lactogen hormone cDNA and the ubiquitin initiation codon, which are presumably UbA₅₂ 5' NCR (Figure

Table 4.7.1: UbA₅₂ Splice Junction Sequences

Intron	Splice Donor	Branch Point (a)	Splice Acceptor
CONSENSUS ^b	CAG/GTR	CTGAC(~28)	YYYYYYYYYYYYNYAG/GTR
?	?	?	TCTTCTTTTTCTTCAG/CGAGG
A	AGCTG/GTT	CTCAC(29)	CGCCCATCCTCCTCAG/ACGCA
B	34 GluG GGAGG/GTG	CTGAG(16)	36 lyIle CCTCCCTCCCCCTCAG/GTATC
C	63 LysG GAAAG/GTA	CTCAG(28)	65 luSer TGGCTGTCTCCTGCAG/AGTCC
D	21 ArgLy CGCAA/GTA	CTCAC(23)	23 sCys TCCTGTCTCTGTGCAG/GTGCT

Notes: (a) Distance in nt from the splice acceptor AG dinucleotide (Keller and Noon, 1984).

(b) Consensus sequence as given by Breathnach and Chambon (1981).

The translation and aa number are given above the exon sequences. Tail aa's are numbered 1 to 52.

4.2.2). The 8bp immediately upstream of the initiation codon are included in exon 2, while the first 10bp are not found until 1400bp upstream in exon 1. However, comparison with the two pseudogenes λ UBA4 and EHD5 suggests that exon 1 is 16bp in length as follows:

		Met
λ P15 cDNA:		GGCCGAGCTGACGCAAACATG
λ UA1 Gene: 973	TCTTCTTTTTCTTCAGCGAGGCGGCCGAGCTG	gttggtggcgg
EHD5: 213	AGCTCTTTCTCTTTAGTGAGGCAGCCAAGCTGACATAGAGATG	
λ UA4: 327	direct repeatGAGTGAGGCGGCTGAGCTGACGCAGAGATG	
		ACA

Notably, the exon (underlined) is preceded by a perfect splice acceptor consensus sequence, (Y)₁₄AG (Breathnach and Chambon, 1981), which is also evident in the EHD5 pseudogene. Whether this represents an *in vivo* splice acceptor site is not known, as cDNAs extending upstream of this region have not been isolated. However, the processed pseudogene λ UA4 contains 3bp between the end of the flanking direct repeat and the exon 1 sequence, GAG (see above). While λ UA4 is a pseudogene, and has already suffered a 3bp insertion in its 5' NCR (see above), these 3bp provide preliminary evidence for the existence of another exon further upstream of exon 1.

In order to obtain an estimation of the length of the UbA₅₂ mRNA 5' NCR, primer extension (Chapter 2.2.16) was performed on poly(A)+ RNA isolated from placenta and a cultured lymphocyte cell line (Chapter 2.2.12). The primer used was a 5'-[γ ³²P]-labelled 25mer being the non-coding strand of the HpaII/BglII fragment from the 5' end of the placental cDNA λ P15 (Figure 4.2.2, nt 55 to 79). This 25mer contains 16 bases of UbA₅₂ 5' NCR and 9bp of ubiquitin coding sequence and thus should be specific for UbA₅₂ transcripts. The

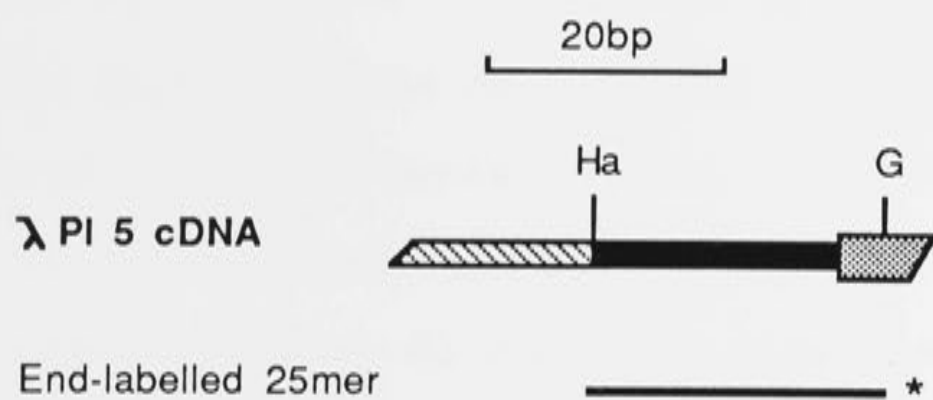
results of primer extension are shown in Figure 4.7.3. Primer extension of lymphocyte poly(A)⁺ RNA produced 45, 46, 48, 49 and 74nt species, while that of placental poly(A)⁺ RNA produced 59, 60 and 73nt species. The approximately similar length of the longest species most likely represents the full-length mRNAs, while shorter species may correspond to shorter mRNAs or to RNA secondary structures causing premature stoppages. This result suggests a 5' NCR of 64 or 65nt. As exons 1 and 2 only contain a total of 24bp of 5' non-coding sequence, then either exon 1 continues a further 40 or 41bp upstream, or another NCR exon of 40 or 41bp exists upstream of "exon 1". An extra 40 or 41bp added to the total length of the known exons (501bp) would give an unpolyadenylated mRNA length of 541 or 542 nucleotides, in agreement with the observed polyadenylated length of ~600 to 650nt (see Figure 4.4.1).

In an attempt to resolve this question, S1 nuclease mapping (Chapter 2.2) was performed using uniformly-labelled probes derived from AvaII/MspI and MspI subclones covering the region 625 to 1037 (Figure 4.7.2; "exon 1" is nt 989 to 1004). However, results were inconclusive, most likely due to the short lengths of the (putative) exons involved.

The most likely conclusion from the results described above is that another purely NCR exon of ~41bp exists upstream of the presently identified exon 1. If the 3bp between the flanking direct repeat and exon 1 sequence of processed pseudogene λ UA4 are representative of the UbA52 mRNA, then the upstream exon should end with GAG, plus the conserved GT

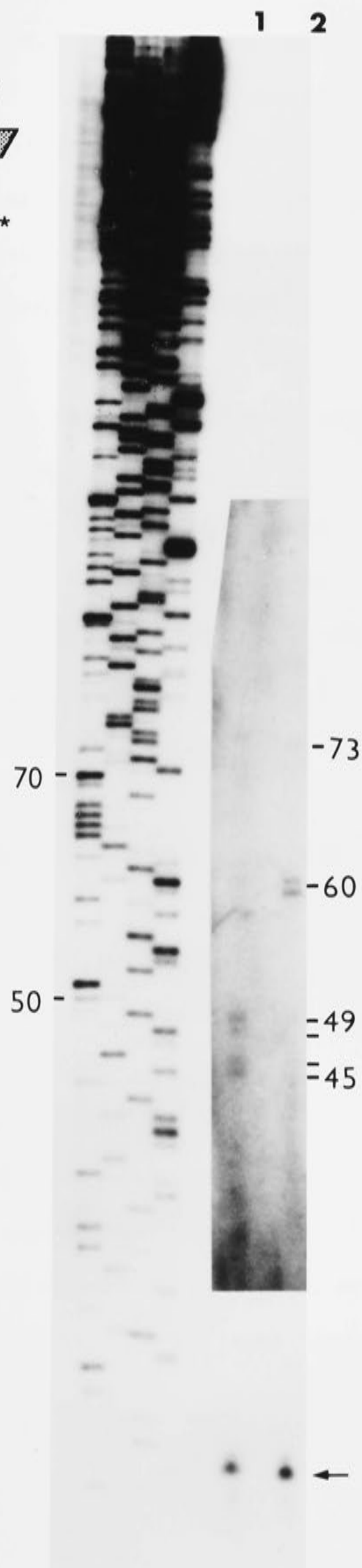
Figure 4.7.3: UbA₅₂ Primer Extension Analysis.

This Figure represents the results of the primer extension analysis described in Chapter 4.7.2. The left panel is a schematic summary of the experiment and shows the derivation of the 25mer primer from a HaeIII (Ha)/BglIII (G) digest of the λ P15 cDNA. The HaeII site is at the junction of the placental lactogen hormone (P.L.H.) and ubiquitin cDNAs. The right panel shows the primer extension products obtained with poly(A)⁺ RNA isolated from a human lymphocyte cell line (Lane 1) and placenta (Lane 2). The unextended 25mer primer is arrowed (bottom). Sizes are in nucleotides and are determined from the sequencing ladder.



Primer Extension Products (Lanes 1 and 2)

- = P.L.H. 5' Non-coding Region
- = Ubiquitin 5' Non-coding Region
- = Ubiquitin Coding Region



at the splice donor (Breathnach and Chambon, 1981), ie, GAGGT. This sequence occurs only once in the 988bp upstream of exon 1: 222 upstream (nt 761 to 766). However, no portion of the AvaII/MspI probe spanning this region (nt 625 to 807) was protected from S1 nuclease digestion by hybridisation to total lymphocyte RNA (not shown), and thus involvement of this region as a splice donor must be considered purely speculative.

A final possibility is that the putative upstream intron may be longer than 988bp, and thus the upstream exon would not be present within the genomic insert of λ UA1. In this case the upstream Alu repeat (Figure 4.7.1) would be within the upstream intron. In the absence of conclusive evidence about the nature of the putative upstream exon, the known 16bp 5' NCR exon will remain as exon 1.

4.7.3 Identification of Putative Promoter Elements Upstream of Exon 1

Several sequences possessing similarity to known promoter elements were identified in the 540bp region between the upstream Alu repeat and exon 1 (Figure 4.7.2).

The most upstream element is a TATA-like sequence TATAATA (nt 570 to 576). This sequence is 419bp upstream of exon 1 and 154bp upstream of a putative 41bp upstream exon (Chapter 4.7.2), clearly outside the consensus distance of \sim 31bp from the mRNA start site (Breathnach and Chambon, 1981), and thus its functionality is unlikely with respect to the currently characterised UbA₅₂ exons.

The second identified sequences are binding sites for the RNA polymerase II transcription factor Sp1. This promoter sequence, originally identified as the GC box upstream of the SV40 virus early promoter, has the expanded consensus G/TGGGCGGRRY which is functional in either orientation. (reviewed by Kadonaga *et al*, 1986). The λ UAl UbA₅₂ gene contains two good matches to the consensus: CGGGCGGGGC (nt 788 to 797) and AGGGCGGGGC (nt 873 to 882). In fact, these to Sp1 boxes are part of a larger direct repeat as follows:

```

785 CGGC-GGGCGGGGCCCA
869 CGGCAGGGCGGGGCCCA

```

In addition, the 19bp surrounding the first Sp1 box matches a 20bp Sp1-containing sequence in the first intron of the human α 1(I) collagen gene (Bornstein *et al*, 1987) as below:

```

 $\lambda$ UAl 781 GACTCGGC-GGGCGGGGCC
Collagen GACTCGGCAGGGCGGGGTCC

```

The most downstream Sp1 box is 106bp upstream of exon 1, but interestingly would be only 65bp upstream of a 41bp-extended exon 1 (see Chapter 4.7.2). Alternatively, the Sp1 boxes may lie within the putative upstream intron. Two other sequences which match the complement consensus Sp1 box at 8 of the 10 positions occur from nt 601 to 610 (TCCCAGCCCC) and nt 702 to 711 (GGCCCGCCAA). Neither of these match the core GC box (CCGCCC or GGGCGG, Kadonaga *et al*, 1986). A third complement consensus match (9 out of 10) occurs upstream of the Alu repeat from 95 to 104 (GTCCCGCCCT). Potential functions for these Sp1 boxes are discussed later (Chapter 4.10.2).

4.8 Comparison of UbA₅₂ Gene and Pseudogene Regions

4.8.1 Comparison of Coding and Non-Coding Regions

The ubiquitin-and tail-like coding units from the gene λ UA1 and the two pseudogenes λ UA4 and EHD5 are compared in Figure 4.8.1 both as DNA sequences and encoded proteins. The third coding unit from the 3 coding unit UbB polyubiquitin gene is also included in the DNA sequence comparison.

At the DNA level, the ubiquitin-like coding units of pseudogenes λ UA4 and EHD5 are respectively 90.4 and 86.8% similar to the gene ubiquitin coding unit, and 85.1% to each other. Notably, the UbA₅₂ coding unit is only 85.1% similar to the third UbB coding unit (Chapter 4.6.1) while both encode identical proteins. The proteins encoded by the λ UA4 and EHD5 pseudogene ubiquitin-like coding units are respectively 85.5 and 82.9% similar to ubiquitin itself, but both include a nonsense codon at the 74th aa as one of their pseudogene traits (Figure 4.8.1).

The tail-coding region of processed pseudogene λ UA4 is 94.0% similar to that of the gene, and has suffered deletions of 4 and 3bp. The λ UA4 tail-like protein is 80.8% similar to the gene-encoded tail, and includes a nonsense codon and the 4bp frame-shift deletion. Corresponding comparison of the EHD5 pseudogene tail-like region reveals that it does not possess an intact tail-like coding region. The first 86bp are 82.6% similar to the gene, while this figure drops to 15.7% for the remaining 70bp. Similarly, the first 29aa (87bp) are 75.9% similar, dropping to 4.3% for the last 23aa (Figure 4.8.1). This difference is not due to a frame shift deletion or insertion, as no other tail-homologous sequences were

Figure 4.8.1: Sequence Comparison of UbA₅₂ Gene and Pseudogenes.

All Panels: Sequences are named at the lower right with the percentage match to the first sequence listed where appropriate. Dots indicate identity with the top sequence. Dashes indicate gaps introduced to maximise alignment.

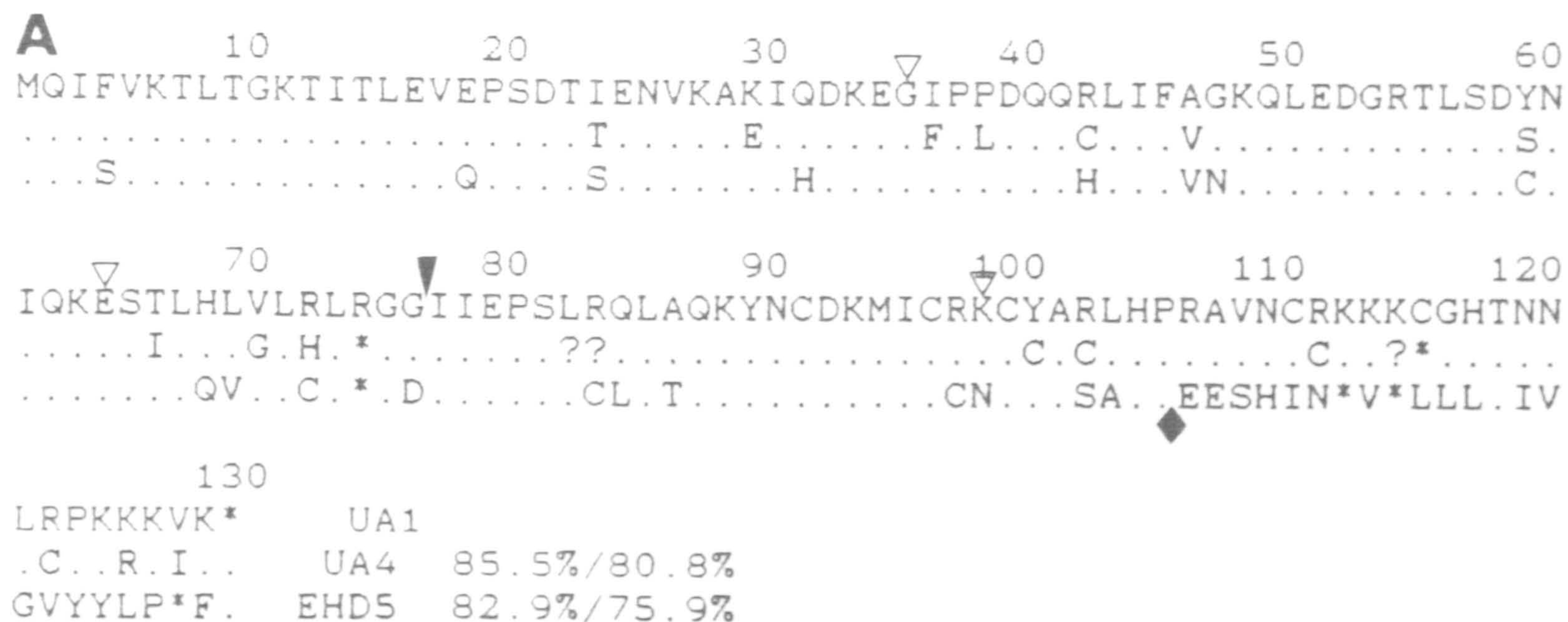
Panel A (below): Comparison of proteins encoded by λ UA1, λ UA4 and EHD5. A black arrowhead indicates the junction of the ubiquitin and tail proteins. Open triangles indicate the positions of introns in the UbA₅₂ gene. A black diamond indicates the breakpoint of EHD5 homology. Percentage similarities are given as ubiquitin portion/tail portion. Only the portion of EHD5 up until the homology breakpoint is included in this calculation.

Facing Page: DNA Sequence Comparisons.

Panel B: Ubiquitin coding region comparisons. The third UbB gene coding unit ("UbB") is also included.

Panel C: Tail coding region comparisons. Only the first 86bp of the EHD5 tail coding region were used in the percentage calculation.

Panel D: 3' NCR comparisons. EHD5 is omitted as it lacks a 3' NCR. The Adrenal cDNA λ Ad2 (Ad2cDNA) is included to show the G/T variation (nt 7) and the alternate polyadenylation site. The 3' NCR of the human leukocyte cDNA determined by Salvesen et al (1987) is also included ("leuk").



B

10	20	30	40	50	60
ATGCAGATCTTTGTGAAGACCCTCACTGGCAAACCATCACCCCTTGAGGTGAGCCCAGT					
.....A.....G.....C.....					
.....C...A.....A.....C...G...					
.....C.....G..C.....G.....T..G.....G.....					
70	80	90	100	110	120
GACACTATTGAGAATGTCAAAGCCAAAATTCAAGACAAGGAGGGTATCCCACCTGACCAG					
.....C.C.....G..G.....CT.T...T.....					
.....C.G...A.....G..C..C.....C.....					
.....C..C..A.....G..G.....G..C.....T..A..A..C.....C..C.....					
130	140	150	160	170	180
CAGCGTCTGATATTTGCCGGCAAACAGCTGGAGGATGGCCGCACTCTCTCAGACTACAAC					
..AT.....TG.....T.....C.....					
...A.....TGAA.....T.....T.....T.....GT...					
...A.G..C..C...A...G.....A.....T..T.....					
190	200	210	220	230	
ATCCAGAAAGAGTCCACCCTGCACCTGGTGTGCGCCTGCGAGGTGGC					
.....T.....G.C...A...T.....					
.....GG.....C.TT...CT.....A.					
.....G.....CC.....A.G.....					
				UA1	
				UA4	90.4%
				EHD5	86.8%
				UbB	85.1%

C

10	20	30	40	50	60
ATTATTGAGCCTTCTCTCCGCCAGCTTGCCCAGAAATACAACCTGCGACAAGATGATCTGC					
.....C...-----.....					
.....C...T...TA..CA.....					
70	80	90	100	110	120
CGCAAGTGCTATGCTCGCCTTCACCCTCGTGCTGTCAACTGCCGCAAGAAGAAGTGTGGT					
.....G...T...G.....C.....T.....---..A..C					
T...T..T.....A.TGCA..T..AGAG.AAAGTC.TATAAATT.AGT.TGA.TATTA					
130	140	150	160		
CACACCAACAACCTGCGTCCCAAGAAGAAGGTCAAA					
.....T.C.....G...A.....					
TTA..A.TAGTTGG.GTATATT.CCT.CCCTGATTC					
				UA1	
				UA4	94.0%
				EHD5	82.6% (86bp)

D

10	20	30	40	50	60
TAAGGTGGTTCTTTCCTTGAAGGGCAGCCTCCTGCCAGGCCCGTGGCCCTGGAGCCTC					
.....T.....					
.....CTC...C...CCC.....A.....G.....					
70	80	90			
AATAAAGTGTCCCTTTCATTGACTGGAGCAGC					
.....-----					
.....-----					
.....-.....					
				UA1	
				Ad2cDNA	
				UA4	90.7%
				Leuk	

observed in the 988bp sequenced downstream of this position (Figure 4.1.2). In addition, the UbA₅₂ tail/3' NCR probe (Figure 4.2.1) did not hybridise to any other region in the entire EHD5 genomic insert, other than those predicted from sequence analysis. Notably, the breakpoint of similarity does not correspond to an exon/intron junction, but is 21bp downstream of the exon 4/intron D junction (Figure 4.8.1). These observations raise questions as to which type of pseudogene EHD5 is. It clearly exhibits some features of processing, in that introns A, B, C and D have been removed. However, it still contains the putative splice acceptor site upstream of exon 1 (see Chapter 4.7.2) and has lost part of the tail-coding- and 3' NCRs. This pseudogene could have arisen in one of three ways from a pre-mRNA still containing the unspliced putative upstream intron. First, the full-length DNA copy was inserted but subsequently lost its 3' portion (and downstream direct repeat) through DNA deletion. Second, through partial duplication of a full-length pseudogene at another locus. Third, from a 3'-truncated mRNA, copied and inserted as a normal processed pseudogene. In this latter case, the sequence immediately downstream of the break in homology should also be present upstream as a flanking direct repeat. This is observed for the hexamer AGGAAA present at nt 39 to 44 and 555 to 560 (Figure 4.1.2), and also with a single transition mismatch at nt 180 to 185, AGAAAA, 15bp upstream of the 5' break in homology to the gene. Whether these short repeats are actual insertion-generated duplications is not known.

As EHD5 lacks a 3' NCR, only those of the gene and λ UA4 can be compared, with a similarity of 90.7%. Notably, λ UA4 is polyadenylated at the same site as the adrenal cDNA clones, 6bp upstream of the site used in the placental cDNA clone.

While this study was in progress, Salvesen et al (1987) reported the sequence of a partial human leukocyte cell line UbA₅₂ cDNA, encoding aa 40 through 76 of ubiquitin plus the 52aa tail and 3' NCR. The latter is included in Figure 4.8.1 and differs in four places from the gene and cDNAs isolated in this study. Two differences correspond to those observed above, the first being the G/T substitution 4bp downstream of the stop codon: the leukocyte cDNA matches the λ UA1 gene at this point. The second difference is the polyadenylation site: the leukocyte clone is polyadenylated at the same position as the placental cDNA, rather than the adrenal site. (Chapter 4.3). The other two differences involve DNA deletions not observed in the placental/adrenal cDNAs (Figure 4.8.1). However, a re-check of the reported leukocyte cDNA sequence suggests that these two differences may in fact be sequence reading errors (Dr Guy Salvesen, personal communication) and thus most likely do not reflect in vivo sequence variation.

