CHARACTERISATION OF A CLUSTER OF NOVEL GENES IN THE HUMAN MAJOR HISTOCOMPATIBILITY COMPLEX

A thesis submitted for the

Degree of Doctor of Philosophy

January 1993

Mark G. Olavesen

Corpus Christi College, Oxford

MRC Immunochemistry Unit

Department of Biochemistry

University of Oxford



To Mum, Dad, Martin

& Nicki

ACKNOWLEDGEMENTS

I would firstly like to thank Duncan Campbell for his guidance and supervision of my work at both undergraduate and graduate levels in Oxford. I am also very grateful to Wendy Thomson for teaching me the basic molecular biology techniques early on in my work.

I should like to thank all the members of the MRC Immunochemistry Unit for the help and advice I have received, in particular the members of lab F1 who have made it such an enjoyable place to work. I am grateful to Tony Day and Tony Gascoyne for the synthesis of oligonucleotides and to Tony Day for running the protein alignments.

I also wish to thank those who have helped me in the preparation of this thesis; Kathleen Nolan for her diligent proofreading of the text, Ken Johnson for the photographic work and all those individuals who have provided helpful discussion on aspects of my work and its presentation, as well as those who helped to assemble it.

I gratefully acknowledge the financial support of the Medical Research Council and thank my parents for all their support throughout my education. I am also very grateful to my college for their generous support to help me to travel abroad to conferences.

My most sincere thanks go to Nicki for her unfailing love and her support for me in my work over the years.

ABBREVIATIONS

Nucleotides containing the bases adenine, cytosine, guanine and thymine are abbreviated A, C, G and T, respectively. The single-letter and triplet codes are used for amino acids and restriction enzymes are referred to by their three letter names, derived from the source microorganism. The nomenclature of the HLA antigens and genes is that recommended by the World Health Organisation Nomenclature Committee (1992). The nomenclature of the complement components is that recommended by the World Health Organisation (1968; 1981). Other abbreviations are listed below or defined in the text.

APS	Ammonium persulphate
rATP	Adenosine triphosphate
BCIG	5-Bromo-4-chloro-3-indolyl-β-galactopyranoside
BCIP	5-Bromo-4-chloro-3-indolyl phosphate
bp	Base pairs
BSA	Bovine serum albumin
cDNA	Complementary DNA
CpG	Cytidine-guanidine dinucleotide
cpm	Counts per minute
CVID	Combined variable immunodeficiency
CYP21	Cytochrome P-450 steroid 21-hydroxylase
DHFR	Dihydrofolate reductase
DMEM	Dulbecco's modification of Eagle's medium
DMF	Dimethyl formamide
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	deoxyribonucleotide triphosphate
DTT	Dithiothreitol
EAO	Experimental autoimmune orchitis
EDTA	Ethylenediaminetetra-acetic acid
EGF	Epidermal growth factor
ELISA	Enzyme-linked immumosorbent assay
FCS	Foetal calf serum
HEPES	N-[2-Hydroxyethyl]piperazine-N'-2-ethanesulphonic acid
HGT	High gelling temperature

HLA	Human leukocyte antigens
HSP	Heat shock protein
HTF	*Hpa*II tiny fragment
IFN	Interferon
Ig	Immunoglobulin
IPTG	Isopropyl-β D-thiogalactoside
kb	Kilobase pairs
kDa	Kilodalton
LGT	Low gelling temperature
Mb	Megabase pairs
MHC	Major Histocompatibility Complex
MOPS	3-(N-Morpholino)propanesulphonic acid
M_r	Relative molecular mass
mRNA	Messenger RNA
MTPBS	Mouse tonicity phosphate buffered saline
OD_{xxx}	Optical density at xxx nm
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
pfu	Plaque forming units
PFGE	Pulsed-field gel electrophoresis
PIPES	Piperazine-N,N'-bis[2-ethanesulphonic acids]
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RNase	Ribonuclease
RT-PCR	Reverse transcribed-PCR
SD	standard deviation
SDS	Sodium dodecylsulphate
SLE	Systemic lupus erythematosus
SSC	Standard citrate saline
TBE	Tris-borate/EDTA
T_d	Dissociation temperature
TE	Tris/EDTA
TEMED	N,N,N',N'-tetramethylethylenediamine
TNP	Trinitrophenol
tRNA	Transfer RNA
Tris	Tris(hydroxymethyl)-aminomethane
U	Units
YAC	Yeast artificial chromosome

ABSTRACT

CHARACTERISATION OF A CLUSTER OF NOVEL GENES IN THE HUMAN MAJOR HISTOCOMPATIBILITY COMPLEX

Mark G. Olavesen
Submitted for the degree of D.Phil.