The 5' NCRs have previously been detailed in Chapter 4.7.2. Over the known 5' NCR of 24bp, λ UA4 and EHD5 respectively share 80.0 and 70.8% similarity to the gene region.

4.8.2 Features of the UbA₅₂ Tail Protein

The human 52aa tail protein is very basic, with 16 Arg and Lys residues (30.8%). Similar observations about the tails of other fusion genes prompted suggestions of a nuclear role for the tail protein (Lund *et al*, 1985; Ozkaynak *et al*, 1987). The UbA₅₂ tail also contains 5 basic aa clustered in a 7aa region at its C-terminus (Arg-Pro-Lys-Lys-Lys-Val-Lys, Figure 4.8.2 below) which is similar to the nuclear location signal of the SV40 large T antigen (see Lund *et al*, 1985). A similar sequence is present in other tail proteins, prompting speculation as to its role in the nuclear localisation of ubiquitin (Lund *et al*, 1985; Ozkaynak *et al*, 1987).

Another domain identified in the tail proteins is a cysteine rich metal-binding, nucleic acid-binding domain. This domain was originally identified in the *Xenopus* transcription factor TFIIIA (Miller *et al*, 1985) and subsequently observed in several other proteins that have been implicated in nucleic acid binding (Berg, 1986; Harrison, 1986; reviewed by Klug and Rhodes, 1987). The UbA₅₂ tail protein contains a sequence that matches the consensus metal/nucleic acid-binding domain (Berg, 1986): Cys-X₂-Cys-X₁₀-Cys-X₄-Cys; X is any aa. Notably, the positions of these Cys residues are conserved between all known tail proteins (Figure 4.8.2 below).

Figure 4.8.2: Tail Protein Comparisons

```
Human: IIEPSLRQLAQKYNCDKMICRKCYARLHPRAVNCRKKKCGHTNNLRPKKKVK
Mouse: .....C.....
Slime: -.....VI..R..K.....S.....LLK
Yeast: .....KA..S.....SV.....P...T...R.....Q.....L.
```

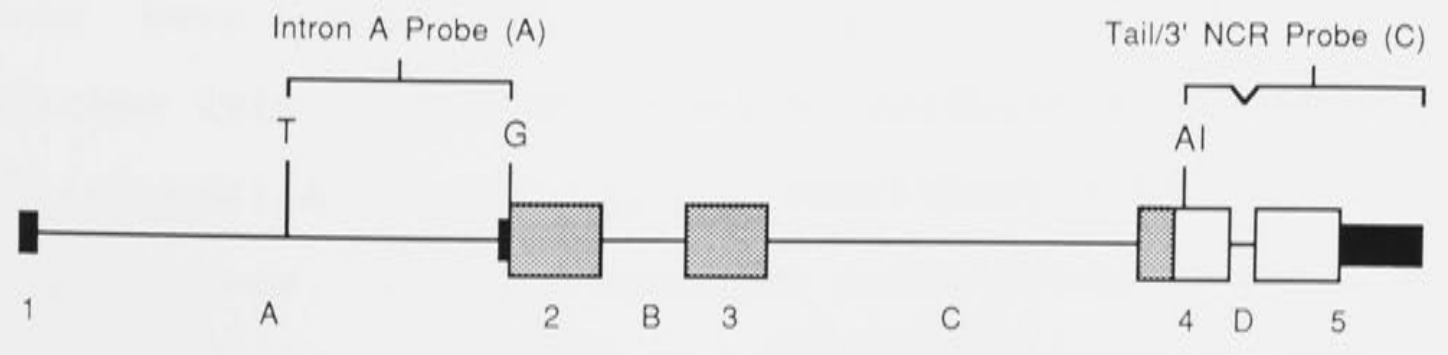
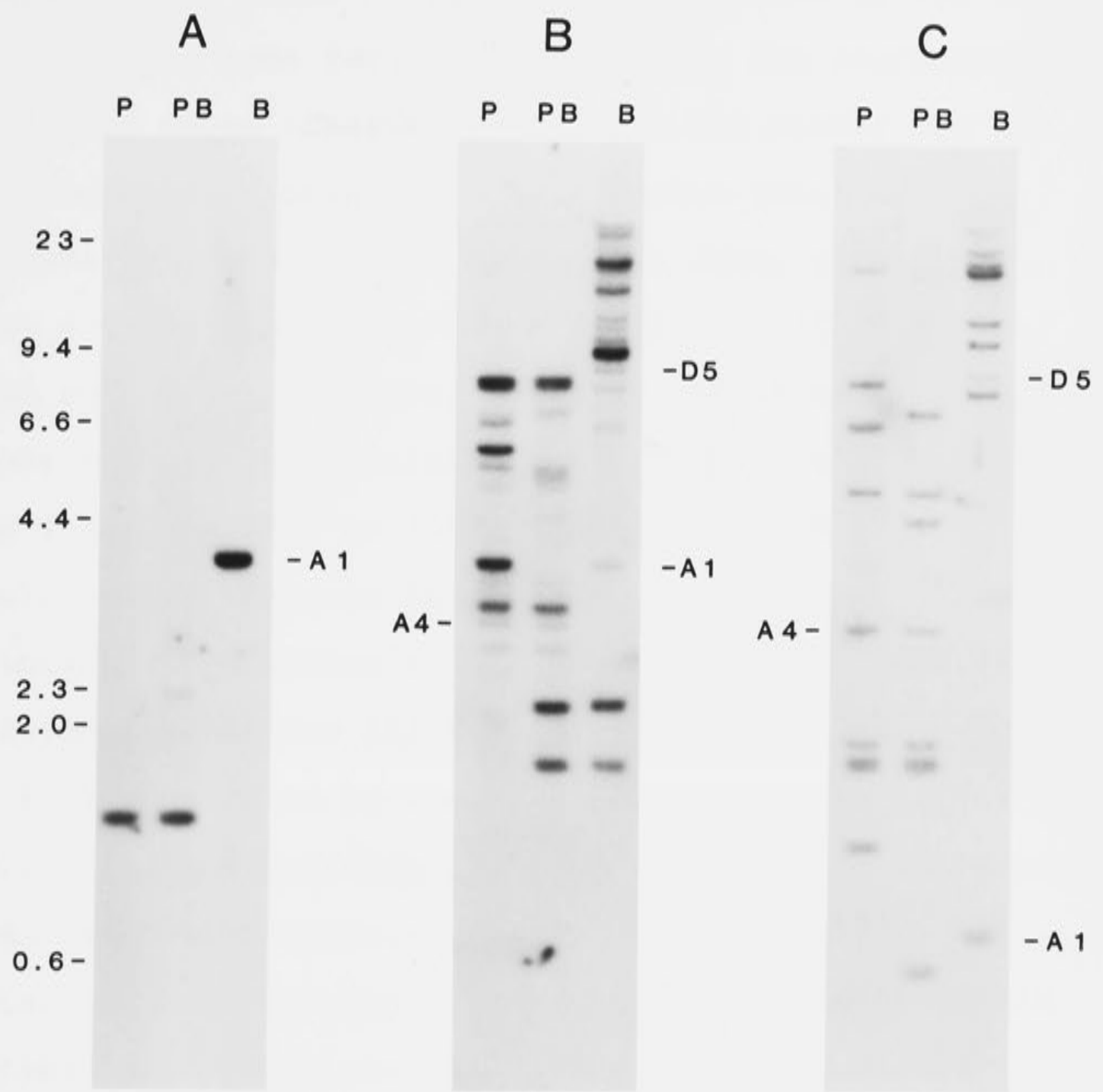
(See Figure 4.8.1 legend for symbols. Cys residues are underlined. See text for references).

While ubiquitin itself has been very strongly conserved in evolution (Chapter 1.1), the 52aa tail proteins are also conserved to a high degree. Figure 4.8.2 shows a comparison of known UbA₅₂-type tail proteins. The yeast (Ozkaynak *et al*, 1987) and slime mould (Westphal *et al*, 1986; as modified by Schlesinger and Bond, 1987) tail proteins respectively exhibit 81 and 83% similarity to the human tail protein, and 73% to each other. In addition, the first 18aa of a putative mouse tail protein deduced from a cDNA clone (St John *et al*, 1986; see Ozkaynak *et al*, 1987) are identical to the corresponding human residues. Such conservation over large evolutionary distances is indicative of a specific function(s) for the tail protein in its own right, rather than merely as a "carrier" protein to transport ubiquitin to the nucleus.

4.9 Hybridisation Analysis of Human Genomic DNA

Hybridisation analysis was performed on PstI, PstI/BamHI and BamHI digested genomic DNA as described in Chapter 3.8 but with probes derived from different regions of the UbA₅₂ gene and cDNAs. Neither enzyme cuts within known UbA₅₂ ubiquitin- or tail-coding regions with the exception of a PstI site near the 3' end of the EHD5 ubiquitin-like coding unit (Figure 4.1.1). Figure 4.9.1 shows the results of hybridisation with the following probes: (A) an intron A probe, being the 674bp TagI/BglII fragment immediately upstream of exon 2 (Figure 4.6.2, nt 1744 to 2401); (B) a coding region probe derived from the UbB gene as described in Chapter 3.2;

Figure 4.9.1: UbA₅₂ Genomic Hybridisation Analysis.
Human genomic DNA was digested with PstI (P), PstI/BamHI (PB) or BamHI (B). Probes used were: A UbA₅₂ intron A probe (panel A); a coding region probe derived from the UbB cDNA (Chapter 3.2) (panel B) and a UbA₅₂ tail/3' NCR probe (panel C). The derivation of the Panel A and C probes from the UbA₅₂ genes is represented schematically below the figure. Introns and exons are not drawn to the same scale in this diagram. The tail/3' NCR probe was derived from the λP15 cDNA (Figure 4.2.2) and thus does not contain intron D (caret). Some hybridising bands are assigned to the known UbA₅₂ subfamily members as described in the text. Size standards are on the left of Panel A and are in kb.



and (C) a tail/3' NCR probe derived from the placental UbA52 cDNA as described in Chapter 4.2.

The large number of hybridising fragments (HFs) produced by the coding region probe (Panel B) is indicative of the size of the ubiquitin gene family. The stronger HFs represent the polyubiquitin genes (Chapter 3.8) while the weaker HFs are most likely single coding unit/tail fusion genes and pseudogenes. The restriction map of the UbA52 gene λ UA1 predicts a 3.7kb BamHI HF (Figure 4.7.2, nt 131 to 3857) with the coding region probe, which is indicated A1 (Panel B, Figure 4.9.1). The adjoining 740bp BamHI fragment of λ UA1 (Figure 4.7.2, nt 3857 to 4592) contains the 39bp of ubiquitin coding sequence in exon 4, which is apparently insufficient to hybridise to the probe. However, this 740bp BamHI fragment contains all of the tail-coding and 3' NCRs, and is thus identified by the tail-specific probe (Panel C). The short PstI and PstI/BamHI HFs on Panel C also correspond to predicted λ UA1 fragments. The pseudogene EHD5 is predicted to produce a BamHI HF of 9kb with both ubiquitin coding and tail probes which is indicated D5 in Figure 4.9.1. Pseudogene λ UA4 should produce a BamHI HF of larger than 8.3kb and a PstI HF of 3.1kb with both probes, which is marked A4 (Panels B and C).

Notably, the intron A probe (Panel A) produces a unique HF of 3.7kb (BamHI) and 1.4kb (PstI), consistent with the λ UA1 sequence. However, the tail specific probe produces about 9 HFs with either enzyme (Panel C), indicating that the UbA52 subfamily is quite large. These results suggest that either the intron A region of duplicated UbA52 genes has diverged

considerably, or that all of the other Panel C HFs represent processed (ie. intronless) pseudogenes (see Chapter 4.10.4). Alternatively, some of these HFs could be other (non-ubiquitin) genes containing the same tail coding region. However, as most Panel C HFs correspond to a Panel B HF of the same size, presumably most are ubiquitin loci. The corresponding analyses with other λ UA1 introns were not performed, due to the presence of Alu repeats (intron C; Figure 4.7.1) or their small size and lack of "convenient" restriction sites (introns B and D).

4.10 DISCUSSION

4.10.1 UbA₅₂ Gene Structure - Exons and Introns

The human UbA₅₂ gene represented by clone λ UA1 consists of 5 known exons separated by 4 introns. The presently identified exon 1 contains 16bp of 5' non-coding sequence and is preceded by a perfect splice acceptor sequence (Chapter 4.7.2). The 128 codons coding for the ubiquitin-52aa tail fusion protein are spread approximately evenly over the remaining 4 exons, with exon 2 also containing 8bp of 5' non-coding sequence and exon 5 containing the 3' NCR of 84 or 90bp. The length of the 3' NCR appears to differ in a tissue-specific manner. While the same polyadenylation signal is employed, the polyadenylation site in the placental cDNA clone λ P15 (Chapter 4.2) is 6bp downstream from that used in three adrenal gland cDNAs (Chapter 4.3). The placental cDNA does not appear to represent an erroneously polyadenylated transcript, as a recently reported partial human UbA₅₂ cDNA from a leukemic cell line encoding

aa 40 to 76 of ubiquitin plus the 52aa tail protein is also polyadenylated at this site (Salvesen et al, 1987). However, these known examples are too few to confirm either alternate polyadenylation sites in all tissues, or tissue-specific polyadenylation.

The 5' NCR of this UbA₅₂ gene has not been completely characterised. Primer extension experiments indicate a 5' NCR length of about 65nt of which only 24 are found in the known exons 1 and 2. While S1 nuclease protection experiments using uniformly labelled probes covering most of the region upstream of exon 1 were inconclusive, it appears likely that another 5' non-coding exon is present in this region. Alternatively, exon 1 may continue upstream for some 40bp, although the presence of a perfect splice acceptor consensus adjacent to the present exon 1 is argumentative for, but not conclusive proof of, the former possibility. A third possibility is that transcription initiates at several positions over a 30 to 40bp region, as primer extension produced species of varying length, due either to different length mRNAs or premature termination of extension at mRNA secondary structures (Figure 4.7.3). This last possibility is noteworthy with respect to the presence of Sp1 promoter sites upstream of exon 1 (see below).

A most interesting feature is the presence of (at least) 4 introns in the UbA₅₂ gene, 3 of which interrupt the coding regions. In this respect the UbA₅₂ gene exhibits the standard structure of higher eukaryotic genes. An intron within a gene's 5' NCR is also a relatively common event, and notably the human UbB (Chapter 3) and chicken UbI (Bond

and Schlesinger, 1986) polyubiquitin genes have a similarly located intron. Although the number of characterised ubiquitin genes is small, introns are absent from the coding regions of all known polyubiquitin genes (Wiborg et al, 1985; Bond and Schlesinger, 1986; Baker and Board, 1987a/Chapter 3; Giorda and Ennis, 1987; Ozkaynak et al, 1987). This list could also include a Drosophila polyubiquitin locus partially characterised by Arribas et al (1986). To date, the only known coding-region intron-containing genes are the yeast Ubi1 and Ubi2 genes, homologues of the human UbA52 gene, which respectively contain a single 434 and 367bp intron between the second and third nucleotides of the third codon of the ubiquitin coding region (Ozkaynak et al, 1987). While this intron position is not conserved between yeast and human, these two species represent the extremes of the ubiquitin spectrum and any meaningful structural comparison should await the characterisation of the homologous gene(s) from intermediate species. Interestingly, the yeast Ubi3 ubiquitin/tail fusion gene, homologue of the human UbA80-type gene, contains no introns (Ozkaynak et al, 1987). Unfortunately our knowledge of the human-UbA80 gene is presently limited to an incomplete cDNA clone (Lund et al, 1985).

A further interesting feature of the positions of the coding-region introns is that they occur at the boundaries of some of ubiquitin's structural domains. The three-dimensional structure of ubiquitin has been determined by X-ray diffraction (Vijay-Kumar et al, 1985, 1987a) and further characterised by two-dimensional ^1H NMR studies (Di Stefano

and Wand, 1987) as described in Chapter 1.1. The structural representations determined by these authors are shown in Figure 4.10.1. Exon 2 encodes two antiparallel 7aa strands of a mixed β -sheet (aa 1/7 and 11/17), a 3.5 turn α -helix (aa 23 to 34), and the two reverse turns between these structures (aa 8/11 and 18/21). Exon 3 also encodes two short antiparallel strands of the β -sheet (aa 41/44 and 48/50), a short helical turn (aa 55 to 59) and three reverse turns (aa 37/40, 45/47 and 51/54). Finally, the ubiquitin-coding portion of exon 4 encodes the final strand of the β -sheet (aa 64 or 66 to 72) and the protruding C-terminus, while the splice junction lies within the last reverse turn (aa 63 to 65). While the tail protein has never been isolated and/or structurally characterised, notably the last exon (5) contains most of the putative metal/DNA binding domain and nuclear localisation signal (Chapter 4.8.2).

These observations suggest that the gene was constructed by the recruitment of pre-existing exons encoding discreet structural domains, rather than the invasion of a preformed intronless gene by introns, as it is unlikely that such insertions would be by chance at structural domain boundaries. This conclusion is in agreement with current theories on the origins of introns as reviewed by Gilbert et al (1986).

A final feature of the exon structure is that ubiquitin and the tail protein are not encoded by completely separate sets of exons, but that exon 4 encodes the 13 C-terminal aa of the former plus the first 22 residues of the latter. It thus

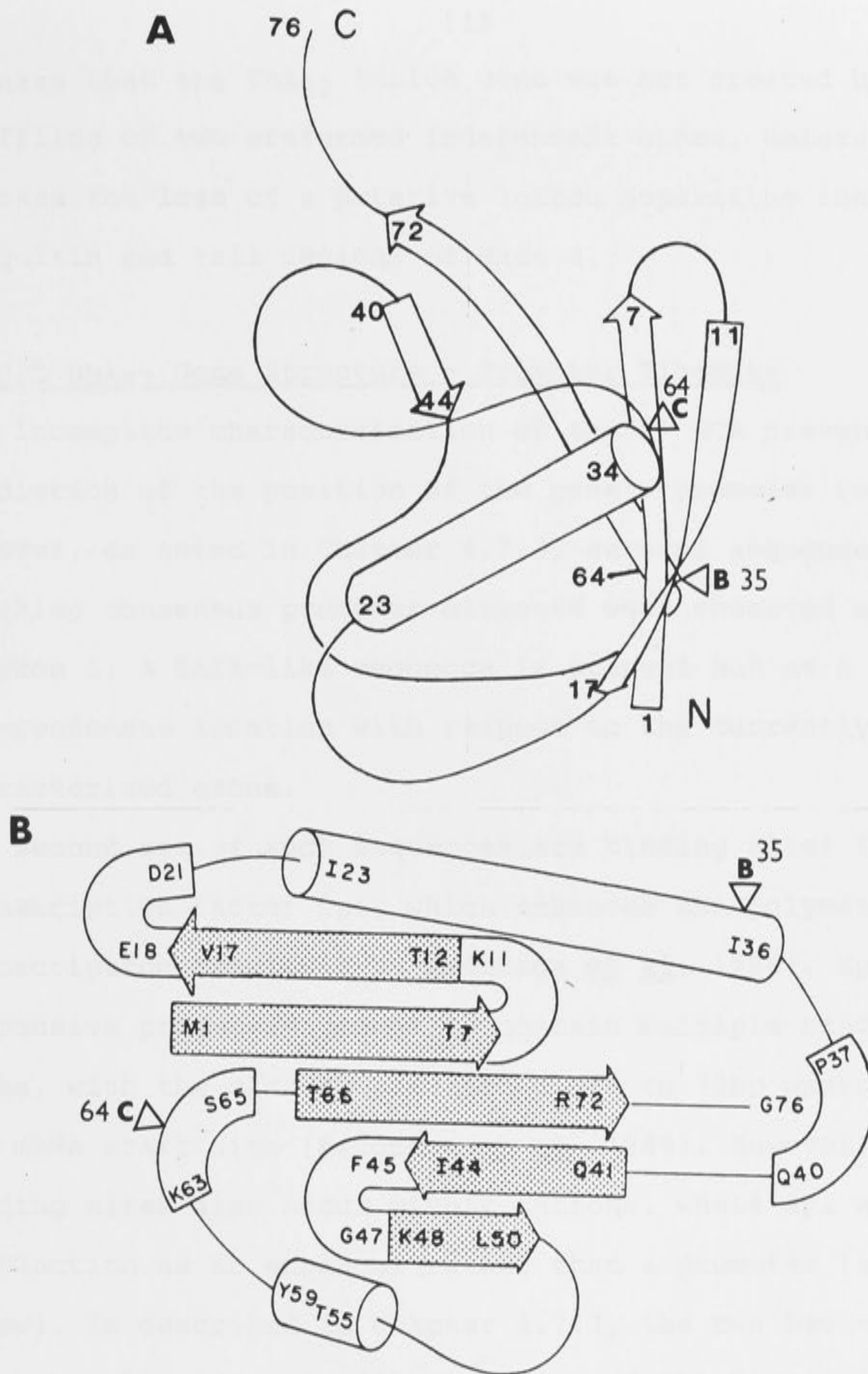


Figure 4.10.1: Ubiquitin Structural Domains and Intron Position.

The structural representations of ubiquitin as determined by Vijay-Kumar et al (1985) and Di Stefano and Wand (1987) are shown in panels A and B respectively. Cylinders and arrows represent helices and β -sheet structures respectively; other shapes represent reverse turns. The positions of introns B and C of the UbA₅₂ gene are indicated by open triangles along with the number of the codon in which they occur. Numbers refer to aa residues.

appears that the UbA₅₂ fusion gene was not created by exon shuffling of two preformed independent genes, unless one invokes the loss of a putative intron separating the ubiquitin and tail regions of exon 4.

4.10.2 UbA₅₂ Gene Structure - Promoter Elements

The incomplete characterisation of the 5' NCR prevents prediction of the position of the gene's promoter region. However, as noted in Chapter 4.7.3, several sequences matching consensus promoter elements were observed upstream of exon 1. A TATA-like sequence is present but at a clearly non-consensus location with respect to the currently characterised exons.

The second set of such sequences are binding sites for the transcription factor Sp1, which enhances RNA polymerase II transcription (reviewed by Kadonaga et al, 1986). Sp1-responsive promoters generally contain multiple binding sites, with the closest positioned ~40 to 70bp upstream of the mRNA start site (Kadonaga et al, 1986). However, Sp1 binding sites also occur within introns, where Sp1 appears to function as an enhancer rather than a promoter (see below). As described in Chapter 4.7.3, the two best-match Sp1 boxes lie 106 and 191bp upstream of the presently characterised exon 1, which is outside the consensus range (Kadonaga et al, 1986). However, if exon 1 were extended a further 41bp upstream (Chapter 4.7.3), the closest Sp1 box would be 65bp upstream, within the consensus range.

A second feature of the Sp1 binding site is its association with the promoter regions of several "housekeeping" genes:

that is, genes whose activity is required in all cells. Genes in this category include adenosine deaminase, hypoxanthine phosphoribosyl transferase, dihydrofolate reductase, hydroxymethyl glutaryl CoA reductase, U1 RNA, and scrapie prion protein genes (reviewed in Basler et al, 1986). Notably, these genes also lack a TATA box. In the case of the latter gene, Sp1-promoted transcription initiates at multiple sites over a 25bp region, as judged by S1 mapping, primer extension and in vitro transcription studies (Basler et al, 1986). This latter finding is interesting in view of the multiple primer extension products observed with the UbA₅₂ gene (Chapter 4.7.2), which may be indicative of a similar heterogeneity in transcription initiation, or may be methodological artefacts. Thus, it is possible that the Sp1 binding sites observed in λ UA1 may function as the UbA₅₂ promoter(s). In addition, the UbA₅₂ gene may well be a housekeeping gene, as recent construction and characterisation of yeast ubiquitin deletion mutants by Finley et al (1987) indicates that the ubiquitin/tail fused genes are sufficient for viability during exponential growth, while the polyubiquitin is specific for the stress response. Extrapolation of these results to the human system implies that the UbA₅₂ gene(s) is basally expressed and is most likely a housekeeping gene.

An alternate role for these Sp1 binding sites may be as enhancer sequences within a putative UbA₅₂ upstream intron (Chapter 4.7.3). Sp1 boxes have been implicated in the enhancement of transcription of the human pro- α 1(I) collagen gene by sequences from its first intron (Rossouw et al,

1987). Notably these sequences also enhance transcription when translocated upstream of the mRNA start site. Other reports indicate that this intron may contain several transcriptional regulatory elements, both positively and negatively acting (Bornstein et al, 1987; Bornstein and McKay, 1987). As noted in Chapter 4.7.3, a 19bp sequence containing one of the UbA₅₂ Sp1 sites is very similar to a 20bp Sp1-containing region of the pro- α 1(I) collagen gene intron. However, it must be noted that the pro- α 1(I) collagen gene contains consensus CCAAT and TATA promoters upon which presumably the Sp1 enhancement acts, whereas the UbA₅₂ promoter region has not been fully characterised.

4.10.3 Features of the Fused Ubiquitin/Tail Protein

The UbA₅₂ gene λ UAI encodes a fusion protein of ubiquitin and a 52aa tail protein. Thus all known transcriptionally active ubiquitin genes and cDNAs encode fusion proteins - either of ubiquitin to itself to produce polyubiquitin (eg UbB, Chapter 3) or to an unrelated tail sequence, as with UbA₅₂. As discussed for the polyubiquitin genes (Chapter 3.11), this implies that ubiquitin is always generated by post-translational proteolytic processing. As also previously discussed, the protease identified by Bachmair et al (1986) responsible for the in vivo cleavage of ubiquitin from engineered ubiquitin- β -galactosidase fusion proteins in yeast may be a candidate for the processing of the fused ubiquitin/tail protein.

The presence of 31% basic aa, a nuclear translocation-type signal and a metal/DNA binding domain (Chapter 4.8.2) argue

strongly for a nuclear location for the tail protein. While the tail protein may function to transport ubiquitin to the nucleus (Lund et al, 1985), its high level of evolutionary conservation, including the absolute maintainance of the metal/DNA binding domain, is indicative of a specific function(s). Whether this would involve the fused ubiquitin/tail protein, the free tail protein or both is not known. Finley et al (1987) have speculated on several possibilities, such as the free tail protein acting as a DNA-binding protein to regulate gene expression, or in forming other protein/tail conjugates. This latter case is interesting in the light of a recent report of the analysis of a 16 kDa protein isolated with anti-synthetic 80aa-tail antisera, which was found to be the 80aa tail protein as a part of a non-ubiquitin protein (see Bond et al, 1988). Alternatively, the fused protein may provide a means of targeting ubiquitin to specific chromosomal regions regulated by the tail's DNA binding properties, followed by the proteolytic release of ubiquitin for further activities confined to that site (Finley et al, 1987). These authors plan further investigations of the functions of the yeast tail proteins including in vitro DNA-binding studies, immunological studies using tail-specific antibodies, and gene deletion analyses of the ubiquitin/tail fusion genes, results of which should be relevant to all eukaryotes.

4.10.4 The UbA₅₂ Subfamily

The members of the UbA₅₂ subfamily described in this chapter include a transcriptionally active gene, a processed

pseudogene, and a second pseudogene which exhibits some features of processing.

A calculation of synonymous mutation rates between the UbA₅₂ gene and pseudogene coding regions similar to that described for the UbB processed pseudogenes (Chapter 3.11.4) suggests that λ UA4 and EHD5 arose approximately 23 ± 7 and 47 ± 11 million years ago respectively, and thus appear to be older than the identified UbB pseudogenes.

Hybridisation analysis indicates that there are several as yet uncharacterised members of this subfamily, as PstI and BamHI each produce at least 9 restriction fragments which hybridise to a tail/3' NCR-specific probe, three of which correspond to the known subfamily members (Chapter 4.9).

Notably, the first intron of the UbA₅₂ gene λ UA1 is a single-copy sequence in the human genome: that is, none of the uncharacterised subfamily members contain a corresponding intronic sequence. Thus either the introns are evolving very rapidly, or the UbA₅₂ subfamily consists of one active gene (λ UA1) and several processed (ie. intronless) pseudogenes. The latter possibility is quite likely, given that two UbA₅₂ processed (or partially-processed; EHD5) pseudogenes were isolated, and that the UbB subfamily contains at least 3 processed pseudogenes (Chapter 3). In addition, the 5 UbA₅₂ cDNAs thus far observed (Chapters 4.2 and 4.3; Salvesen et al, 1987) all appear to have arisen from the λ UA1 gene, as they differ from the gene by a maximum of two nucleotides, which may represent artefactual or allelic differences (Chapter 4.8). Conversely in yeast, which contains two UbA₅₂-type genes, the coding regions vary

at the DNA level by ~15%, and considerably more in the NCRs (Ozkaynak et al, 1987). Thus the available evidence suggests that λ UA1 may represent the only transcriptionally active UbA₅₂ gene amongst a subfamily composed otherwise of processed pseudogenes. Further resolution of this situation must await the characterisation of the remaining subfamily members.

CHAPTER 5

LENGTH POLYMORPHISM AT THE HUMAN Ubc LOCUS

CHAPTER 5. LENGTH POLYMORPHISM AT THE HUMAN UbC LOCUS5.1 Identification of a Putative UbC mRNA Length Polymorphism

During Northern analysis with ubiquitin coding region, UbB- and UbA₅₂ gene-specific probes (Chapters 3.6.5 and 4.4), the lymphocyte RNA isolated from one individual (proband) contained an extra ~2200nt species which hybridised to the coding-region probe. This extra species does not appear to be an artefact of the RNA isolation procedure, as it was observed in RNA prepared on 3 separate days from freshly drawn blood, and was not present in parallel preparations from other individuals (see Figure 5.2.1, Panel D). This species did not hybridise with either the UbB-specific probe (Figure 3.6.5) or the UbA₅₂ specific probe (Figure 4.4.1), and is therefore either a UbC variant, or an abnormal UbA₈₀ transcript. The latter case would require an unspliced UbA₈₀ mRNA, as the observed species is ~1600nt longer than the mature UbA₈₀ mRNA. While the structure of the UbA₈₀ gene is unknown, if it is similar to the UbA₅₂ gene (Chapter 4), then this length discrepancy could be explained by an unspliced intron(s). However, this would require a consistent UbA₈₀ splicing defect in this individual not seen in others, while the UbB transcript has been spliced correctly. The former possibility of a variant UbC transcript seems the more likely, as the new species is ~200 to 250nt shorter than the UbC transcript, which could be explained by alternate RNA processing, or more interestingly, loss of a single coding unit.

The observation of this putative UbC mRNA length polymorphism (mRLP) prompted further Southern and Northern analysis to further characterise the polymorphism.

5.2 Correlation of the UbC mRLP with a Restriction Fragment Length Polymorphism (RFLP), Identification of a Second UbC mRLP/RFLP and Demonstration of Mendelian Inheritance

Fortunately, material was available for family studies, including both parents and three sibs of the proband. The restriction enzyme chosen for Southern analysis was HaeIII (recognition sequence GGCC) which, according to the partial UbC gene sequence reported by Wiborg et al (1985) cleaves 25bp upstream of the first initiation codon, does not cleave within any of the 9 UbC ubiquitin coding units, and does not cleave within the reported 116bp downstream of the stop codon. Therefore, HaeIII cleavage should produce a single hybridising fragment (HF) of at least 2199bp with a ubiquitin coding region probe. Conversely, as a result of DNA sequence variation, HaeIII cleaves 7 times within the UbB gene and its flanks (Figure 3.6.2, nt 1519, 2163, 2239, 2391, 2467, 2619, 2695). Thus, UbB-originating HFs would be of 644bp or less. Similarly, the other known ubiquitin genes and pseudogenes (Chapters 3 and 4) would produce small and/or weak HFs. Hybridisation of a coding-region probe to HaeIII digested genomic DNA prepared from random blood donors generally produced a constant medium intensity ~2.5kb HF, a single strongly hybridising ~2.3kb HF and several smaller weak HFs (see Figure 5.2.1). Occasional extra HFs were observed as described later. Presumably this 2.3kb HF represents the UbC

gene, implying that a HaeIII site is present ~ 100 bp downstream of the reported sequence (Wiborg et al, 1985). Thus most individuals appear homozygous for a 9 coding unit UbC gene. The corresponding analysis with the family of the proband produced the following result. First, the proband exhibited an RFLP at the UbC locus, with ~ 2.3 and ~ 2.05 kb HFs (Figure 5.2.1). Second, the proband's mother exhibited the same RFLP, suggesting heterozygosity at this locus. As the two HFs differ by ~ 200 to 250bp, about the length of a ubiquitin coding unit (228bp), one possibility is that the ~ 2.05 kb HF only contains 8 coding units. This phenotype will thus be termed 9,8. Third, the proband's father exhibited a different RFLP, with HFs of ~ 2.3 and ~ 1.85 kb. Again, one possibility is that the ~ 1.85 kb HF has lost a further coding unit to contain only 7. This phenotype will thus be termed 9,7. Fourth, these HFs are inherited in Mendelian fashion. The two male sibs (proband and his brother) have the 9,8 phenotype, inheriting the type 9 allele paternally and the type 8 maternally. Their sister also clearly demonstrates the Mendelian inheritance of these alleles, being a compound heterozygote of the type 8 (maternal) and type 7 (paternal) alleles (Figure 5.2.1).

Notably, the mRFLP observed in the proband corresponds to the RFLP observed at the DNA level. This correlation was further investigated by Northern analysis of lymphocyte RNA prepared from family members. The results of this study are included in Figure 5.2.1. Unfortunately, some samples were degraded in transit. However, in each family member studied, the mRFLP corresponds with the RFLP. Notably, the compound heterozygote

Figure 5.2.1: Characterisation of the UbC Polymorphisms
by Southern and Northern analysis.

Panels A and B (Top): Southern analysis of HaeIII digested genomic DNA. Lane C: control 9,9 homozygote. Lanes 1 to 6: family samples. P = Proband (Lane 5). Lanes 7, 8 and 9: Unrelated individuals. Arrows at left indicate the type 9, 8 and 7 alleles. Size standards at right are in kb.

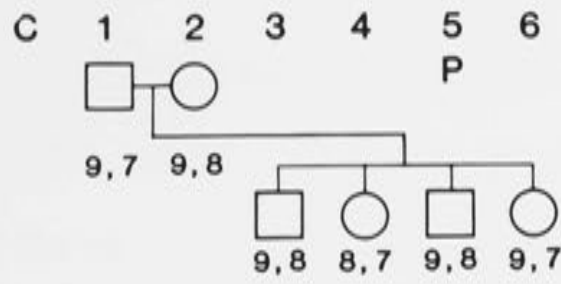
Panels C and D (Bottom): Northern analysis of total lymphocyte RNA. Lane designations are as above. Lanes 10 and 11: RNA prepared from unrelated individuals. Species are identified UbA, UbB and UbC after Wiborg et al (1985). Arrows to the right of panel C indicate the type 9, 8 and 7 UbC mRNAs.

Panels A and C: Family studies. The pedigree is given above Panel A. Square = male, circle = female.

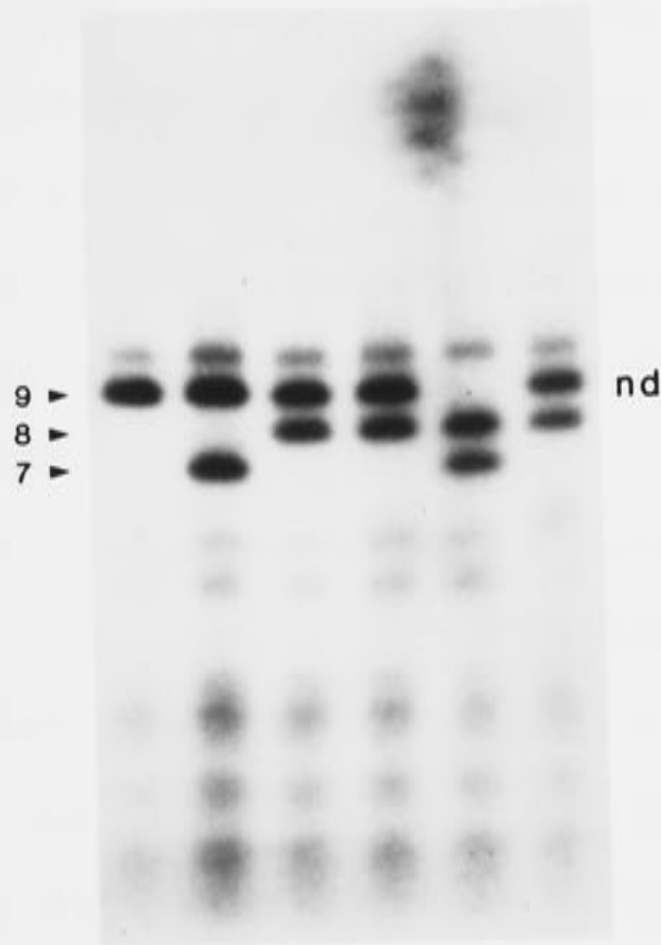
Phenotypes are shown under the pedigree symbols. nd = not determined.

Panels B and D: Examples of population studies. Panel B shows some phenotypes observed: 9,7 (lane 7), 9,8 (lane 8), 8,8 (lane 9), and 9,9 (lane C). The 9,7 individual was selected initially as a BamHI UbC heterozygote from a set of 39 samples and thus has^{not} been included in subsequent frequency calculations (Chapter 5.3) for statistical reasons.

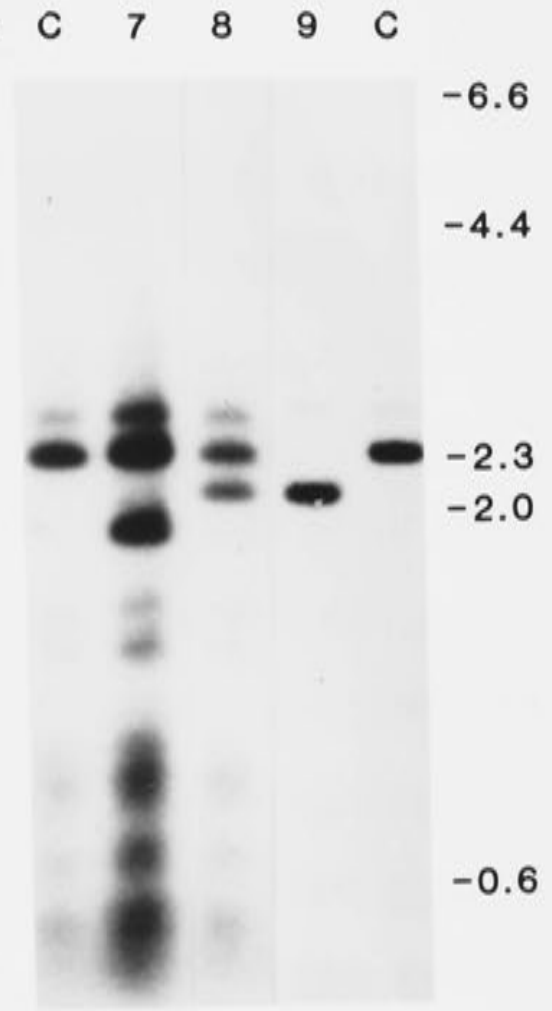
All panels represent hybridisation with a ubiquitin coding region probe. The constant ~2.5kb band observed in Panels A and B is presumably a non-UbC ubiquitin-positive restriction fragment.



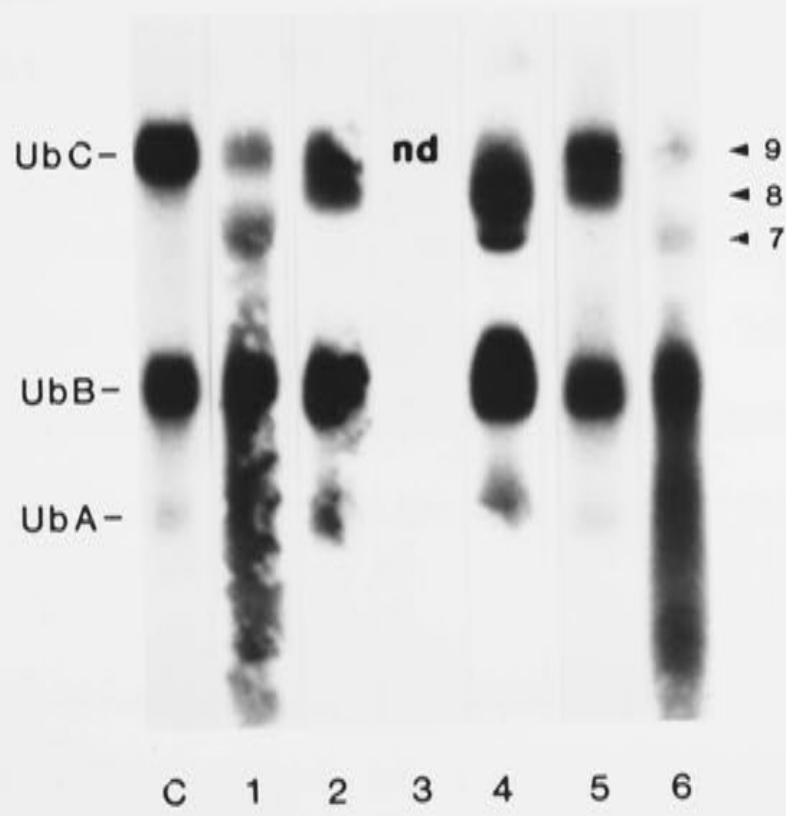
A



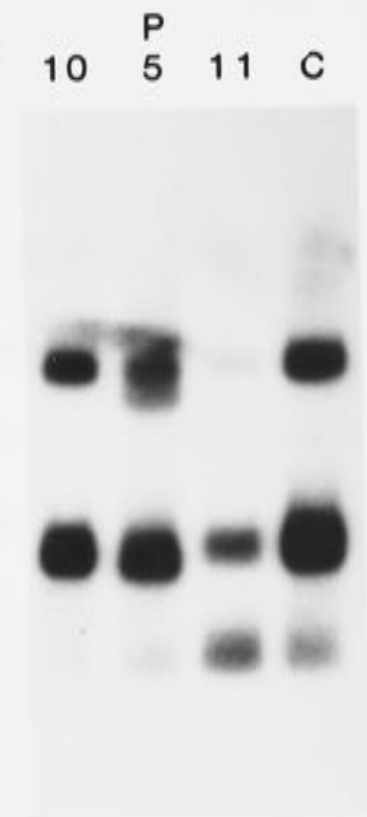
B



C



D



completely lacks the normal length Ubc transcript. A second sister of the proband for whom insufficient material was available for DNA preparation was found by Northern analysis to have the same mRFLP pattern as the father, and is thus predicted to have the 9,7 RFLP phenotype.

Significantly, the HaeIII RFLP is also identified by other restriction enzymes. For example, PstI, BamHI and HindIII each show an RFLP in individuals positive for the HaeIII RFLPs (not shown). These RFLPs are less easy to detect, as the putative Ubc HFs (Chapter 3.8) are ~8.5kb (PstI) and ~10kb (BamHI), and ~8kb (HindIII; not shown), and thus a ~230 or ~460bp difference causes a markedly smaller change in mobility than in the ~2.3kb HaeIII HF. In fact, the individual studied in Chapters 3.8 and 4.9 is a 9,8 heterozygote, which is just detectable in the PstI digest (eg Figure 4.9.1, Panel B), but is easily detectable with HaeIII (Figure 5.2.1, lane 8). Identification with multiple enzymes is indicative of a variation in the length of DNA between two unchanged restriction sites, rather than the loss or creation of a restriction site, which is very unlikely to occur for four enzymes simultaneously. Tight linkage disequilibrium is an alternate explanation, but is less likely, given the correlation between mRFLP and RFLP. This observation supports the possibility that the RFLP represents the loss of one or two coding units relative to the 9 coding unit gene.