Corpus Christi College
Hilary Term, 1993

The human MHC has been shown to be associated with susceptibility to a large number of diseases. Within the MHC class III region, which consists of a ~1,100kb segment of DNA located between the class I and class II multigene families, at least 38 genes have been identified, some of which could be involved in disease susceptibility. Analyses of the homologous mouse *H-2S* region have localised two loci responsible for susceptibility to experimental autoimmune orchitis, *Orch-1*, and corticosteroid-induced cleft palate, *Cps-1*, to an ~100kb segment of DNA. The characterisation of four genes located within the homologous region of the human MHC class III region is described here.

The G7 gene is a novel gene located 58kb telomeric of the HSP70-Hom locus. G7 has been sequenced at the cDNA and genomic levels. The G7 gene, which contains 27 exons that span ~22kb of DNA, is expressed as an ~3kb mRNA. This mRNA was shown to be present at low levels in all of the cell types analysed, and was most abundant in monocytes and macrophages. Sequence analysis of cDNA clones revealed that the G7 mRNA contains two alternatively spliced exons which alter the putative G7 protein at the C-terminus. The different mRNA species encode putative protein products of 747 and 788 amino acids. The putative G7 proteins appear to be intracellular although their function remains to be determined. However, the putative G7 protein sequences do share significant similarity with bacterial DNA mismatch repair proteins and other similar proteins in yeast, mouse and man. A *Pvu*II RFLP associated with the G7 gene was shown to result in a potential Thr→Ala amino acid substitution in ~25% of the haplotypes analysed.

The G7b gene is another novel gene which is located between the HSP70-Hom and valyl-tRNA synthetase (VARS2) genes and may be homologous to other genes elsewhere in the human genome. G7b has been sequenced at the cDNA and genomic levels. The G7b gene contains 5 exons that span ~10kb. The G7b gene expressed an ~0.9kb mRNA that was present in all of the cell types analysed and was most abundant in monocytes and macrophages. The G7b cDNA clones encode a putative G7b protein of 95 amino acids. Polyclonal antisera have been raised against the putative G7b protein which detected an ~5kDa protein in U937 cell lysates. Therefore, the putative G7b protein appears to be an intracellular protein for which the function remains to be determined. However, the putative G7b protein shares significant sequence similarity to the human small ribonucleoprotein Sm-D for which autoantibodies are present in patients with SLE.

A new gene G7c located between the G7 and valyl-tRNA synthetase (VARS2) genes has been identified as a result of the presence of non-repetitive sequence in the intergenic region that were shown to be conserved between species. Exon trapping analysis has identified two discrete exons from within this region.

The mouse homologue of the human HSP70-Hom gene, *Hsc70t*, is a potential candidate gene for the *Orch-1* locus. As a result it was been sequenced in *H-2* congenic mouse strains known to be resistant and susceptible to experimental autoimmune orchitis. Sequence polymorphism was detected although this did not correlate with resistance or susceptibility to the disease.

CONTENTS

		Page No.
ABBREVIATIONS		iv
ABSTRACT		vi
CONTENTS		
CHAPTER 1: INTRODUCTION		
1.1	IDENTIFICATION OF THE MAJOR HISTOCOMPATIBILITY COMPLEX	1
1.2	THE ORGANISATION OF THE MHC	2
1.2.1	Structure of the mouse MHC	2
1.2.2	Structure of the human MHC	3
1.3	THE MHC CLASS I REGION	3
1.3.1	The class I genes	3
1.3.2	Expression and function of the classical class I antigens	4
1.3.3	Structure of the classical class I antigens	5
1.4	THE MHC CLASS II REGION	7
1.4.1	The class II genes	7
1.4.2	Expression and function of the class II antigens	8
1.4.3	Structure of the class II antigens	9
1.4.4	Novel non-HLA class II genes	10
1.4.5	Functions of the novel HLA class II genes	11
1.5	THE MHC CLASS III REGION	13
1.5.1	Organisation of the MHC class III region	13
1.5.2	Characterisation of the MHC class III region	14
1.5.3	The complement gene cluster	15
1.5.4	The CYP21A and CYP21B genes	16
1.5.5	The TNF genes	17
1.5.6	The HSP70 gene cluster	17
1.5.7	The B144 and RD genes	18
1.5.8	Other novel genes in the MHC class III region	18
1.5.9	G12, OSG/XB and Tenascin	21