5.3 Calculation of Gene Frequency

The gene frequencies of the type 7, 8 and 9 alleles were calculated from the hybridisation patterns of 35 randomly se-

lected blood donors. It must be stressed that this is a small sample size for statistical analysis but should provide an estimate of frequency. The data is presented in Table 5.3.1

5.4 Discussion

The results presented in this Chapter describe RFLPs at the human Ubc locus which is correlated with mRLPs of the Ubc transcript. Polymorphism in mRNA length may arise for one of several reasons. These include: alternate polyadenylation sites, such as the murine dihydrofolate reductase gene, which has 11 known polyadenylation sites over a 5.5kb range (Hook and Kellems, 1988); use of alternate promoters and 5' non-coding exons, such as the human aldolase A gene transcripts, which differ by up to 127nt (Maire et al, 1987); or alternate splicing of pre-mRNAs, such as the human fibronectin gene, where the different type fibronectins are produced from mRNAs alternatively spliced at 3 coding-region exons (Mardon et al, 1987). Another novel example is the human apolipoprotein (apo) B-100 hepatic mRNA, which is normally ~14000nt long. However, in the intestine, co-or post-translational events change a CAA Gln codon to UAA stop codon approximately midway through the mRNA, translation of which produces the smaller apo B-48 protein (Chen et al, 1987, Powell et al, 1987). This event also results in a shorter (~7-8000nt) mRNA due to usage of cryptic AATAAA and G-T signals downstream of the new stop codon. However, all of these polymorphisms do not involve changes at the genomic DNA level and thus cannot be correlated with an RFLP. One mRLP that has been correlated with an RFLP is seen with the human leukocyte antigen DQ α ,

Allele:	9	8	7
Frequency:	0.857	0.129	0.014

N = 35

Phenotype	Frequency	
	Predicted	Observed
9,9	25.7	26
8,8	0.6	1
7,7	0.007	0
9,8	7.7	7
9,7	0.86	1
8,7	0.13	0

TABLE 5.3.1: Frequencies of the UbC Type 7, 8, and 9 Alleles.

TOP: Observed frequencies of each allele from a randomly selected caucasian population sample of 35 individuals.

BOTTOM: Observed and predicted frequencies of UbC phenotypes. The predicted phenotype frequencies were calculated from the observed allele frequencies. The observed phenotype frequencies are in Hardy-Weinberg equilibrium and are not significantly different from the predicted values by the chi-squared test.

and is indicative of differences in gene structure (Loiseau et al, 1986). Notably, the UbC DNA RFLPs correlate with mRLPs observed in the same individuals, indicating that the variant alleles are transcriptionally active. As ubiquitin coding sequences comprise ~90% of the expected ~2.3kb 9 coding unit HaeIII fragment being studied, it is most likely that the deletions occur in the coding region. A deletion occurring in the 5' or 3' flank would be expected to disrupt transcription of the allele, whereas discrete mRNAs are observed, especially in the case of the compound heterozygous 8,7 sib who completely lacks the common type 9 allele. Combined with the observation that in both the RFLP and mRLP, the length differences are in multiples of ~230bp, the most likely explanation for the length polymorphism is variation in the number of ubiquitin coding units. The lengths of the identified polymorphic fragments suggests that some alleles contain 7 and 8 coding units compared to the more common 9. An important conclusion from these results is that either the UbC gene itself is dispensable, or that the number of coding units is not critical. The former possibility is the least likely considering the conservation of the protein encoded by each coding unit. With respect to the latter case, clearly 9 coding units are not required, as exemplified by the observed 8,7 heterozygote and 8,8 homozygote (Figure 5.2.1). This result is in accordance with recent findings of Finley et al (1987), whereby the yeast 5 coding unit UBI4 polyubiquitin gene was found by construction of a gene deletion mutant (ubi4) to be essential for viability under stress conditions. However, the stress-sensitive ubi4 phenotype could be

complemented by an in vitro constructed UBI4 "minigene" containing only a single ubiquitin coding unit and UBI4 flanking regions (including the heat shock promoter), indicating that the function of UBI4 under stress conditions is to provide ubiquitin monomers rather than a polyubiquitin protein (Finley et al, 1987). While it is not known whether the human UbC gene contains heat shock promoters (Wiborg et al, 1985), the situation appears analogous to the yeast system whereby the number of coding units is not essential for the proper function of the gene.

The final feature of the UbC RFLP and mRFLP relates to the evolution of the gene. While only a small sample size has been studied, the only alleles observed appear to contain 7, 8 and 9 coding units. This observation is most simply explained by an unequal crossover event of two 8 coding unit alleles misaligned by a single coding unit, producing 7 and 9 coding unit alleles. Other crossovers are also possible as discussed below. Unequal crossover events have been suggested to explain the substantial ubiquitin RFLP and mRFLP observed in Xenopus (Dworkin-Rastl et al, 1984) and Drosophila (Arribas et al, 1986), and have been proposed as a general mechanism for ubiquitin gene evolution (Sharp and Li, 1987a, 1987b). The high frequency of the type 9 allele could be explained by its potential selective advantage in producing more ubiquitin monomers, especially under stress conditions (Ozkaynak et al, 1987), although it is not known if the UbC gene is a heat shock gene. The observed frequencies of the type 9, 8 and 7 alleles (Table 5.3.1) are supportive of

selective pressure maintaining alleles with a higher number of coding units.

Interestingly, a partial human UbC-type cDNA sequence recently reported by Einspanier et al (1987b) may in fact correspond to the type 7 allele. This cDNA clone pHGR21 contained ~3.5 coding units and a UbC-like extra Val codon and 3' flank. While the coding unit immediately upstream of the termination codons of pHGR21 and the UbC gene were practically identical, the next 2.5 coding units did not match, prompting the authors to conclude that pHGR21 may represent another yet unknown polyubiquitin gene (Einspanier et al, 1987b). However, a closer inspection of the pHGR21 coding units reveals that they almost perfectly match a UbC gene deleted for coding units 7 and 8. This comparison is presented schematically in Figure 5.4.1. As noted by Einspanier et al (1987b), the first half coding unit is identical to the fourth UbC coding unit. If the sequences are then compared from this point, the second coding unit of pHGR21 differs by only one nt from the 5th of UbC, the third is identical to UbC's 6th, and the last differs by only one nt from UbC's 9th coding unit. Thus pHGR21 appears to lack the 7th and 8th UbC coding units. While such a deletion could have occurred during cDNA construction by a similar mechanism to that proposed for the UbB processed pseudogenes (Chapter 3.11.4) (although the inverted repeats are not as well conserved), it could also be explained by a deletion of these two coding units at the genomic level; ie, a type 7 allele. Such a deletion could be explained by an unequal crossover of two 8 coding unit alleles at the junction of their second-

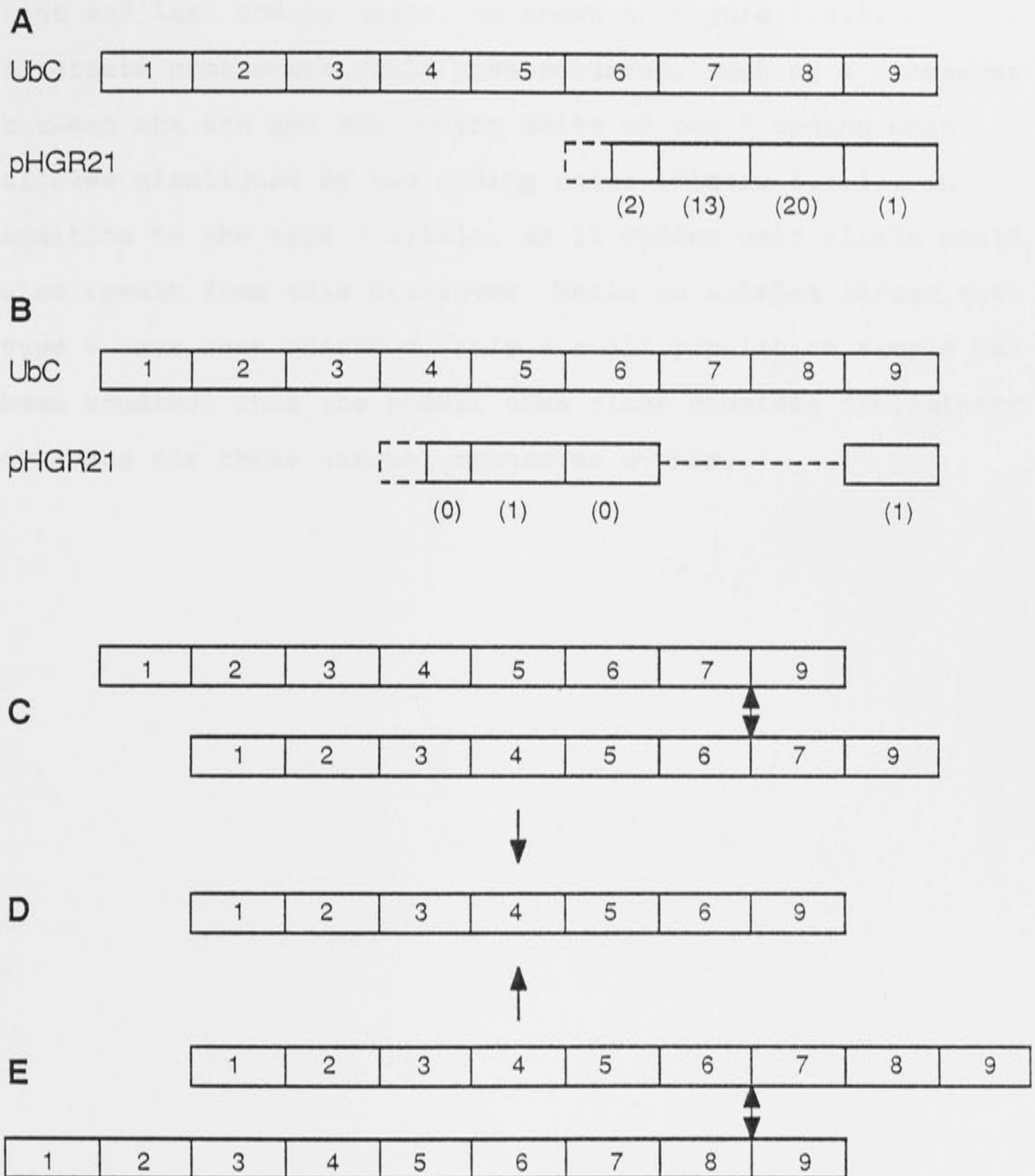


Figure 5.4.1: Relationship of the UbC Gene and the pHGR21 cDNA.

(A): Direct comparison aligned at 3' ends. UbC coding units (boxes) are numbered, while the number of differences between a pHGR21 coding unit and the UbC coding unit directly above are shown in brackets.

(B): Alignment after introduction of a 2-coding unit gap (dotted line) markedly improves the similarity.

(C) and (E): Two possible unequal crossover events (double-headed arrows) leading to a pHGR21-like 7 coding unit allele (D). Crossover (C) is between two 8 coding unit alleles (numbered 1-7 and 9 for ease of comparison), while (E) is between two 9 coding unit alleles. Each crossover will produce another allele which is not shown.

last and last coding units, as shown in Figure 5.4.1. Alternate crossovers could have occurred, such as a crossover between the 6th and 9th coding units of two 9 coding unit alleles misaligned by two coding units (Figure 5.4.1). In addition to the type 7 allele, an 11 coding unit allele would also result from this crossover. While no alleles larger than type 9 have been observed, only a small population sample has been studied. Thus the pHGR21 cDNA clone provides preliminary evidence for these unequal crossover events.

The 15th and 16th chromosomes were employed as reference chromosomes for the identification of the human UbB gene. The results of the FISH analysis are shown in Figure 6.1. The results of the FISH analysis are shown in Figure 6.1. The results of the FISH analysis are shown in Figure 6.1.

CHAPTER 6

CHROMOSOMAL LOCALISATION OF THE HUMAN UBIQUITIN UbB GENE

The human UbB gene was localised to the 15th chromosome by FISH analysis. The results of the FISH analysis are shown in Figure 6.1. The results of the FISH analysis are shown in Figure 6.1. The results of the FISH analysis are shown in Figure 6.1.

CHAPTER 6: CHROMOSOMAL LOCALISATION OF THE HUMAN UBIQUITINUbB GENE

Note: The in situ hybridisation techniques employed in determining the chromosomal location of the human UbB gene were performed by Dr Graham C. Webb, Human Genetics Group, Division of Clinical Sciences, John Curtin School of Medical Research, Australian National University, Canberra, Australia, to whom I am deeply grateful. My contribution to these investigations was limited to constructing and supplying the purified recombinant plasmids used as probes, and the interpretation of the results. As such, these techniques have not been included in Chapter 2, but are described briefly below.

6.1 Introduction, Materials and Methods

In situ hybridisation is a technique whereby intact human chromosomes are hybridised with radioactively labelled DNA probes, washed under conditions to remove non-specifically bound probe, exposed to photographic emulsion, developed, stained to produce a chromosome banding pattern allowing individual chromosome identification, and then scored for the position(s) of silver grains over the chromosomes. The general principles of the technique including technical considerations have recently been reviewed by Buckle and Craig (1986).

Recombinant plasmids used for in situ hybridisation were: (1) pUb26, being the intact UbB cDNA insert of pRBL26 (Chapter 3.1) subcloned into pUC18; and (2) pINT, being the 750bp BamHI/BglII fragment from the UbB gene (Chapter 3.6)

containing 22bp of 5' NCR, the entire 715bp intron, the 6bp of the 5' NCR encoded on the second exon, and 9bp of coding sequence (Figure 3.6.2, nt 1340 to 2091) subcloned into pUC18. Plasmids were labelled with [³H]dATP, [³H]dCTP and [³H]dTTP (Amersham) using a nick-translation kit (Amersham; N5500). The experimental procedure employed was largely as described by Board and Webb (1987), with modifications applied to denaturing of the pINT probe and chromosomes plus the stringency rinses as described by Dorlon (1986).

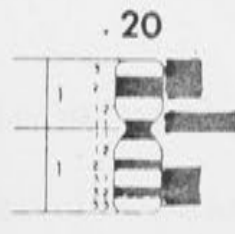
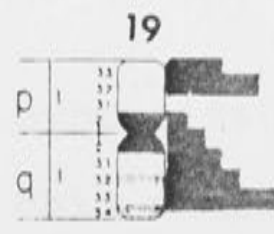
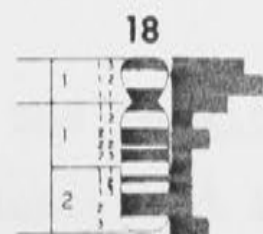
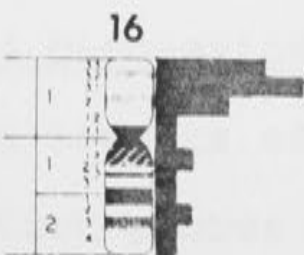
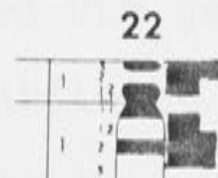
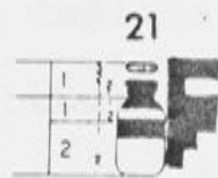
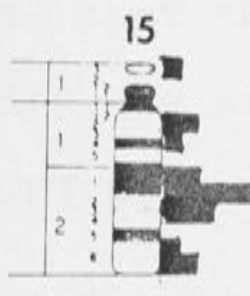
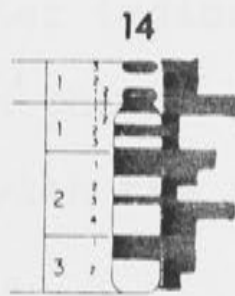
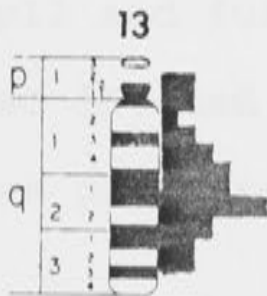
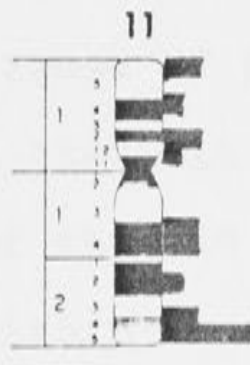
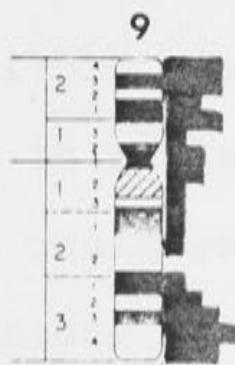
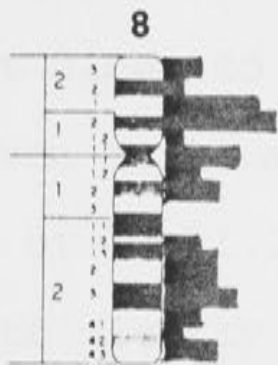
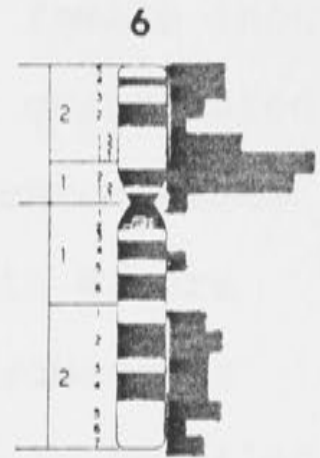
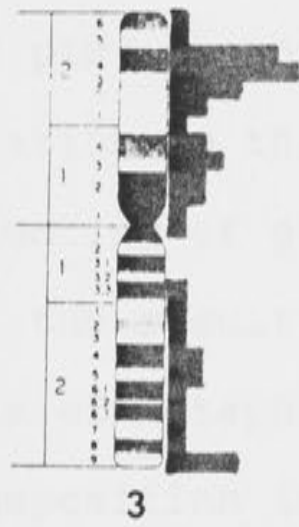
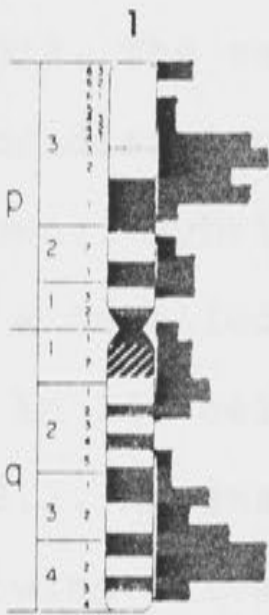
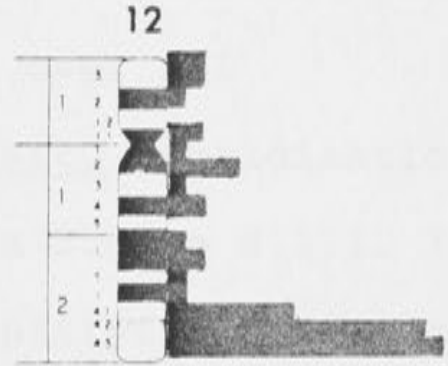
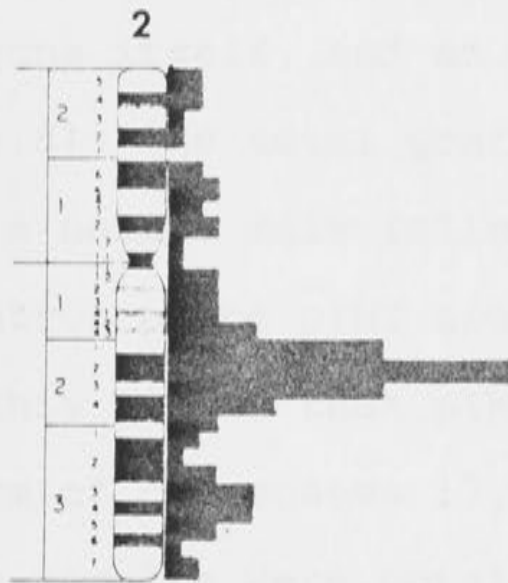
6.2 In Situ Hybridisation of a UbB cDNA to Human Chromosomes

The results of the in situ hybridisation of the UbB cDNA probe pUb26 to human chromosomes are shown in Figure 6.2.1 as a diagram of total grain counts over all chromosomes. The most obvious concentration of grains occurs over the short arm of chromosome 17, centred approximately over band 17p12, which is therefore a likely candidate for the location of the UbB gene. However, several other peaks were also observed, including "medium" signal peaks over 2q22 and 12q24-qter, and "weaker" peaks over 6p12, 16p12-p13, 1q4, 1p32, 3p23-p24 and 2p2. While the probe used is a UbB cDNA, it contains 2-3 ubiquitin coding units which are ~80 to 95% similar to the coding units of other ubiquitin genes and pseudogenes (Chapters 3, 4 and 5), and as such should hybridise to all members of the ubiquitin gene family. Therefore, in situ hybridisation using a UbB gene-specific probe was performed as described below.

Figure 6.2.1: In Situ Hybridisation With a UbB cDNA Clone.

In situ hybridisation with plasmid pUb26 containing the pRBL26 UbB cDNA (see Chapter 6.1) followed by the counting of grains in both prophasic and metaphasic cells produced this plot of grain distribution over all chromosomes. Sample peak heights (below) are given in numbers of grains.

17p12	:	65
2q22	:	20
12q24	:	16
6p12	:	8
16p13	:	8
3p2	:	7
1p32	:	6
1q4	:	6
19q13	:	6



6.3 In situ Hybridisation of the UbB Intron to Human Chromosomes

The UbB intron was chosen as a UbB gene-specific probe as it hybridised to only two restriction fragments on a Southern blot: the UbB gene itself, and an apparently duplicated UbB gene (Chapter 3.8). The total grain counts over all chromosomes of a normal male following in situ hybridisation with the UbB intron probe pINT are shown in Figure 6.3.1. It is clear from this Figure that pINT is identifying a region on the short arm of chromosome 17, with the peak over band 17p12. The same results were obtained by in situ hybridisation of pINT to chromosomes of a normal female (not shown). Hybridisation to this region was further quantitated by a detailed counting of grains over prophasic chromosomes 17 in 145 cells, the results of which are shown in Figure 6.3.2. An example of metaphasic and prophasic chromosomes showing silver deposition is given in Figure 6.3.3. Detailed counting revealed that the grain peak was centred over band 17p11.2, but close to 17p12. However, the range of β particles from tritium in emulsion is such that either of the bands 17p12 or 17p11.2 could be the site of the UbB gene. These results clearly place the UbB gene within band 17p11.2-17p12 and further indicate that the duplicated UbB gene is also located in this region, as no other significant peaks were produced with the pINT probe (Figure 6.3.1).

6.4 Discussion

The results of the in situ hybridisations with UbB cDNA and intron probes described in this chapter localise the

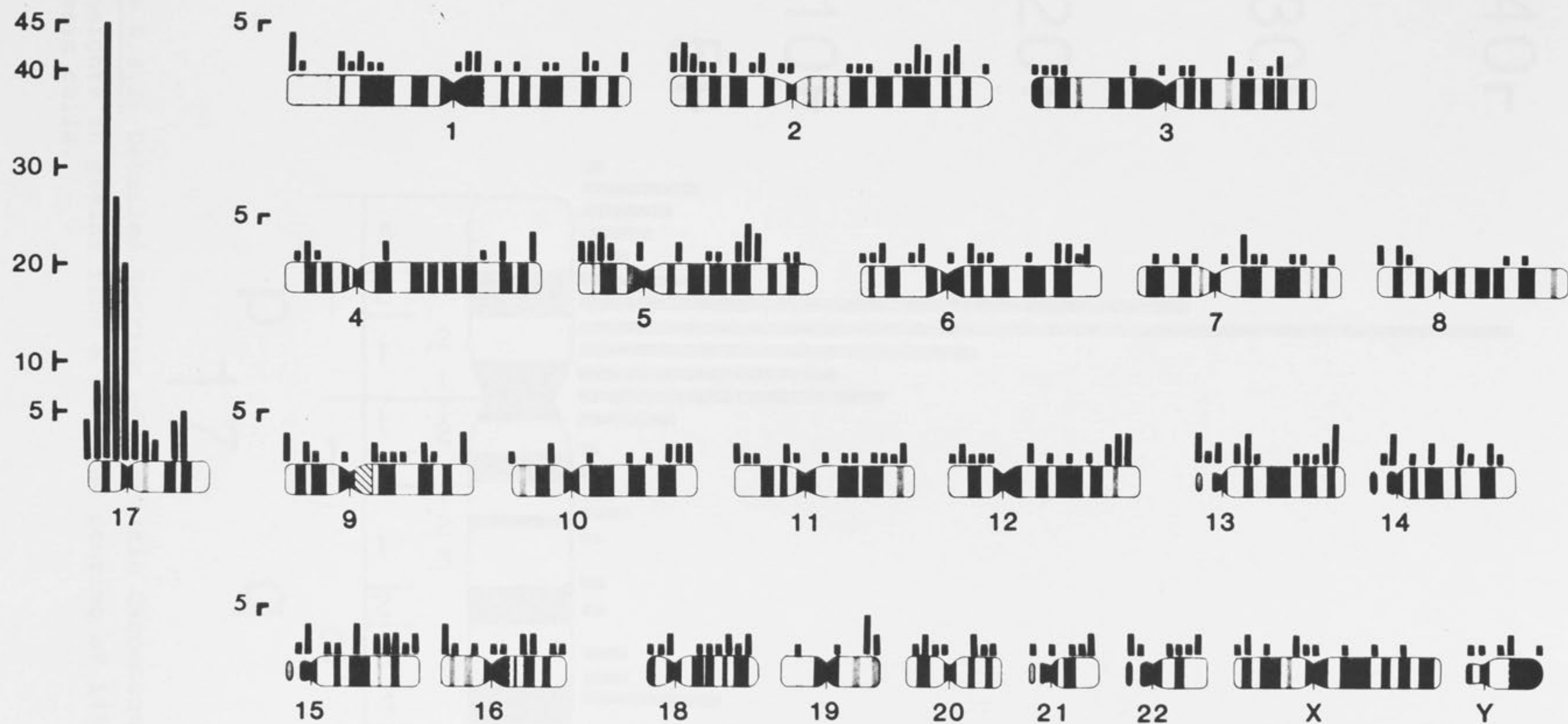


Figure 6.3.1: In Situ Hybridisation with the UbB Gene Intron Probe pINT.

Plot of grain distribution over all chromosomes. Peak heights are in numbers of grains.

40r

30r

20r

10r

5r

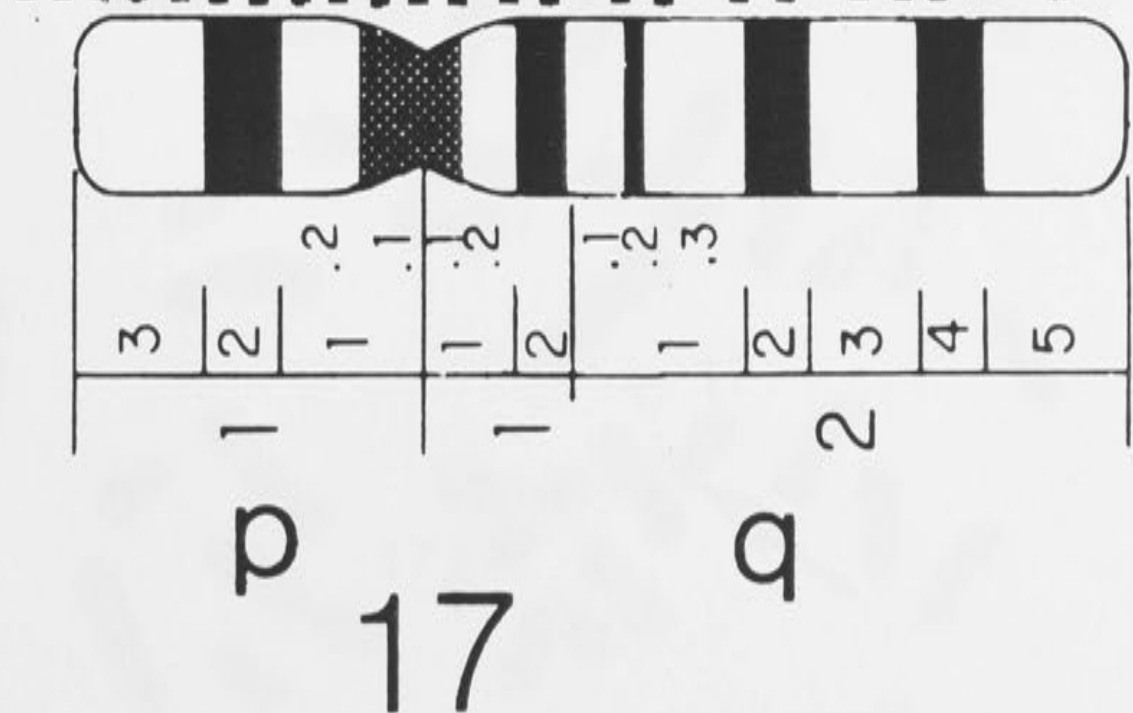
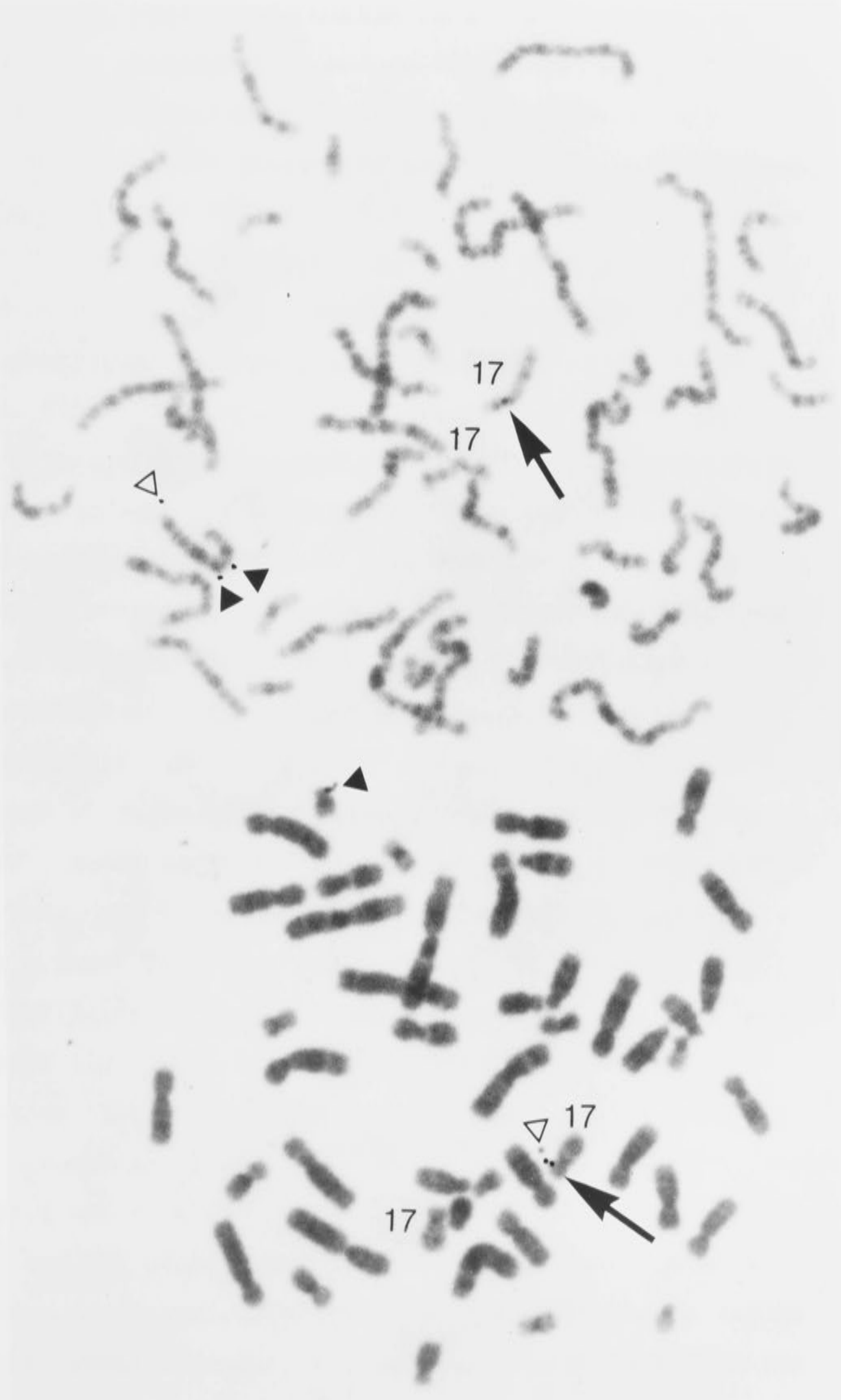


Figure 6.3.2: Detailed Scoring of Prophasic Chromosomes 17. Peak heights in grains from a detailed scoring of 145 prophasic cells.

Figure 6.3.3: In situ hybridisation.

In situ hybridisation of the chromosomes of a normal male with the UbB intron subclone pINT. Adjacent cells in prophase (top) and metaphase (bottom) of mitosis are shown. The chromosome 17 pairs are numbered. The arrows point to label grains over the target band 17p11.2 in the prophasic cell and 17p12 (2 grains) in the metaphasic cells. Scoreable background grains are indicated by solid triangles and unscorable grains, more than one chromatid width from the chromosomes, by open triangles.



apparently duplicated human UbB gene to the short arm of chromosome 17, most likely within band 17p11.2-17p12. As expected, the gene-specific nature of the UbB intron allowed positive identification of the UbB gene amongst a large family of potentially cross-hybridising genes, and indicated that the duplicated UbB genes lie within the same chromosomal region. Conversely, the more general UbB cDNA clone identified many peaks of varying intensity (Figure 6.2.1). Interpretation of peak height may be difficult for several reasons. First, Southern blots indicate that hybridisation signal is proportional to coding unit number, at least with a single coding unit probe (Figures 3.8.1 and 4.9.1). Second, several processed pseudogenes have been identified (Chapters 3 and 4) which will presumably occur at different chromosomal locations to their parent gene. Third, the ubiquitin gene and pseudogene coding units differ by up to 20% which would affect hybridisation efficiency. Finally, the probe used was a UbB cDNA which would have more affinity for UbB-related loci. With these considerations in mind, tentative analysis can be made. The 9 coding unit UbC gene would be expected to produce a major peak, whereas the strongest peak observed is over 17p11.2-17p12, the location of the UbB genes. The second major peak (2q22) is considerably smaller, prompting speculation that the UbC gene is also located on the short arm of chromosome 17 and contributes to the peak height. Such a location would be in agreement with the possibility that the UbB and UbC genes arose from an early unequal crossover event of a single polyubiquitin precursor gene (Sharp and Li, 1987a and 1987b). However, the same authors suggest that the

human ubiquitin loci may have different chromosomal locations based on an apparent lack of inter-locus gene conversion. In this respect, it must be noted that the peak observed on chromosome arm 17p by in situ hybridisation covers thousands of kilobases, and thus the UbB and UbC loci may still be sufficiently isolated to reduce the frequency of gene conversion events.

The UbB processed pseudogenes would be expected to be dispersed through the chromosomes and may correspond to some of the peaks on chromosomes other than 17. The second largest peak (band 2q22) could potentially represent the two coding unit pseudogene EHB4, while some of the smaller peaks may correspond to the single coding unit pseudogenes EHD1 and EHB7, and other as yet uncharacterised fragments revealed by the UbB 3' NCR probe (Figure 3.9.1). The chromosomal location of the UbA₅₂ gene λ UA1 has not yet been determined, although its gene-specific intron A probe (Figure 4.9.1) is an obvious candidate for future in situ hybridisation. However, some of the other minor peaks observed may correspond to the UbA₅₂ gene, its processed pseudogenes (including λ UA4 and EHD5), and the UbA₈₀ subfamily gene(s) (and pseudogenes?), all of which contain a single coding unit of varying percentage similarity to the UbB cDNA probe.

The assignment of the UbB intron to chromosome bands 17p11.2 to 17p12 may provide a useful marker for linkage studies in a chromosomal region where few markers have yet been localised (O'Brien, 1987).

CHAPTER 7. ASSOCIATION OF THE REPETITIVE DNA SEQUENCES WITH
UBIQUITIN GENES AND PSEUDOGENES

The Association of the Repetitive DNA Sequences with the Ubiquitin
Gene

During the course of the study of the ubiquitin gene, several
pseudogenes were identified. The first of these was a sequence
located 5' to the ubiquitin gene. This sequence is an important
component of the 5' untranslated region of the ubiquitin
mRNA, distinguishing it from other mRNAs. The second
pseudogene was located 3' to the ubiquitin gene.

CHAPTER 7

ASSOCIATION OF Alu REPETITIVE DNA SEQUENCES
WITH UBIQUITIN GENES AND PSEUDOGENES

The Ubiquitin Gene and Pseudogenes
The Ubiquitin gene contains two Alu repetitive sequences of the
proximal region (Figure 1.4.1). The first sequence is a
complete Alu repeat in the opposite transcriptional
orientation to the Ubiquitin gene (Figure 1.4.2). It is
flanked by an imperfect 15bp direct repeat of sequence
80% similar to the consensus Alu sequence (Cairns et al.,
1987; see Figure 1.1.1). The proximal tail is composed of 5
repeats of the consensus Alu, with the third repeat missing
in some. This repetitive structure in the proximal tail is a
common feature of some Alu repeats (Cairns and Powell, 1984;
Cairns et al., 1987). The second Alu sequence is located
5' to the Ubiquitin gene and is in the same transcriptional
orientation as the Ubiquitin gene (Figure 1.4.3). It is

CHAPTER 7: ASSOCIATION OF Alu REPETITIVE DNA SEQUENCES WITH
UBIQUITIN GENES AND PSEUDOGENES

7.1 Analysis of Alu Repeats Associated with the UbB Gene and
Pseudogenes

During sequence analysis of the UbB gene and its processed pseudogenes, several members of the Alu family of repetitive DNA sequences were identified by similarity to the consensus Alu repeat sequence (Kariya *et al*, 1987). The human Alu sequence is an imperfect dimer composed of two direct repeats of a ~130bp monomer with a 31bp insertion in the second monomer, distinguishing it from the first (reviewed by Jelinek and Schmid, 1982). Each monomer terminates with an A-rich sequence. Approximately 500,000 Alu repeats are present in the human genome and thus should occur once every 5kb on average (Jelinek and Schmid, 1982).

The UbB gene EHB8 contains two Alu members upstream of its promoter region (Figure 3.6.1). The most upstream member is a complete Alu repeat in the opposite transcriptional orientation to the UbB gene (Figure 3.6.2, nt 109 to 402). It is flanked by an imperfect 16bp direct repeat and exhibits 90% similarity to the consensus Alu sequence (Kariya *et al*, 1987; see Figure 7.1.1). Its poly(A) tail is composed of four repeats of the tetramer AACA, with the third repeat mutated to GACA. This repetitive structure in the poly(A) tail is a common feature of some Alu repeats (Jelinek and Schmid, 1982; Kariya *et al*, 1987). The second UbB Alu repeat is truncated by 34bp at its 5' end and is in the same transcriptional orientation as the UbB gene (Figure 3.6.2, nt 510 to 776). It

is 91% similar to the consensus Alu, has a poly(A) tail consisting mainly of A residues, and appears to lack a flanking direct repeat (Figure 7.1.1), indicating that it may have been inserted as a full-length Alu and subsequently lost the 5' repeat and 34bp of the Alu unit. These two Alu's are arranged head-to-head and spaced by only 108bp, and thus represent a large inverted repeat located only 195bp upstream of the UbB gene heat shock elements, and 465 bp upstream of the TATA box.

Other Alu repeats were identified downstream of the EHB7 processed pseudogene (Figure 3.5.1). One complete Alu repeat totally occupies a 303bp EcoRI fragment 530bp downstream of the pseudogene (Figure 3.5.2, nt 1128 to 1421). In fact, the EcoRI sites are part of the flanking direct repeat, and provide a convenient Alu-specific hybridisation probe (see below). This Alu repeat is in the same transcriptional orientation as the pseudogene, is 88% similar to the consensus repeat, and possesses a simple poly(A) tail (Figure 7.1.1). A further 1.5kb downstream there is a 584bp Alu-homologous region in the opposite transcriptional orientation to the pseudogene (Figure 3.5.1). Sequence analysis indicates that this region arose from the integration of one Alu repeat into the central A-rich region of a pre-existing Alu member. Thus, the central Alu repeat is flanked by the 12bp direct repeat ATACAAAAAATT, while the whole 584bp region is flanked by a different repeat, AGGGARTGGGTCA (Figure 7.1.1). The central and composite outer Alu members respectively show 89 and 86% similarity to the consensus. The outer Alu has suffered a 30bp deletion in the second monomer, which notably

does not correspond to the 31bp insertion distinguishing it from the first monomer. While the central Alu has a simple poly(A) tail, that of the outer unit consists of the tetramer AATA repeated 11 times (Figure 7.1.1). As noted above, Alu poly(A) tails often exhibit this repeat structure, but not usually with as many or as faithful repeats (Kariya *et al*, 1987). The overall structure of this region is best seen by the dot-matrix comparison presented in Figure 7.1.2.

The 303bp EcoRI Alu-containing fragment described above was used to probe restriction digests of the 5 UbB genomic clones and their plasmid subclones to determine the approximate location of other Alu repeats relative to the UbB gene and pseudogenes. This analysis indicated the presence of Alu repeats: far upstream of the UbB gene; within the 1kb BamHI/PstI fragment immediately downstream of the UbB gene (Figure 3.6.1); the 1.9kb PstI/BglII fragment immediately upstream of the EHB4 pseudogene (Figure 3.3.1); and at the far downstream end of the EHD1 genomic insert (Figure 3.4.1). Additionally, no further Alu repeats were identified within either of the EHB6 or EHB7 plasmid subclones (Figure 3.5.1). The Alu repeat thus identified downstream of the UbB gene was partially characterised by sequence determination and is in the opposite transcription orientation with respect to the gene (Figure 3.6.1) The PstI site at the downstream end of the pB8.3 subclone occurs 45bp into the first monomer of the Alu repeat, and thus its complete sequence was not determined. This Alu has not been included in Figure 7.1.1. The other hybridisation-located Alu repeats were not

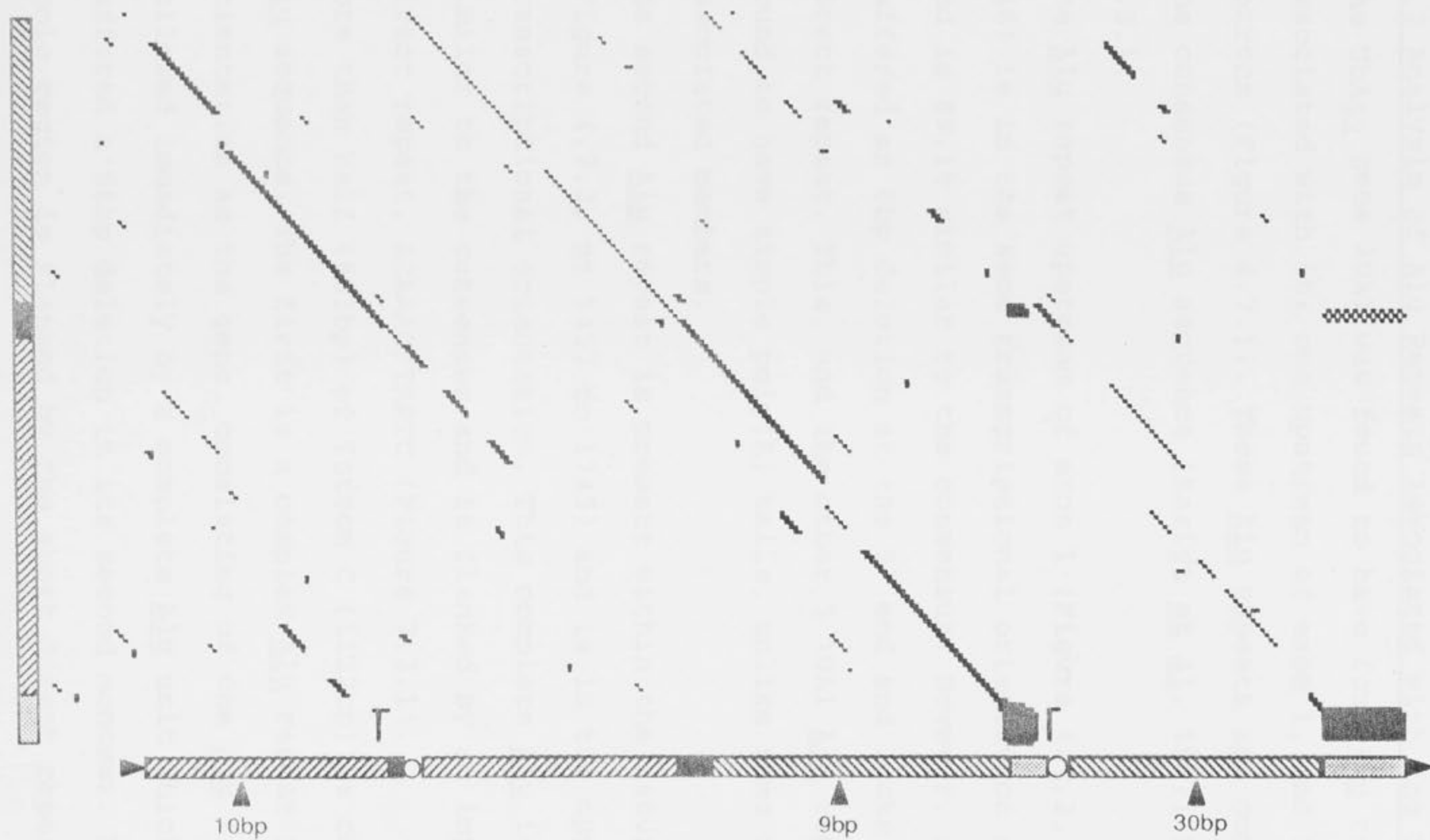






Figure 7.1.2: Dot-Matrix Analysis of the EHB7 Far-Downstream *Alu* Repeat Region.