1.6	POLYMORPHISM IN THE MHC	23
1.6.1	The class I antigens	23
1.6.2	The class II antigens	23
1.6.3	The MHC class III region	23
1.6.3.1	*The complement genes*	23
1.6.3.2	*Other genes in the MHC class III region*	24
1.6.3.3	*RFLPs and microsatellite repeats*	24
1.6.4	Linkage disequilibrium and extended haplotypes	25
1.7	MHC AND DISEASE	25
1.7.1	Diseases in which immunological mechanisms are not involved	25
1.7.2	Autoimmune diseases	26
1.7.3	Diseases of uncertain pathogenesis	28
1.8	GENETIC TRAITS AND LOCI MAPPED WITHIN THE MHC	28
1.8.1	Loci mapped within the human MHC	29
1.8.2	Loci mapped in the mouse H-2 complex	29
1.9	EXPERIMENTAL AUTOIMMUNE ORCHITIS AND THE *Orch-1* LOCUS	30
1.9.1	Infertility and orchitis	30
1.9.2	Immunopathology of orchitis	31
1.9.3	Immunogenetics of EAO	31
1.10	AIMS OF THE PROJECT	33

CHAPTER 2: MATERIALS & METHODS

2.1	MATERIALS	34
2.1.1	Chemicals and Reagents	34
2.1.2	Enzymes	35
2.1.3	Reagent Kits	35
2.1.4	*E.coli* Cloning Vectors	36
2.1.5	*E. coli* Strains	36
2.1.6	Standard buffers and solutions	36
2.1.7	Bacterial culture media	38
2.2	STORAGE AND GROWTH OF ESCHERICHIA COLI BACTERIAL STRAINS	40
2.2.1	Storage of bacterial strains and clones	40

2.2.2	Overnight cultures	40
2.2.3	Preparation of competent bacterial cells for immediate use	40
2.2.4	Preparation of frozen competent bacterial cells	40
2.3	PREPARATION OF NUCLEIC ACIDS	41
2.3.1	Recovery of nucleic acids from solution	41
2.3.1.1	*Organic extraction*	41
2.3.1.2	*Ethanol precipitation of nucleic acids*	41
2.3.2	Small scale preparation of plasmid and cosmid DNA	41
2.3.2.1	*Alkali lysis method*	41
2.3.2.2	*Magic minipreps DNA purification*	42
2.3.3	Large scale preparation of plasmid DNA	42
2.3.4	Preparation of genomic DNA	43
2.3.5	Preparation of total RNA	43
2.3.6	Preparation of oligonucleotide primers	44
2.3.7	Reverse transcription of RNA to cDNA	44
2.3.8	Polymerase chain reaction (PCR) protocols	45
2.3.9	Estimation of nucleic acid concentration	45
2.4	**GENERAL NUCLEIC ACID MANIPULATIONS**	45
2.4.1	Restriction enzyme digests	45
2.4.2	Fractionation of DNA by agarose gel electrophoresis	46
2.4.3	Isolation of DNA probes using LGT agarose	46
2.4.4	End-filling of digested DNA with recessed 3′ termini	46
2.4.5	Fractionation of DNA by polyacrylamide gel electrophoresis	47
2.4.6	Recovery of DNA fragments from polyacrylamide gel slices	47
2.4.7	Fractionation of RNA by agarose gel electrophoresis	47
2.5	TRANSFER OF NUCLEIC ACIDS TO NITROCELLULOSE MEMBRANES	48
2.5.1	Transfer of DNA from agarose gels (Southern blotting)	48
2.5.2	Transfer of RNA from agarose gels (Northern blotting)	48
2.5.3	Transfer of bacterial DNA to nitrocellulose membranes	48
2.5.4	Transfer of bacteriophage M13 DNA to nitrocellulose membranes	49
2.6	HYBRIDISATION TECHNIQUES	49
2.6.1	Radiolabelling of double-stranded DNA probes	49
2.6.2	Hybridisation of radiolabelled DNA	49

2.6.3	Washing of nitrocellulose membranes and autoradiography	50
2.7	SUBCLONING METHODS	50
2.7.1	Preparation of plasmid vectors	50
2.7.2	Preparation of insert DNA	51
2.7.3	Ligation reactions	51
2.7.4	Religation reactions	51
2.7.5	Transformation procedures	51
2.7.6	Transfection procedures	52
2.7.7	Identification of recombinant clones	52
2.7.7.1	*Colour selection*	52
2.7.7.2	*PCR amplification of inserts*	53
2.7.7.3	*Hybridisation using a radiolabelled probe*	53
2.7.7.4	*Restriction enzyme digestion of plasmid DNA*	53
2.8	ISOLATION OF CLONES FROM cDNA LIBRARIES	53
2.8.1	Screening and isolation of positive clones	53
2.8.2	Rescreening of positive clones	54
2.9	DNA SEQUENCING	54
2.9.1	Preparation of single-stranded DNA from bacteriophage M13	54
2.9.2	Preparation of single-stranded DNA from phagemid clones	55
2.9.3	Dideoxy sequencing of single-stranded DNA templates	55
2.9.4	Dideoxy sequencing of double-stranded DNA templates	55
2.9.5	Cycle sequencing	56
2.9.6	Denaturing polyacrylamide gel electrophoresis	56
2.9.7	Analysis of sequence data	56
2.10	DETERMINATION OF TRANSCRIPTION START SITES	57
2.10.1	Primer extension analysis	57
2.10.1.1	*Preparation of radiolabelled primer*	57
2.10.1.2	*Hybridisation and reverse transcription*	57
2.10.2	RNase protection analysis	58
2.10.2.1	*Preparation of radiolabelled riboprobe*	58
2.10.2.2	*Hybridisation and RNase digestion*	59
2.11	EXON AMPLIFICATION	59
2.11.1	Cell culture and electroporation	60
2.11.2	RNA isolation and RT-PCR	60

2.12	GENERAL PROTEIN TECHNIQUES	61
2.12.1	Computer analysis of protein sequences	61
2.12.2	Analysis of proteins by SDS-PAGE	61
2.12.3	Transfer of proteins onto nitrocellulose membranes (Western blotting)	61
2.12.4	Determination of protein concentration	62
2.12.5	Expression of polypeptides in *E.coli*	62
2.12.6	Isolation of fusion proteins from *E.coli*	63
2.12.6.1	*Small scale preparations*	63
2.12.6.2	*Large scale preparations*	63
2.12.7	Direct thrombin cleavage of purified fusion protein	64
2.12.8	Elution of fusion proteins from glutathione agarose beads	64
2.13	PREPARATION AND PURIFICATION OF POLYCLONAL ANTIBODIES	65
2.13.1	Immunisation of rabbits with antigen	65
2.13.2	Preparation and storage of rabbit serum	65
2.13.3	Purification of antibodies by affinity chromatography	65
2.13.3.1	*Preparation of affinity chromatography columns*	65
2.13.3.2	*Affinity chromatography*	66
2.14	CHARACTERISATION OF ANTIBODIES	66
2.14.1	ELISA assays	66
2.14.2	Detection of antigens on Western blots	67

CHAPTER 3: CHARACTERISATION OF G7

3.1	INTRODUCTION	68
3.2	ISOLATION AND CHARACTERISATION OF G7 cDNA CLONES	69
3.2.1	Isolation of G7 cDNA clones	68
3.2.2	Characterisation of cDNA inserts	70
3.2.3	Nucleotide sequence analysis of the cDNA clones	70
3.2.3.1	*G7-1*	70
3.2.3.2	*G7-2*	71
3.2.3.3	*G7-3*	71
3.2.3.4	*Alignment of cDNA sequences*	72
3.2.4	Confirmation of U937 sequence at position 1971	72