The consensus *Alu* repeat (vertical axis) is compared to the complement of nt 41 to 680, Figure 3.5.2 (lower portion) by dot-matrix analysis using a window of 10bp of which 8 must match to register a dot. The consensus *Alu* repeat and the EHB7 region are represented schematically. Horizontal arrows and circles represent flanking direct repeats, while vertical arrows point to deletions (in bp) resulting in discontinuities in the dot-matrix. The *Alu* first monomer, second monomer, central A-rich and terminal A-rich regions are respectively shown by:    

characterised by sequence determination, and thus nothing is known of their number, orientation or similarity to the consensus repeat sequence.

7.2 Analysis of Alu Repeats Associated with the UbA₅₂ Gene

The UbA₅₂ gene λ UA1 was found to have four Alu repeats associated with it: one upstream of exon 1, and 3 within its introns (Figure 4.7.1). These Alu repeats are compared with the consensus Alu sequence (Kariya *et al*, 1987) in Figure 7.2.1.

The Alu repeat upstream of exon 1 (Figure 4.7.2, nt 156 to 448) is in the same transcriptional orientation as the gene and is 89.1% similar to the consensus. However, it has suffered an 8bp deletion at the 5' end and lacks a flanking direct repeat. This, and the other 3 λ UA1 Alu repeats were found to have simple poly(A) tails, unlike some of the UbB-associated members.

The second Alu repeat is present within the 1400bp intron A (Figure 4.7.2, nt 1457 to 1745) and is in the opposite transcriptional orientation. This complete Alu is 86.9% similar to the consensus and is flanked by an imperfect 10bp direct repeat, ACAAAG/TARTC (Figure 7.2.1).

More than half (653bp) of Intron C (1122bp) is composed of Alu sequence. The first is a complex Alu repeat in the same orientation as the gene, consisting of one Alu first monomer, followed immediately by a complete Alu unit which has suffered a 68bp deletion in its second monomer. However, the whole region is flanked by the short direct repeat ATAATG, suggesting its insertion as one entity. The first monomer

```

      10          20          30          40          50          60
-----GGCCGGGCGCGGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGG
      .A.....T.....
ACAAAGAATC.A.....A.G.....T.....A...
ATAATGG.A.T...A.A.....T.....G.....
      A..T...T..A.....G.....C.....
GACTGAAGTCTA...G...A.CA.....T.....AA...A

      70          80          90          100          110          120
CCGAGGTGGGTGGATCACCTGAGGTCAGGAGTTCAAGACCAGCCTGGCCAACATGGTGAA
..A...C...C.....GC.....TG.....A.....A...
.....CA.AC.....TT.....A.....TG.....T.....
.TT...C.....A.....G.....
.T...C...CA.....-A.....A..A.TG...T..T.A..A.....
-----G.A.....

      130          140          150          160          170          180
ACCCCGTCTCTACTAAAAATACAAAA---TTAGCCGGGCGTGGTGGCGCGCGCCTGTAA
.....G...-G...A...A.....T...A.A.T.....
..T.....C.....C.A.T.....C.AG.....A...
.T...T...T.....C
.T...A.....TA...A.....T...T.....G
....T.....AAA...T.....G.T.....

      190          200          210          220          230          240
TCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGTGGAGGTTGC
..G.....T.....A.....C.....
.....C...A.....C.G.....A...
.....A.....A.....T.....T.....A-----
.....G.....AC.....CA.

      250          260          270          280          290          300
AGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGAC-AGAGCGAGACTCCGTCTC
G.....T.....A.....A.G.....A.....
.....C..A.....A.....T.A.....A.....A...G
-----
.....T.....T.....TA.....T.....

      310          320          330
-----
AAAAAAAAAAAAAAAAAAAAA UBA1 ALU#1 % MATCH
AAAAAAAAATACAAATAGTC UBA1 ALU#2 89.1
AAAAAAAAAAAGTCATAATG UBA1 ALU#3 1ST UNIT 86.9
ACAAAAAAGGACTGAATTCT UBA1 ALU#3 2ND UNIT 88.9
ACAAAAAAGGACTGAATTCT UBA1 ALU#4 84.1
ACAAAAAAGGACTGAATTCT UBA1 ALU#4 87.6

```

Figure 7.2.1: Uba52-Associated Alu Repeats.

Alu repeats are identified as in Figure 4.7.2. Other details are as in Figure 7.1.1. Note that the Alu#3 direct repeat flanks the composite 1st and 2nd unit Alu sequence.

unit is 88.9% similar to the consensus, while the second unit shows 84.1% homology. The structure of this Alu region is best seen by dot-matrix comparison (Figure 7.2.2). The second intron C Alu member is in the opposite orientation to the gene, is 87.6% similar to the consensus, and is flanked by a 14bp direct repeat GCACTGAAG/TTCT (Figure 7.2.1). It has also suffered a large deletion; 38bp from the first monomer. Most notably, these two intron C Alu repeats are spaced by only 77bp and thus form a large inverted repeat within the intron.

No Alu-homologous sequences were observed within either of the sequenced regions of the processed pseudogenes EHD5 or λ UA4. However, unlike the UbB genomic clones, the UbA52 clones were not further analysed by hybridisation with an Alu-specific probe, and thus the proximity of Alu repeats to these pseudogenes is unknown.

7.3 Association of Alu Repetitive DNA Sequences with the UbB and UbA52 Subfamilies

As described above, several Alu repeats are located in the vicinity of the UbB gene and some of its processed pseudogenes, while four repeats are present within and around the UbA52 gene. The Alu repeat family comprises the most abundant family of middle repetitive DNA sequences in the human genome with one member occurring every 5kb on average (see Jelinek and Schmid, 1982; Kariya *et al*, 1987). Thus by chance, some genes and pseudogenes will be in close proximity to an Alu member, and so the significance of this association maybe no more than a random event. Many Alu sequences co-

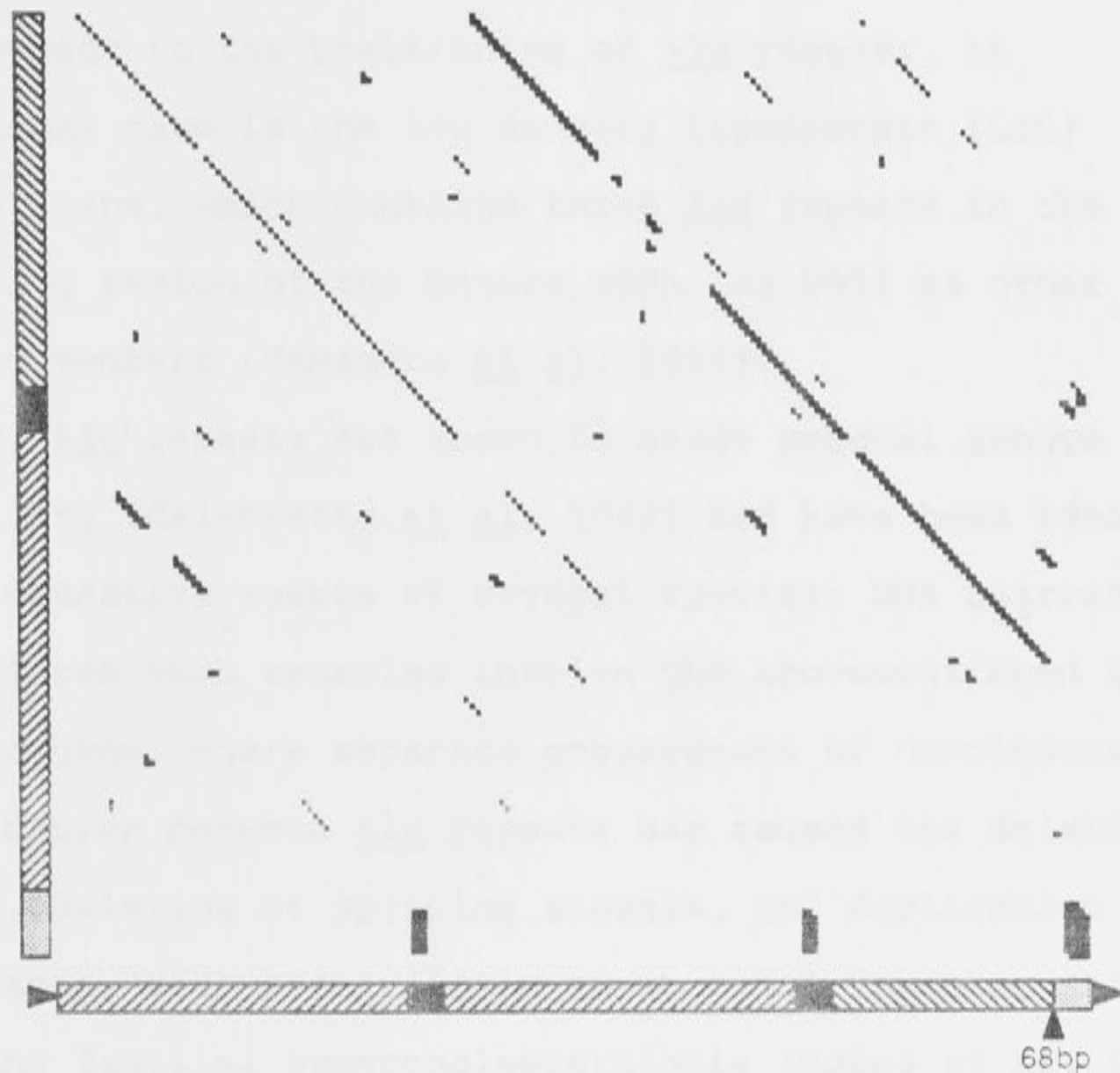


Figure 7.2.2: Dot-Matrix Analysis of the Intron C Alu Repeat.

The consensus Alu repeat (vertical axis) is compared to nt 2961 to 3360 of the UbA52 gene intron C (Figure 4.7.2) by dot-matrix analysis using a window of 10bp of which 8 must match to register a dot. Horizontal arrows represent the flanking direct repeat, while the vertical arrow points to a 68bp deletion causing the discontinuity in the dot-matrix. Shaded boxes represent the various Alu repeat regions as described in the legend to Figure 7.1.2.

exist stably within transcriptionally active genes: for example ten lie within one intron of the β -tubulin gene (Lee *et al*, 1984) while two introns of the prealbumin gene contain Alu members (Sasaki *et al*, 1985). Thus these Alu members are present in the pre-mRNA and are removed during mRNA maturation. The U_BA₅₂ gene is thus similar to these genes with respect to the positioning of Alu repeats. An exceptional case is the low density lipoprotein (LDL) receptor gene, which contains three Alu repeats in the 3' non-coding region of the mature mRNA, as well as other intronic members (Yamamoto *et al*, 1984).

However, Alu repeats are known to cause general genome instability (Calabretta *et al*, 1982) and have been identified as the causative agents of several specific DNA rearrangements. Three such examples involve the abovementioned LDL receptor gene, where separate occurrences of homologous recombination between Alu repeats has caused the deletion of an exon, deletion of splicing signals, and duplication of seven exons, each resulting in an abnormal receptor and producing familial hypercholesterolemia (Hobbs *et al*, 1986; Lehrman *et al*, 1987a, 1987b). However, considering the presumed essential function of the U_BA₅₂ gene and the likelihood that it may be the only transcriptionally active member of its subfamily suggests that any similar Alu-mediated rearrangements of the U_BA₅₂ gene may result in a non-viable phenotype.

A second type of rearrangement mediated by Alu-Alu recombination is gross chromosomal interchange. Rouyer *et al* (1987) report a case of human XX maleness resulting from an

interchange of terminal parts of the X and Y chromosomes, with the breakpoint within an Alu repeat. These authors also speculate on the general involvement of Alu-Alu recombination in other chromosomal rearrangements such as translocations and inversions.

The effect of the UbB-proximal Alu repeats on the stability of the UbB locus is unknown. However, it is notable that the gene is flanked by Alu repeats and that it has recently been duplicated, suggesting a possible mechanism for duplication. As mentioned above, precedents exist for Alu-mediated DNA duplication (eg. Lehrman et al, 1987b). The relative proximity of Alu repeats to some of the UbB processed pseudogenes may stem from mechanistic requirements of the insertion of both types of elements into the genome. An analysis of sequences flanking Alu repeats by Kariya et al (1987) suggests that Alu sequences are inserted basically at random but preferentially into A/T rich regions. As processed pseudogenes and Alu repeats are inserted by similar mechanisms (Weiner et al, 1986), both may be targeted to the same genomic region of favourable sequence composition.

The variance observed between the Alu repeats and the consensus is due to a total of 212 transitions, 56 transversions and 31 deletion/insertion events. Kariya et al (1987) noted that Alu nucleotide substitutions consisted of 78% transitions and 28% transversions, very similar values to those observed above (79 and 21% respectively).

CHAPTER 8

This study has been the most comprehensive of the
available of the human voice in the field. There are
and a group of related studies which are
members of the field and the field is
available in the field and the field is
approximately 2 hours of material in the field
with carefully selected material in the field
transcriptionally active and the field is
at least in the field and the field is

CHAPTER 8

GENERAL DISCUSSION AND CONCLUSIONS

The study has been the most comprehensive of the
available of the human voice in the field. There are
and a group of related studies which are
members of the field and the field is
available in the field and the field is
approximately 2 hours of material in the field
with carefully selected material in the field
transcriptionally active and the field is
at least in the field and the field is

The study has been the most comprehensive of the
available of the human voice in the field. There are
and a group of related studies which are
members of the field and the field is
available in the field and the field is
approximately 2 hours of material in the field
with carefully selected material in the field
transcriptionally active and the field is
at least in the field and the field is

CHAPTER 8: GENERAL DISCUSSION AND CONCLUSIONS8.1 The Human Ubiquitin Gene Family

This thesis describes the extensive characterisation of two subfamilies of the human ubiquitin gene family. Chapters 3 and 4 present the detailed sequence analysis of several members of the UbB and UbA₅₂ subfamilies respectively. Both subfamilies are quite large, with each containing approximately 9 members detectable by cross-hybridisation with subfamily-specific probes. However, only one transcriptionally active gene was detected in each subfamily, at least in the tissues studied. In fact, these subfamilies appear to be comprised primarily of processed pseudogenes, based on the general lack of hybridisation with intronic sequences from their respective transcriptionally active gene. The single exception is an apparently duplicated UbB gene, which from the limited evidence available does not appear to be transcriptionally active in liver tissue, but may function in other tissues or be a non-processed pseudogene. Gene families comprised mainly of processed pseudogenes are relatively common in the higher vertebrates.

The human UbC gene was also studied by hybridisation analysis following the observation of an apparent UbC mRNA length polymorphism (mRLP). Chapter 5 describes the subsequent identification of a second mRLP, the correlation of both mRLPs with a restriction fragment length polymorphism (RFLP) at the UbC locus, and the Mendelian inheritance of both mRLP and RFLP. The most likely explanation for these polymorphisms

is a variation in the number of coding units at the UbC locus, resulting from unequal crossover events.

Chapter 6 describes in situ hybridisation aimed at determining the chromosomal location of the human ubiquitin genes. This work unequivocally localised the duplicated UbB gene to the short arm of chromosome 17, most likely over bands 17p11.2 to 17p12. Although several other chromosomal locations were identified by hybridisation with a probe capable of detecting all ubiquitin loci, assignment cannot be made without further analysis with gene- and subfamily-specific probes.

The only subfamily not directly investigated in this study is the UbA₈₀ subfamily. Presently our knowledge of this subfamily is restricted to a partial cDNA clone (Lund et al, 1985). However, there is evidence that there are approximately 7 UbA₈₀ subfamily genomic loci (see Schlesinger and Bond, 1987). In addition, Wiborg et al (1985) have reported the flanking regions of two apparent UbA₈₀ pseudogenes which have suffered nucleotide deletions compared to the UbA₈₀ cDNA. These genomic sequences are not interrupted by introns and thus may be processed pseudogenes. While the presence or absence of introns in the UbA₈₀ gene(s) is unknown, it is quite possible that the UbA₈₀ subfamily is similar in composition to the UbB and UbA₅₂ subfamilies, containing several processed pseudogenes.

Finally, Chapter 7 describes several of the members of the Alu family of repetitive DNA sequences which were found in

association with the UbA₅₂ gene and the UbB gene and pseudogenes.

8.2 Ubiquitin Gene Evolution

The characterisation of the human ubiquitin gene subfamilies described herein adds significant new information to that already accumulated on ubiquitin gene structure. Most importantly, a representative from each gene type has now been identified for an organism other than yeast. That these two organisms differing by vast evolutionary distances contain very similarly structured gene families is testament to the remarkable conservation of the ubiquitin system. This raises the fact that these gene structures must predate the human/yeast divergence, implying that the ubiquitin system was present very early in eukaryotic evolution.

A second point relevant to ubiquitin gene evolution is the presence of at least four introns in the human UbA₅₂ gene, three of which interrupt the coding region. The coding region of the homologous yeast genes are also interrupted by an intron (Ozkaynak et al, 1987). Conversely, no introns interrupt the coding region of any known polyubiquitin gene. These observations encourage the following speculations on ubiquitin gene evolution. The UbA₅₂ exon structure suggests that this gene was constructed by recruitment of pre-existing exons encoding functional domains. As such, this may reflect the formation of the original ubiquitin gene. The observation that the polyubiquitin gene is not essential for normal cell growth (Finley et al, 1987) suggests that it may have appeared at a later date and then been recruited for the

stress response. The intronless nature and putative later appearance of the polyubiquitin gene could both be explained by a single event - the generation of a functional retroposon from a ubiquitin-tail fusion gene. This event would require the generation of the retroposon from an aberrantly long transcript including the gene's promoter region, or alternatively, the fortuitous insertion of the retroposon adjacent to a promoter sequence. This second possibility is interesting in light of the intron in the 5' non-coding regions of the human UbB and chicken UbI polyubiquitin genes, perhaps arising from retroposon insertion near a promoter/NCR exon domain. A third possibility arises from the hypothesis of Fink (1987), who speculates that most yeast genes may in fact be functional retroposons, as a result of homologous recombination between each gene and a reverse transcribed cDNA of its transcript. This hypothesis explains the observed general lack of introns in yeast, and the asymmetric location of the few known introns at the extreme 5' ends of the gene (see Fink, 1987), as is the case with the yeast UBI1 and UBI2 ubiquitin genes. Thus the intronless retroposon would replace the intron-bearing gene without disturbing the promoter region.

The tail coding region could then be inactivated by a mutation (substitution, insertion or deletion) leading to a stop codon one aa downstream of the end of the ubiquitin coding unit. This event would explain the single aa extension observed in the present day polyubiquitin genes. The tail coding region would then degenerate to become the polyubiquitin 3' non-coding region. Such a process would be

presumably detrimental to the organism if the retroposon had integrated by the hypothesis of Fink (see above). However, if recombination occurred with one gene of a gene pair (eg the yeast UBI1/UBI2 gene pair), the other gene would still produce the intact fusion protein.

These speculated events would generate a ubiquitin gene with a single intronless coding unit, presumably the precursor of present-day polyubiquitin genes. However, the generation of a polyubiquitin gene from a monoubiquitin precursor is harder to explain. The present-day polyubiquitin genes can be readily explained by unequal crossover events of a precursor two coding-unit ubiquitin gene, whereas the step from the single to the double coding unit gene is less obvious, and necessitates a precise crossover between two single coding unit alleles at the end of one coding unit and the start of the other, with no repetitive structure to promote misalignment.

This speculation is based on the assumption that introns were present in the original genes and have been lost from simpler organisms during evolution. If one subscribes to the alternate view that higher organisms acquired introns late in evolution, then a reverse explanation could apply: a single coding unit ubiquitin gene "acquiring" a tail coding region through a mutation inactivating the stop codon, and a subsequent invasion by introns. However, it is highly unlikely that the tail protein thus acquired would contain the DNA binding and nuclear translocation signals present in the current tail proteins, and even less unlikely that this would occur for the two distinct tails. Current evidence

however, supports the view that ancestral genes contained introns (Gilbert et al, 1986), and thus the former hypothesis explaining the two ubiquitin gene structural types appears at present to be the more probable.

8.3 Future Studies

This study has provided detailed structural analysis of one transcriptionally active gene from each of the human UbB and UbA₅₂ subfamilies, as well as several processed pseudogenes of each. In addition, evidence for a polymorphism in the number of coding unit repeats at the UbC locus has been obtained. However, further work is required in several areas to completely characterise each subfamily.

First, each subfamily appears to contain other processed pseudogenes in addition to those characterised, which need to be isolated and sequenced for completeness. Suitable probes for this isolation would be the 3' NCR (UbB) and the tail/3' NCR (UbA₅₂) fragments already employed, followed by screening of positives with the gene intron probes to remove clones containing the already characterised genes.

Second, the UbB gene appears to be duplicated, and in situ hybridisation studies suggest that both genes are relatively close together. A suitable probe for the isolation of the duplicated gene would be the UbB intron, followed by restriction mapping of positives to remove the presently known UbB gene. In addition, overlapping genomic clones could be obtained to determine the distance between the duplicated genes. An alternative method would be to screen a genomic sub-library constructed from ~2.3kb BamHI fragments. However,

this approach would not be informative as to the proximity of the two genes. Furthermore, it is not known if the duplicated gene is transcriptionally active. Sequencing of the duplicated gene and UbB cDNA clones isolated from libraries constructed from tissues other than liver may allow the determination of its transcriptional activity.