3.2.5	Mapping the 5′ end of G7 mRNA by primer extension	73
3.2.6	Determination of the G7 ORF	73
3.2.7	Correction of a G7 cDNA clone	74
3.2.8	Southern and Northern blot analysis	75
3.3	DETERMINATION OF G7 GENE STRUCTURE	76
3.3.1	Mapping and subcloning the G7 gene	76
3.3.2	Sequencing G7 exons	77
3.3.3	G7 gene structure	78
3.3.3.1	*Exonic organisation of G7*	78
3.3.3.2	*Alu repeats*	79
3.3.4	Explanation of cDNA sequence variations	80
3.3.5	Upstream sequence analysis of the G7 gene	81
3.4	EXPRESSION OF G7 mRNA	82
3.4.1	Tissue distribution of G7 expression	82
3.4.2	Exon usage in G7 mRNA	83
3.5	CHARACTERISATION OF THE G7 GENE PRODUCTS	84
3.5.1	Characteristics of the G7 protein sequence	84
3.5.2	Determination of G7 protein sequence identity	85
3.6	POLYMORPHIC ANALYSIS OF G7	88
3.6.1	RFLP analysis	88
3.6.2	Characterisation of the G7 *PvuII* polymorphism	89
3.6.3	Dinucleotide $(CA)_n$ repeats	90
3.7	EXPRESSION OF G7 POLYPEPTIDES	91
3.7.1	Expression of G7 fusion proteins	91
3.7.2	Purification of G7 fusion proteins	92
3.8	DISCUSSION	93
3.8.1	The G7 gene	93
3.8.2	Functions of the G7 proteins	94
3.8.3	Possible roles of G7 in disease	95

CHAPTER 4: CHARACTERISATION OF G7b

4.1	INTRODUCTION	97
4.2	ISOLATION AND CHARACTERISATION OF G7b cDNA CLONES	98

4.2.1	Isolation and identification of G7b cDNA clones	98
4.2.2	Mapping the location of the cDNA inserts in cosmid 10S	99
4.2.3	Southern and Northern blot analysis	99
4.2.4	Nucleotide sequence analysis of the G7b cDNA clones	100
4.2.5	Mapping the 5′ end of G7b mRNA by RNase mapping	101
4.2.6	G7b cDNA structure	102
4.3	DETERMINATION OF G7b GENE STRUCTURE	102
4.3.1	Mapping of the G7b gene	103
4.3.2	Nucleotide sequence analysis of G7b exons	104
4.3.2.1	*Sequencing the 5′ region of G7b*	104
4.3.2.2	*Sequencing the 3′ region of G7b*	104
4.3.3	G7b gene structure	104
4.3.4	Identification of an additional HTF island at the 5′ end of G7b	105
4.3.5	Upstream sequence analysis of G7b	105
4.3.6	*Alu* repeat sequences in G7b	106
4.3.7	RFLP analysis of G7b	106
4.4	CHARACTERISATION OF THE G7b GENE PRODUCT	107
4.4.1	Features of the putative G7b polypeptide sequence	107
4.4.2	Determination of G7b protein sequence identity	107
4.5	EXPRESSION OF THE G7b POLYPEPTIDE	109
4.5.1	Expression and isolation of G7b fusion protein	109
4.5.2	Preparation of G7b antibodies and Western blot analysis	110
4.5.2.1	*ELISA analysis*	110
4.5.2.2	*Western blot analysis of* Sj26-G7b *fusion protein*	110
4.5.2.3	*Affinity chromatography purification of anti-*G7b *IgG*	111
4.5.2.4	*Western blot analysis of total* U937 *protein*	111
4.6	DISCUSSION	112
4.6.1	The G7b gene	112
4.6.2	Functions of the G7b protein and possible roles in disease aetiology	113