Third, while the UbB gene contains possible heat shock promoters, it is clearly transcribed at relatively high levels in placenta and lymphocytes. A controlled study of its transcription in human cell lines under various stress conditions should reveal if it is in fact a heat shock gene.

Fourth, the 5' NCR and promoter region of the UbA₅₂ gene have not been fully characterised. The mRNA start site could be determined from a combination of sequencing both full-length cDNA clones and the primer-extended products described in Chapter 4.7.2, and comparing these sequences to the known gene sequence. Genomic fragments extending further upstream from the present λ UA1 clone may be required if the determined sequences do not appear in λ UA1. Determination of the mRNA start site(s) would then allow an evaluation of the activity of the putative TATA and Sp1 promoter sites observed upstream of the known UbA₅₂ exons.

Fifth, while the UbB gene has been localised to chromosome band 17p11.2 to 17p12, the location of other ubiquitin genes is unknown. The UbA₅₂ intron A probe is specific for this gene and should allow its unequivocal localisation: these studies are presently underway. Similar localisations of the UbC and UbA₈₀ genes would require gene-specific probes derived from the structural gene, which are not yet available

for the latter subfamily. Localisation of the UbC gene would determine if it is clustered with the UbB genes on chromosome 17.

Sixth, the putative UbC coding unit number polymorphism requires further population studies to accurately determine the allele frequencies. The nature of the polymorphism could be determined by constructing genomic libraries from polymorphic individuals (eg the 8,7 heterozygote and 8,8 homozygote) and isolating and characterising the UbC alleles. These studies may also reveal the site of the proposed crossover events. A further technique would be to probe genomic DNA partially digested with a restriction enzyme cleaving only once within each coding unit (eg BglII) to determine the number of coding units at the UbC locus as performed with Xenopus by Dworkin-Rastl et al (1984). This analysis would only reveal the longest UbC allele which would mask the patterns from shorter alleles, and thus would be most informative with the 8,7 and 8,8 phenotypes.

Finally, the UbA₈₀ subfamily (not investigated in this study) remains the least well characterised, with our knowledge limited to a partial cDNA sequence. However, this cDNA provides a subfamily-specific probe, which is presumably being used to isolate the structural gene(s). Once isolated, this gene will allow structural comparison with the UbA₅₂ gene and the relevant yeast genes, especially with respect to the positioning of introns, if any.

8.4 Conclusions

The results and discussion constituting this thesis have fulfilled the initial aim of the structural characterisation of the human UbB and UbA₅₂ subfamilies, and have also provided an interesting insight into the UbC locus with respect to a putative coding unit number polymorphism. These results also reinforce the power of the application of molecular biological techniques to the study of the ubiquitin system in general, which has resulted in the identification of several properties previously unknown: the novel polyubiquitin gene structure and its implications for ubiquitin synthesis; the fused ubiquitin-tail gene, again a novel structure with functional implications; the identification of ubiquitin as a heat-shock protein and the location of heat-shock promoters upstream of some polyubiquitin genes; the identification of other ubiquitin-protein conjugates, such as cell surface receptors; and the identification of previously known proteins (ie yeast RAD6 and possibly CDC34) as functional components of the ubiquitin system.

It is hoped that the contribution of the results discussed in this thesis will lead to a fuller understanding of ubiquitin gene structure, regulation, and evolution, which in turn will aid in the functional characterisation of the ubiquitin system, surely one of the most remarkable components of the eukaryotic organism.

... (1981) ...

... (1982) ...

... (1983) ...

... (1984) ...

... (1985) ...

... (1986) ...

... (1987) ...

... (1988) ...

... (1989) ...

... (1990) ...

REFERENCES

... (1981) ...

... (1982) ...

... (1983) ...

... (1984) ...

... (1985) ...

... (1986) ...

... (1987) ...

... (1988) ...

... (1989) ...

... (1990) ...

- Arribas, C., Sampedro, J. and Izquierdo, M. (1986). The ubiquitin genes in *D. melanogaster*: transcription and polymorphism. *Biochim. Biophys. Acta* 868, 119-127.
- Aviv, H. and Leder, P. (1972). Purification of biologically active globin mRNA by chromatography on oligo thymidylic acid cellulose. *Proc. Natl. Acad. Sci. USA* 69, 1408-1412.
- Bachmair, A., Finley, D. and Varshavsky, A. (1986). *In vivo* half-life of a protein is a function of its amino-terminal residue. *Science* 234, 179-186.
- Baker, R.T. and Board, P.G. (1987a). The human ubiquitin gene family: structure of a gene and pseudogenes from the UbB subfamily. *Nucl. Acids Res.* 15, 443-463.
- Baker, R.T. and Board, P.G. (1987b). Nucleotide sequence of a human ubiquitin UbB processed pseudogene. *Nucl. Acids Res.* 15, 4352.
- Baker, R.T. and Board, P.G. (1988). An improved method for mapping recombinant λ phage clones. *Nucl. Acids Res.* 16, 1198.
- Ball, E., Karlik, C.C., Beall, C.J., Saviile, D.L., Sparrow, J.C., Bullard, B. and Fyrberg, E.A. (1987). Arthrin, a myofibrillar protein of insect flight muscle, is an actin-ubiquitin conjugate. *Cell* 51, 221-228.
- Barsoum, J. and Varshavsky, A. (1985). Preferential localization of variant nucleosomes near the 5'-end of the mouse dihydrofolate reductase gene. *J. Biol. Chem.* 260, 7688-7697.
- Basler, K., Oesch, B., Scott, M., Westaway, D., Walchli, M., Groth, D.F., McKinley, M.P., Prusiner, S.B. and Weissman, C. (1986). Scrapie and cellular PrP isoforms are encoded by the same chromosomal gene. *Cell* 46, 417-428.
- Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980). The ovalbumin gene - sequence of putative control regions. *Nucl. Acids Res.* 8, 127-142.
- Benton, W.D. and Davis, R.W. (1977). Screening λ gt recombinant clones by hybridisation to single plaques *in situ*. *Science* 196, 180-181.
- Berg, J.M. (1986). Potential metal-binding domains in nucleic acid binding proteins. *Science* 232, 485-487.
- Berget, S.M. (1984). Are U4 small nuclear ribonucleoproteins involved in polyadenylation? *Nature* 309, 179-182.
- Berk, A.J. and Sharp, P.A. (1977). Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease digested hybrids. *Cell* 12, 721-732.

- Bienz, M. and Pelham, H.R.B. (1987). Mechanisms of heat-shock gene activation in higher eukaryotes. *Adv. Genet.* 24, 31-72.
- Board, P.G. (1984). Genetic heterogeneity of the B subunit of coagulation factor XIII: resolution of type 2. *Ann. Hum. Genet.* 48, 223-228.
- Board, P.G. and Webb, G.C. (1987). Isolation of a cDNA clone and localisation of human glutathione S-transferase 2 genes to chromosome band 6p12. *Proc. Natl. Acad. Sci. USA* 84, 2377-2381.
- Bond, U. and Schlesinger, M.J. (1985). Ubiquitin is a heat shock protein in chicken embryo fibroblasts. *Mol. Cell. Biol.* 5, 949-956.
- Bond, U. and Schlesinger, M.J. (1986). The chicken ubiquitin gene contains a heat shock promoter and expresses an unstable mRNA in heat-shocked cells. *Mol. Cell. Biol.* 6, 4602-4610.
- Bond, U., Agell, N., Haas, A., Radman, K. and Schlesinger, M.J. (1988). Ubiquitin in stressed chicken embryo fibroblasts. *J. Biol. Chem.* 263, 2384-2388.
- Bond, U. and Schlesinger, M.J. (1987). Heat-shock proteins and development. *Adv. Genet.* 24, 1-29.
- Bornstein, P. and McKay, J. (1988). The first intron of the $\alpha 1(I)$ collagen gene contains several transcriptional regulatory elements. *J. Biol. Chem.* 263, 1603-1606.
- Bornstein, P., McKay, J., Morishima, J.K., Devarayalu, S. and Gelinas, R.E. (1987). Regulatory elements in the first intron contribute to transcriptional control of the human $\alpha 1(I)$ collagen gene. *Proc. Natl. Acad. Sci. USA* 84, 8869-8873.
- Breathnach, R. and Chambon, P. (1981). Organization and expression of eukaryotic split genes coding for proteins. *Ann. Rev. Biochem.* 50, 349-383.
- Buckle, V.J. and Craig, I.W. (1986). *In situ* hybridization in Human Genetic Diseases: a practical approach. K.E. Davies (ed). IRL Press, Oxford, UK. pp85-100.
- Burdon, R.H. (1986). Heat shock and the heat shock proteins. *Biochem. J.* 240, 313-324.
- Burke, J.F. (1984). High-sensitivity S1 mapping with single-stranded [32 P]DNA probes synthesized from bacteriophage M13mp templates. *Gene* 30, 63-68.
- Calabretta, B., Robberson, D., Barrera-Saldana, H.A., Lambrou, T.L. and Saunders, G.F. (1982). Genome instability in a region of human DNA enriched in Alu repeat sequences. *Nature* 296, 219-225.

- Cary, P.D., King, D.S., Crane-Robinson, C., Bradbury, E.M., Rabbini, A., Goodwin, G.H. and Johns, E.W. (1980). Structural studies on two high-mobility-group proteins from calf thymus, HMG-14 and HMG-20 (ubiquitin), and their interaction with DNA. *Eur. J. Biochem.* 112, 577-580.
- Casadaban, M.J. and Cohen, S.N. (1980). Analysis of gene control signals by DNA fusion and cloning in *Escherichia coli*. *J. Mol. Biol.* 138, 179-207.
- Chen, S.-H., Habib, G., Yang, C.-Y., Gu, Z.-W., Lee, B.R., Weng, S.-A., Silberman, S.R., Cai, S.-J., Deslypere, J.P., Rosseneu, M., Gotto, A.M., Li, W.-H. and Chan, L. (1987) Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* 238, 363-366.
- Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979). Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* 18, 5294-5299.
- Chomczynski, P. and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* 162, 156-159.
- Ciechanover, A. (1987). Regulation of the ubiquitin-mediated proteolytic pathway: role of the substrate α -NH₂ group and of transfer RNA. *J. Cell. Biochem.* 34, 81-100.
- Ciechanover, A., Elias, S., Heller, H., Ferber, S. and Hershko, A. (1980). Characterisation of the heat-stable polypeptide of the ATP-dependent proteolytic system from reticulocytes. *J. Biol. Chem.* 255, 7525-7528.
- Ciechanover, A., Elias, S., Heller, H. and Hershko, A. (1982). "Covalent affinity" purification of ubiquitin-activating enzyme. *J. Biol. Chem.* 257, 2537-2542.
- Ciechanover, A., Finley, D. and Varshavsky, A. (1984). Ubiquitin dependence of selective protein degradation demonstrated in the mammalian cell cycle mutant ts85. *Cell* 37, 57-66.
- Ciechanover, A., Heller, H., Katz-Etzion, R. and Hershko, A. (1981). Activation of the heat-stable polypeptide of the ATP-dependent proteolytic system. *Proc. Natl. Acad. Sci. USA* 78, 761-765.
- Ciechanover, A., Wolin, S.L., Steitz, J.A. and Losish, H.F. (1985). Transfer RNA is an essential component of the ubiquitin- and ATP-dependent proteolytic system. *Proc. Natl. Acad. Sci. USA* 82, 1341-1345.
- Clark, B.D., Collins, K.L., Gandy, M.S., Webb, A.C. and Auron, P.E. (1986). Genomic sequences for human prointerleukin 1 beta: possible evolution from a reverse transcribed prointerleukin 1 alpha gene. *Nucl. Acids Res.* 14, 7897-7914.

- Costanzo, F., Colombo, M., Staempfli, S., Santoro, C., Marone, M., Frank, R., Delius, H. and Cortese, R. (1986). Structure of gene and pseudogenes of human apoferritin H. *Nucl. Acids Res.* 14, 721-736.
- Coulson, F.S.A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A. (1980). Cloning in single-stranded Bacteriophage as an aid to rapid DNA sequencing. *J. Mol. Biol.* 143, 161-178.
- Cox, M.J., Haas, A.L. and Wilkinson, K.D. (1986). Role of ubiquitin conformations in the specificity of protein degradation: iodinated derivatives with altered conformations and activities. *Arch. Biochem. Biophys.* 250, 400-409.
- Dagert, M. and Ehrlich, S.D. (1977). Prolonged incubation in calcium chloride improves the competence of *Escherichia coli* cells. *Gene* 6, 23-28.
- Di Stefano, D.L. and Wand, A.J. (1987). Two-dimensional ^1H NMR study of human ubiquitin: a main chain directed assignment and structure analysis. *Biochemistry* 26, 7272-7281.
- Dorlon, T.A. (1986). Practical approaches to in situ hybridisation. *Karyogram* 12, 3-10.
- Duerksen-Hughes, P.J., Xu, X. and Wilkinson, K.D. (1987). Structure and function of ubiquitin: evidence for differential interactions of arginine-74 with the activating enzyme and the proteases of ATP-dependent proteolysis. *Biochemistry* 26, 6980-6987.
- Dworkin-Rastl, E., Shrutkowski, A. and Dworkin, M.B. (1984). Multiple ubiquitin mRNAs during *Xenopus laevis* development contain tandem repeats of the 76 amino acid coding sequence. *Cell* 39, 321-325.
- Ecker, D.J., Khan, M.I., Marsh, J., Butt, T.R. and Crooke, S.T. (1987a). Chemical synthesis and expression of a cassette adapted ubiquitin gene. *J. Biol. Chem.* 262, 3524-3527.
- Ecker, D.J., Butt, T.R., March, J., Sternberg, E.J., Margolis, N., Monia, B.P., Jonnalagadda, S., Khan, M.I., Weber, P.L., Mueller, L. and Crooke, S.T. (1987b). Gene synthesis, expression, structures, and functional activities of site-specific mutants of ubiquitin. *J. Biol. Chem.* 262, 14213-14221.
- Eckert, R.L. and Green, H. (1986) Structure and evolution of the human involucrin gene. *Cell* 46, 583-589.
- Efstratiadis, A., Kafatos, F.C., Maxam, A.M. and Maniatis, T. (1976). Enzymatic in vitro synthesis of globin genes. *Cell* 7, 279-288.

- Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980). The structure and evolution of the human β -globin gene family. *Cell* 21, 653-668.
- Einspanier, R., Sharma, H.S. and Scheit, K.H. (1987a). An mRNA encoding poly-ubiquitin in porcine corpus luteum: identification by cDNA cloning and sequencing. *DNA* 6, 395-400.
- Einspanier, R., Sharma, H.S. and Scheit, K.H. (1987b). Cloning and sequence analysis of a cDNA encoding poly-ubiquitin in human ovarian granulosa cells. *Biochem. Biophys. Res. Commun.* 147, 581-587.
- Farrel, P., Sharp, S. and DeFranca, D. (1981). *Laboratory Manual*, Yale University, Connecticut, USA.
- Ferber, S. and Ciechanover, A. (1986). Transfer RNA is required for conjugation of ubiquitin to selective substrates of the ubiquitin- and ATP-dependent proteolytic system. *J. Biol. Chem.* 261, 3128-3134.
- Fields, S. and Winter, G. (1981). Nucleotide-sequences heterogeneity and sequence rearrangements in influenza virus cDNA. *Gene* 15, 207-214.
- Fink, G.R. (1987). Pseudogenes in yeast? *Cell* 49, 5-6.
- Finley, D., Ciechanover, A. and Varshavsky, A. (1984). Thermolability of ubiquitin-activating enzyme from the mammalian cell cycle mutant ts85. *Cell* 37, 43-55.
- Finley, D. and Varshavsky, A. (1985). The ubiquitin system: functions and mechanisms. *Trends. Biochem. Sci.* 10, 343-347.
- Finley, D., Ozkaynak, E. and Varshavsky, A. (1987). The yeast polyubiquitin gene is essential for resistance to high temperatures, starvation, and other stresses. *Cell* 48, 1035-1046.
- Frischauf, A.-M., Lehrach, H., Poustka, A. and Murray, N. (1983). Lambda replacement vectors carrying polylinker sequences. *J. Mol. Biol.* 170, 827-842.
- Fusauchi, Y. and Iwai, K. (1985). Tetrahymena ubiquitin-histone conjugate uH2A. Isolation and structural analysis. *J. Biochem.* 97, 1467-1476.
- Gallatin, W.M., Weissman, I.L. and Butcher, E.C. (1983). A cell-surface molecule involved in organ-specific homing of lymphocytes. *Nature* 304, 30-34.
- Gallatin, M., St John, T.P., Siegelman, M., Reichert, R., Butcher, E.C. and Weissman, I.L. (1986). Lymphocyte homing receptors. *Cell* 44, 673-680.

- Gausung, K. and Barkardottir, R. (1986). Structure and expression of ubiquitin genes in higher plants. *Eur. J. Biochem.* 158, 57-62.
- Gavilanes, J.G., Gonzalez de Buitrago, G., Perez-Castells, R. and Rodriguez, R. (1982). Isolation, characterization, and amino acid sequence of a ubiquitin-like protein from insect eggs. *J. Biol. Chem.* 257, 10267-10270.
- Gianelli, F., Choo, K.H., Rees, D.J.G., Body, Y., Pizza, C.R. and Brownlee, G.G. (1983). Gene deletions in patients with haemophilia B and anti-factor IX antibodies. *Nature* 303, 181-182.
- Giebel, L.B., Dworniczak, B.P. and Bautz, E.F.K. (1987). Nucleotide sequence of a processed human hsc70 pseudogene. *Nucl. Acids Res.* 15, 9605.
- Gilbert, W., Marchionni, M. and McKnight, G. (1986). On the antiquity of introns. *Cell* 46, 151-154.
- Giorda, R. and Ennis, H.L. (1987). Structure of two developmentally regulated *Dictyostelium discoideum* ubiquitin genes. *Mol. Cell. Biol.* 6, 2097-2103.
- Goldknopf, I.L. and Busch, H. (1977). Isopeptide linkage between nonhistone and histone 2A polypeptides of chromosomal conjugate-protein A24. *Proc. Natl. Acad. Sci. USA* 74, 864-868.
- Goldstein, G., Scheid, M., Hammerling, U., Boyse, E.A., Schlesinger, D.H. and Niall, H.D. (1975). Isolation of a polypeptide that has lymphocyte-differentiating properties and is probably represented universally in living cells. *Proc. Natl. Acad. Sci. USA* 72, 11-15.
- Graves, B.J., Johnson, P.F. and McKnight, S.L. (1986). Homologous recognition of a promoter domain common to the MSV LTR and the HSV tk gene. *Cell* 44, 565-576.
- 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). Carrier detection of hemophilia B by using a restriction site polymorphism associated with the coagulation factor IX gene. *J. Clin. Invest.* 73, 1491-1495.
- Gruskin, K.D., Smith, T.F. and Goodman, M. (1987). Possible origin of a calmodulin gene that lacks intervening sequences. *Proc. Natl. Acad. Sci. USA* 84, 1605-1608.
- Gubler, U. and Hoffman, B.J. (1983). A simple and very efficient method for generating cDNA libraries. *Gene* 25, 263-269.
- Harrison, S.C. (1986). Fingers and DNA half-turns. *Nature* 322, 597-598.

- Hattori, M. and Sakaki, Y. (1986). Dideoxy sequencing method using denatured plasmid templates. *Anal. Biochem.* 152, 232-238.
- Hayashida, H. and Miyata, T. (1983). Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatibility complex. *Proc. Natl. Acad. Sci. USA* 80, 2671-2675.
- Hershko, A. and Ciechanover, A. (1986). The ubiquitin pathway for the degradation of intracellular proteins. *Progr. Nucl. Acid Res. Mol. Biol.* 33, 19-56.
- Hershko, A., Ciechanover, A., Heller, H., Haas, A.L. and Rose, I.A. (1980). Proposed role of ATP in protein breakdown: conjugation of proteins with multiple chains of the polypeptide of ATP-dependent proteolysis. *Proc. Natl. Acad. Sci. USA* 77, 1783-1786.
- Hershko, A., Ciechanover, A. and Rose, I.A. (1981). Identification of the active amino acid residue of the polypeptide of ATP-dependent protein breakdown. *J. Biol. Chem.* 256, 1525-1528.
- Hershko, A., Heller, H., Elias, S. and Ciechanover, A. (1983). Components of ubiquitin-protein ligase system: resolution, affinity purification, and role in protein breakdown. *J. Biol. Chem.* 258, 8206-8214.
- Hershko, A., Heller, H., Eytan, E., Kaklij, G. and Rose, I.A. (1984a). Role of the α -amino group of protein in ubiquitin-mediated breakdown. *Proc. Natl. Acad. Sci. USA* 81, 7021-7025.
- Hershko, A., Leshinsky, E., Ganoth, D. and Heller, H. (1984b). ATP-dependent degradation of ubiquitin-protein conjugates. *Proc. Natl. Acad. Sci. USA* 81, 1619-1623.
- Hobbs, H.H., Brown, M.S., Goldstein, J.L. and Russell, D.W. (1986). Deletion of exon encoding cysteine-rich repeat of low density lipoprotein receptor alters its binding specificity in a subject with familial hypercholesterolemia. *J. Biol. Chem.* 261, 13114-13120.
- Hook, A.G. and Kellems, R.E. (1988). Localization and sequence analysis of poly(A) sites generating multiple dihydrofolate reductase mRNAs. *J. Biol. Chem.* 263, 2337-2343.
- Hough, R., Pratt, G. and Rechsteiner, M. (1986). Ubiquitin-lysozyme conjugates: identification and characterisation of an ATP-dependent protease from rabbit reticulocyte lysates. *J. Biol. Chem.* 261, 2400-2408.
- Hough, R., Pratt, G. and Rechsteiner, M. (1987). Purification of two high molecular weight proteases from rabbit reticulocyte lysate. *J. Biol. Chem.* 262, 8303-8313.

- Huang, S.-Y., Barnard, M.B., Xu, M., Matsui, S.I., Rose, S.E. and Garrard, W.T. (1986). The active immunoglobulin κ chain gene is packaged by non-ubiquitin-conjugated nucleosomes. *Proc. Natl. Acad. Sci. USA* 83, 3738-3742.
- Hunt, L.T. and Dayhoff, M.O. (1977). Amino-terminal sequence identity of ubiquitin and the nonhistone component of nuclear protein A24. *Biochem. Biophys. Res. Commun.* 74, 650.
- Huynh, R.V., Young, R.A. and Davies, R.W. (1985). Constructing and screening cDNA libraries in λ gt10 and λ gt11. In *DNA Cloning: A Practical Approach*, Volume 1, D.M. Glover, ed. (IRL Press, Oxford, U.K.). pp 49-78.
- Ish-Horowicz, D. and Burke, J.F. (1981). Rapid and efficient cosmid cloning. *Nucl. Acids Res.* 9, 2989-2998.
- Jelinek, W.R. and Schmid, C.W. (1982). Repetitive sequences in eukaryotic DNA and their expression. *Ann. Rev. Biochem.* 51, 813-844.
- Jentsch, S., McGrath, J.P. and Varshavsky, A. (1987). The yeast DNA repair gene RAD6 encodes a ubiquitin-conjugating enzyme. *Nature* 329, 131-134.
- Kadonaga, J.T., Jones, K.A. and Tijan, R. (1986). Promoter specific activation of RNA polymerase II transcription by Sp1. *TIBS* 11, 20-23.
- Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S. and Matsubara, K. (1987). Revision of consensus human Alu repeats - a review. *Gene* 53, 1-10.
- Keller, E.B. and Noon, W.A. (1984). Intron splicing: a conserved internal signal in introns of animal pre-mRNAs. *Proc. Natl. Acad. Sci. USA* 81, 7417-7420.
- Kemp, D.J., Coppel, R.L., Cowman, A.F., Saint, R.B., Brown, G.V. and Anders, R.F. (1983). Expression of Plasmodium falciparum blood-stage antigens in Escherichia coli: detection with antibodies from immune humans. *Proc. Natl. Acad. Sci. USA* 80, 3787-3791.
- Klug, A. and Rhodes, D. (1987). 'Zinc fingers': a novel protein motif for nucleic acid recognition. *Trends. Biochem. Sci.* 12, 464-469.
- Kobilka, B.K., Frielle, T., Dohlman, H.G., Bolanowski, M.A., Dixon, R.F., Keller, P., Caron, M.G. and Lefkowitz, R.J. Delineation of the intronless nature of the genes for the human and hamster β 2-adrenergic receptor and their putative promoter regions. *J. Biol. Chem.* 262, 7321-7327.
- Koltunow, A.M., Gregg, K. and Rogers, G.E. (1986). Intron sequences modulate feather keratin gene transcription in Xenopus oocytes. *Nucl. Acids Res.* 14, 6375-6392.

- Latchman, D.S., Estridge, J.K. and Kemp, L.M. (1987). Transcriptional induction of the ubiquitin gene during herpes simplex virus infection is dependent upon the viral immediate-early protein ICP4. *Nucl. Acids Res.* 15, 7283-7293.
- Lee, M.G.-S., Loomis, C. and Cowan, N.J. (1984). Sequence of an expressed human β -tubulin gene containing ten Alu family members. *Nucl. Acids Res.* 12, 5823-5836.
- Lehrman, M.A., Russell, D.W., Goldstein, J.L. and Brown, M.S. (1987a). Alu-Alu recombination deletes splice acceptor sites and produces secreted low density lipoprotein receptor in a subject with familial hypercholesterolemia. *J. Biol. Chem.* 262, 3354-3361.
- Lehrman, M.A., Goldstein, J.L., Russell, D.W. and Brown, M.S. (1987b). Duplication of seven exons in LDL receptor gene caused by Alu-Alu recombination in a subject with familial hypercholesterolemia. *Cell* 48, 827-835.
- Lennox, E.S. (1955). Transduction of linked genetic characters of the host by bacteriophage P1. *Virology* 1, 190-206.
- Levinger, L. and Varshavsky, A. (1982). Selective arrangement of ubiquitinated and D1 protein-containing nucleosomes within the Drosophila genome. *Cell* 28, 375-385.
- Li, W.-H., Wu, C.-I. and Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2, 150-174.
- Lin, H.-C., Lei, S.-P. and Wilcox, G. (1985). An improved DNA sequencing strategy. *Anal. Biochem.* 147, 114-119.
- Lindquist, S. (1986). The heat-shock response. *Ann. Rev. Biochem.* 55, 1151-1191.
- Linial, M. (1987). Creation of a processed pseudogene by retroviral infection. *Cell* 49, 93-102.
- Loiseau, P., Lehn, P., Dautry, F., Lepage, V., Colombani, J., Cohen, D., Dausset, J. and Degos, L. (1986). Correlation between an HLA-DQ α length polymorphism of messenger RNA and serologically defined specificities (DQw1, DRw53, DR+5). *Immunogenetics* 23, 111-114.
- Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. and Tizard, R. (1979). The structure and evolution of the two nonallelic rat preproinsulin genes. *Cell* 18, 545-558.
- Low, T.L.K. and Goldstein, A.L. (1979). The chemistry and biology of thymosin. II. Amino acid sequence analysis of thymosin α 1 and polypeptide β 1. *J. Biol. Chem.* 254, 987-995.

- Low, T.L.K., Thurman, G.B., McAdoo, M., McGlure, J., Rossio, J.L., Naylor, P.H. and Goldstein, A.L. (1979). The chemistry and biology of thymosin. I. Isolation, characterisation, and biological activities of thymosin α 1 and polypeptide β 1 from calf thymus. *J. Biol. Chem.* 254, 981-986.
- Lund, P.K., Moats-Staats, B.M., Simmons, J.G., Hoyt, E., D'Ercole, A.J., Martin, F. and Van Wyk, J.J. (1985). Nucleotide sequence analysis of a cDNA encoding human ubiquitin reveals that ubiquitin is synthesized as a precursor. *J. Biol. Chem.* 260, 7609-7613.
- Maire, P., Gautron, S., Hakim, V., Gregori, C., Mennecier, F. and Kahn, A. (1987). Characterisation of three optional promoters in the 5' region of the human aldolase gene. *J. Mol. Biol.* 197, 425-438.
- Maniatis, T., Hardison, R.C., Lacy, E., Lauer, J., O'Connell, D., Quon, D., Sim, G.K. and Efstratiadis, A. (1978). *Cell* 15, 687-701.
- Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982). *Molecular cloning: a laboratory manual*. Cold Spring Harbour Laboratory, Cold Spring Harbour, New York.
- Mardon, H.J., Sebastio, G. and Baralle, F.E. (1987). A role for exon sequences in alternative splicing of the human fibronectin gene. *Nucl. Acids Res.* 15, 7725-7733.
- Matsui, S., Sandberg, A.A., Negoro, S., Seon, B.K. and Goldstein, G. (1982). Isopeptidase: a novel eukaryotic enzyme that cleaves isopeptide bonds. *Proc. Natl. Acad. Sci. USA* 79, 1535-1539.
- Matsui, S.I., Seon, B.K. and Sandberg, A.A. (1979). Disappearance of a structural chromatin protein A24 in mitosis: implications for molecular basis of chromatin condensation. *Proc. Natl. Acad. Sci. USA* 76, 6386-6390.
- McCallum, F.S. and Maden, B.E.H. (1985). Human 18S ribosomal RNA sequence inferred from DNA sequence. *Biochem J.* 232, 725-733.
- McLauchlan, J., Gaffney, D., Whitton, J.L. and Clements, J.B. (1985). The consensus sequence YGTGTTY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. *Nucl. Acids Res.* 13, 1347-1368.
- Messing, J., Crea, R. and Seeburg, P.H. (1981). A system for shotgun DNA sequencing. *Nucl. Acids Res.* 9, 309-321.
- Messing, J. and Viera, J. (1982). A new pair of M13 vectors for selecting either strand of a double-digest restriction fragment. *Gene* 19, 269-276.
- Messing, J. (1983). New M13 vectors for M13 cloning. *Methods Enzymol.* 101, 20-79.

- Meyer, E.M., West, C.M. and Chau, V. (1986). Antibodies directed against ubiquitin inhibit high affinity [³H]choline uptake in rat cerebral cortical synaptosomes. *J. Biol. Chem.* 261, 14365-14368.
- Meyer, E.M., West, C.M., Stevens, B.R., Chau, V., Nguyen, M.-T. and Judkins, J.H. (1987). Ubiquitin-directed antibodies inhibit neuronal transporters in rat brain synaptosomes. *J. Neurochem.* 49, 1815-1819.
- Mezquita, J., Oliva, R. and Mezquita, C. (1987). New ubiquitin mRNA expressed during chicken spermiogenesis. *Nucl. Acids Res.* 15, 9604.
- Miller, J.H. (1972). *Experiments in molecular genetics*. Cold Spring Harbour Laboratory, Cold Spring Harbour, New York.
- Miller, J., McLachlan, A.D. and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *EMBO J.* 4, 1609-1614.
- Mita, S., Yasuda, H., Marunouchi, T., Ishiko, S. and Yamada, M. (1980). A temperature-sensitive mutant of cultured mouse cells defective in chromosome condensation. *Exp. Cell Res.* 126, 407-416.
- Molloy, P.L., Powell, B.C., Gregg, K., Barone, E.D. and Rogers, G.E. (1982). Organisation of feather keratin genes in the chick genome. *Nucl. Acids Res.* 10, 6007-6021.
- Mori, H., Kondo, J. and Ihara, Y. (1987). Ubiquitin is a component of paired helical filaments in Alzheimer's disease. *Science* 235, 1641-1644.
- Mueller, R.D., Yasuda, H., Hatch, C.L., Bonner, W.M. and Bradbury, E.M. (1985). Identification of ubiquitinated histones 2A and 2B in *Physarum polycephalum*: disappearance of these proteins at metaphase and reappearance at anaphase. *J. Biol. Chem.* 260, 5147-5153.
- Munro, S. and Pelham, H. (1985). What turns on heat shock genes? *Nature* 317, 477-478.
- Murray, N.E., Brammar, W.J. and Murray, K. (1977). Lambdoid phages that simplify the recovery of *in vitro* recombinants. *Mol. Gen. Genet.* 150, 53-61.
- Norrande, J., Kempe, T. and Messing, J. (1983). Improved M13 vectors using oligonucleotide-directed mutagenesis. *Gene* 26, 101-106.
- O'Brien, S.J., ed (1987). *Genetic maps 1987: a compilation of linkage and restriction maps of genetically studied organisms*, Vol. 4. Cold Spring Harbour Laboratory, Cold Spring Harbour, New York.

- Ozaki, L.S. and Sharma, S. (1984). DEAE-cellulose method for rapid lambda DNA mini-preparation. in Morel, C.M. (ed), *Genes and Antigens of Parasites: A Laboratory Manual*. Fundacao Oswaldo Cruz, Rio De Janeiro, Brazil, 2nd edn, pp. 172-173.
- Ozkaynak, E., Finley, D. and Varshavsky, A. (1984). The yeast ubiquitin gene: head-to-tail repeats encoding a polyubiquitin precursor protein. *Nature* 312, 663-666.
- Ozkaynak, E., Finley, D., Solomon, M.J. and Varshavsky, A. (1987). The yeast ubiquitin genes: a family of natural gene fusions. *EMBO J.* 6, 1429-1439.
- Peacock, A.C. and Dingman, C.W. (1967). Resolution of multiple ribonucleic acid species by polyacrylamide gel electrophoresis. *Biochemistry* 6, 1818-1827.
- Pelham, H.R.B. and Bienz, M. (1982). A synthetic heat-shock promoter element confers heat-inducibility on the herpes simplex virus thymidine kinase gene. *EMBO J.* 1, 1473-1477.
- Pickart, C.M. and Rose, I.A. (1985a). Functional heterogeneity of ubiquitin carrier proteins. *J. Biol. Chem.* 260, 1573-1581.
- Pickart, C.M. and Rose, I.A. (1985b). Ubiquitin carboxyl-terminal hydrolase acts on ubiquitin carboxyl-terminal amides. *J. Biol. Chem.* 260, 7903-7910.
- Powell, L.M., Wallis, S.C., Pease, R.J., Edwards, Y.H., Knott, T.J. and Scott, J. (1987). A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* 50, 831-840.
- Proudfoot, N.J. and Brownlee, G.G. (1976). 3' Non-coding region sequences in eukaryotic messenger RNA. *Nature* 263, 211-214.
- Raboy, B., Parag, H.A. and Kulka, R.G. (1986). Conjugation of [¹²⁵I]ubiquitin to cellular proteins in permeabilized mammalian cells: composition of mitotic and interphase cells. *EMBO J.* 5, 863-869.
- Rackwitz, H.-R., Zehetner, G., Frischauf, A.-M. and Lehrach, H. (1984). Rapid restriction mapping of DNA cloned in phage lambda vectors. *Gene* 30, 195-200.
- Rapoport, S., Dubiel, W. and Muller, M. (1985). Proteolysis of mitochondria in reticulocytes during maturation is ubiquitin-dependent and is accompanied by a high rate of ATP hydrolysis. *FEBS Lett.* 180, 249-252.
- Rechsteiner, M. (1987). Ubiquitin-mediated pathways for intracellular proteolysis. *Ann. Rev. Cell Biol.* 3, 1-30.
- Reed, K.C. and Mann, D.A. (1985). Rapid transfer of DNA from agarose gels to nylon membranes. *Nucl. Acids Res.* 13, 7207-7221.

- Reid, K.B.M., Bentley, D.R. and Wood, K.J. (1984). Cloning and characterisation of the cDNA for the β -chain of normal serum Clq. *Phil. Trans. R. Soc. Lond.* 306, 345-354.
- Remaut, E., Tsao, H. and Fiers, W. (1983). Improved plasmid vectors with thermoinducible expression and temperature-regulated runaway regulation. *Gene* 22, 103-113.
- Richardson, C.C. (1965). Phosphorylation of nucleic acid by an enzyme from T4 bacteriophage-infected *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 54, 158-165.
- Rose, I.A. and Warms, J.V.B. (1983). An enzyme with ubiquitin carboxy-terminal esterase activity from reticulocytes. *Biochemistry* 22, 4234-4237.
- Rossouw, C.M.S., Vergeer, W.P., du Plooy, S.J., Bernard, M.P., Ramirez, F. and de Wet, W.J. (1987). DNA sequences in the first intron of the human pro- α 1(I) collagen gene enhance transcription. *J. Biol. Chem.* 262, 15151-15157.
- Rouyer, F., Simmler, M-C., Page, D.C. and Weissenbach, J. (1987). A sex chromosome rearrangement in a human XX male caused by Alu-Alu recombination. *Cell* 51, 417-425.
- Salvesen, G., Lloyd, C. and Farley, D. (1987). cDNA encoding a human homolog of yeast ubiquitin 1. *Nucl. Acids Res.* 15, 5485.
- Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A. (1980). Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J. Mol. Biol.* 143, 161-178.
- Sanger, F., Nicklen, S. and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
- Sasaki, H., Yoshioka, N., Takagi, Y. and Sakaki, Y. (1985). Structure of the chromosomal gene for human serum prealbumin. *Gene* 37, 191-197.
- Schlesinger, D.H. and Goldstein, G. (1975). Molecular conservation of 74 amino acid sequence of ubiquitin between cattle and man. *Nature* 255, 423-424.
- Schlesinger, D.H., Goldstein, G. and Niall, H.D. (1975). The complete amino acid sequence of ubiquitin, an adenylate cyclase stimulating polypeptide probably universal in living cells. *Biochemistry* 14, 2214-2218.
- Schlesinger, M.J. (1986). Heat-shock proteins: the search for functions. *J. Cell Biol.* 103, 321-325.
- Schlesinger, M.J. and Bond, U. (1987). Ubiquitin genes. in *Oxford Survey of Eukaryotic Genes*, in press.
- Seeburg, P.H. (1982). The human growth hormone gene family: nucleotide sequences show recent divergence and predict a new polypeptide hormone. *DNA* 1, 239-249.

- Sharp, P.A. (1983). Conversion of RNA to DNA in mammals: Alu-like elements and pseudogenes. *Nature* 301, 471-472.
- Sharp, P.M. and Li, W.-H. (1987a). Ubiquitin genes as a paradigm of concerted evolution of tandem repeats. *J. Mol. Evol.* 25, 58-64.
- Sharp, P.M. and Li, W.-H. (1987b). Molecular Evolution of Ubiquitin Genes. *Trends Ecol. Evol.* 2, 328-332.
- Siegelman, M., Bond, M.W., Gallatin, W.M., St John, T., Smith, H.T., Fried, V.A. and Weissman, I.L. (1986). Cell surface molecule associated with lymphocyte homing is a ubiquitinated branched-chain glycoprotein. *Science* 231, 823-829.
- Soares, M.B., Schon, E., Henderson, A., Karathanasis, S.K., Cate, R., Zeitlin, S., Chirgwin, J. and Efstratiadis, A. (1985). RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. *Mol. Cell. Biol.* 5, 2090-2103.
- Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98, 503-517.
- Srikantha, T., Landsman, D. and Bustin, M. (1987). Retropseudogenes for human chromosomal protein HMG-17. *J. Mol. Biol.* 197, 405-413.
- Stanley, K.K. (1983). Solubilization and immune-detection of β -galactosidase hybrid proteins carrying foreign antigenic determinants. *Nucl. Acids Res.* 11, 4077-4092.
- Stanley, K.K. and Luzio, J.P. (1984). Construction of a new family of high efficiency bacterial expression vectors: identification of cDNA clones coding for human liver proteins. *EMBO J.* 3, 1429-1434.
- St. John, T., Gallatin, W.M., Siegelman, M., Smith, H.T., Fried, V.A. and Weissman, I.L. (1986). Expression cloning of a lymphocyte homing receptor cDNA: ubiquitin is the reactive species. *Science* 231, 845-850.
- Thorne, A.W., Sautiere, P., Briand, G. and Crane-Robinson, C. (1987). The structure of ubiquitinated histone H2B. *EMBO J.* 6, 1005-1010.
- Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973). Improved estimation of secondary structure in ribonucleic acids. *Nature New Biol.* 246, 40-41.
- Uher, L. (1986). Electrophoresis conditions for high molecular weight DNA markers. *Focus (Bethesda Research Laboratories)* 8:1, 10-11.

- Urano, Y., Watanabe, K., Sakai, M. and Tamaoki, T. (1986). The human albumin gene: characterisation of the 5' and 3' flanking regions and the polymorphic gene transcripts. *J. Biol. Chem.* 261, 3244-3251.
- Van den Hondel, C.A. and Schoenmakers, J.G.G. (1975). Studies on bacteriophage M13 DNA. I. A cleavage map of the M13 genome. *Eur. J. Biochem.* 53, 547-558.
- Vierstra, R.D., Langan, S.M. and Schaller, G.E. (1986). Complete amino acid sequence of ubiquitin from the higher plant *Avena sativa*. *Biochemistry* 25, 3105-3108.
- Vijay-Kumar, S., Bugg, C.E., Wilkinson, K.D. and Cook, W.J. (1985). Three-dimensional structure of ubiquitin at 2.8 Å resolution. *Proc. Natl. Acad. Sci. USA* 82, 3582-3585.
- Vijay-Kumar, S., Bugg, C.E. and Cook, W.J. (1987a). Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* 194, 531-544.
- Vijay-Kumar, S., Bugg, C.E., Wilkinson, K.D., Vierstra, R.D., Hatfield, P.M. and Cook, W.J. (1987b). Comparison of the three-dimensional structures of human, yeast, and oat ubiquitin. *J. Biol. Chem.* 262, 6396-6399.
- Volckaert, G., Tavernier, J., Derynck, R., Devos, R. and Fiers, W. (1981). Molecular mechanism of nucleotide sequence rearrangements in cDNA clones of human fibroblast interferon mRNA. *Gene* 15, 215-223.
- Watson, D.C., Levy W., B. and Dixon, G.H. (1978). Free ubiquitin is a non-histone protein of trout testis chromatin. *Nature* 276, 196-198.
- Weber, P.L., Brown, S.C. and Mueller, L. (1987). Sequential ¹H NMR assignments and secondary structure identification of human ubiquitin. *Biochemistry* 26, 7282-7290.
- Weiner, A.M., Deininger, P.L. and Efstratiadis, A. (1986). Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Ann. Rev. Biochem.* 55, 631-661.
- West, M.H.P. and Bonner, W.M. (1980). Histone 2B can be modified by the attachment of ubiquitin. *Nucl. Acids Res.* 8, 4671-4680.
- Westphal, M., Muller-Taubenberger, A., Noegel, A. and Gerisch, G. (1986). Transcript regulation and carboxyterminal extension of ubiquitin in *Dictyostelium discoideum*. *FEBS Lett.* 209, 92-96.
- Wiborg, O., Pedersen, M.S., Wind, A., Berglund, L.E., Marcker, K.A. and Vuust, J. (1985). The human ubiquitin multigene family: some genes contain multiple directly repeated ubiquitin coding sequences. *EMBO J.* 4, 755-759.

- Wilkinson, K.D. and Audhya, T. (1981). Stimulation of ATP-dependent proteolysis requires ubiquitin with the C-terminal sequence arg-gly-gly. *J. Biol. Chem.* 256, 9235-9241.
- Wilkinson, K.D., Cox, M.J., O'Connor, L.B. and Shapira, R. (1986). Structure and activities of a variant ubiquitin sequence from bakers' yeast. *Biochemistry* 25, 4999-5004.
- Wilkinson, K.D., Urban, M.K. and Haas, A.L. (1980). Ubiquitin is the ATP-dependent proteolysis factor I of rabbit reticulocytes. *J. Biol. Chem.* 255, 7529-7532.
- Wu, R., Kohn, K.W. and Bonner, W.M. (1981). Metabolism of ubiquitinated histones. *J. Biol. Chem.* 256, 5916-5920.
- Yamamoto, T., Davis, C.G., Brown, M.S., Schneider, W.J., Casey, M.L., Goldstein, J.L. and Russell, D.W. (1984). The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. *Cell* 39, 27-38.
- Yanisch-Perron, C., Vieira, J. and Messing, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* 33, 103-119.
- 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). Structure of the receptor for platelet-derived growth factor helps define a family of closely related growth factor receptors. *Nature* 323, 226-232.
- Yasuda, H., Matsumoto, Y., Mita, S., Marunouchi, T. and Yamada, M. (1981). A mouse temperature-sensitive mutant defective in H1 histone phosphorylation is defective in deoxyribonucleic acid synthesis and chromosome condensation. *Biochemistry* 20, 4414-4419.
- Young, R.A. and Davis, R.W. (1983a). Efficient isolation of genes by using antibody probes. *Proc. Natl. Acad. Sci. USA* 80, 1194-1198.
- Young, R.A. and Davis, R.W. (1983b). Yeast RNA polymerase II genes: isolation with antibody probes. *Science* 222, 778-782.