CHAPTER 5: EVIDENCE FOR A NEW GENE, G7c

5.1	INTRODUCTION	115

5.2	IDENTIFICATION OF POSSIBLE LOCATIONS FOR NEW GENES	117
5.2.1	Mapping non-repetitive DNA sequences between HSP70-Hom and BAT5	117
5.2.2	Characterisation of the G7–G7a intergenic region	118
5.3	IDENTIFICATION OF THE PUTATIVE GENE, G7c	120
5.3.1	Use of exon amplification to identify coding sequences in the putative gene, G7c	120
5.3.2	Mapping the exon amplification products in the religated cosmid, 10P rE	121
5.3.3	Nucleotide sequence analysis of the exon amplification products	122
5.3.4	Zoo blot analysis of the amplified exons	123
5.4	DISCUSSION	124
5.4.1	Strategies for gene detection	124
5.4.2	G7c, another novel gene in the human MHC	125

CHAPTER 6: POLYMORPHIC ANALYSIS OF Hsc70t

6.1	INTRODUCTION	127
6.2	EXPRESSION OF HSP70-Hom and Hsc70t	129
6.2.1	Northern blot analysis of HSP70-Hom	129
6.2.2	Expression of *Hsc70t* in *H-2* congenic mouse strains	130
6.3	POLYMORPHIC SEQUENCE ANALYSIS OF Hsc70t	130
6.3.1	Cloning and sequencing of *Hsc70t*	130
6.3.2	*Hsc70t* nucleotide sequence analysis	131
6.3.3	Analysis of *Hsc70t* protein sequences	132
6.3.4	Polymorphic analysis of *Hsc70t* in *H-2* congenic mouse strains	133
6.4	DISCUSSION	134
6.4.1	*Hsc70t* – a candidate for the *Orch-1* locus?	134
6.4.2	Polymorphic analysis of *Hsc70t*	134
6.4.3	Heat shock proteins and the immune system	135
6.4.4	Heat shock proteins as autoantigens	136

CHAPTER 7: SUMMARY & CONCLUSIONS

7.1	CHARACTERISATION OF NOVEL GENES	137
7.1.1	The G7 gene	137
7.1.2	The G7b gene	138
7.1.3	The new novel gene G7c	139
7.2	NON-REPETITIVE SEQUENCES AS MARKERS FOR MAPPING NEW GENES	140
7.3	COLINEARITY OF GENES IN THE HUMAN AND MOUSE CLASS III REGIONS	141
7.4	THE MHC CLASS III REGION AND DISEASE ASSOCIATION	143

REFERENCES

CHAPTER 1

INTRODUCTION

1.1 IDENTIFICATION OF THE MAJOR HISTOCOMPATIBILITY COMPLEX

The major histocompatibility antigens play a central role in tissue compatibility and were originally identified during studies of alloreactivity and allograft rejection in mice (Gorer *et al.* 1948). The genetic loci that encode these cell surface molecules in mice were found to be closely linked and were defined as the histocompatibility-2, *H-2*, complex. Analysis of recombinant congenic mouse strains indicated that there were two major loci within this region, *H-2K* and *H-2D*, that encode multiple allelic variants (Klein 1975). In man, the use of alloantisera from polytransfused individuals and multiparous women led to the identification of the human leukocyte antigens, HLA (Bach and van Rood 1976), which were related to the mouse alloantigens. These polymorphic HLA antigens have been designated the MHC class I antigens and in man are encoded by the HLA-A, -B and -C loci of the MHC.

The class II antigens were discovered when inbred mouse and guinea pig strains were found to vary in their antibody responses to antigens, which have a limited range of antigenic determinants. The genetic control of the T cell dependent immune response, *Ir* genes (McDevitt and Chinitz 1969), was later mapped to the *I* region of the *H-2* complex, close to the *H-2K* region (McDevitt *et al.* 1972). The human HLA–D/DR region was identified as responsible for the 'mixed lymphocyte reaction' (Niederhuber and Frelinger 1976) and was shown to encode polymorphic cell surface molecules corresponding to the mouse *Ir* gene products of the *H-2I* region. The HLA-D/DR loci were later shown to be linked to the HLA-A and HLA-B regions in the human MHC (Dausset 1981).

To further characterise the class II region, recombinant congenic mice were analysed, resulting in subdivision of the *I* region into the *I-A* and *I-E* subregions. In man, primed lymphocyte typing (Shaw *et al.* 1981) and DNA cloning (Trowsdale *et al.* 1985) led to the identification of additional loci subsequently defined as the DP

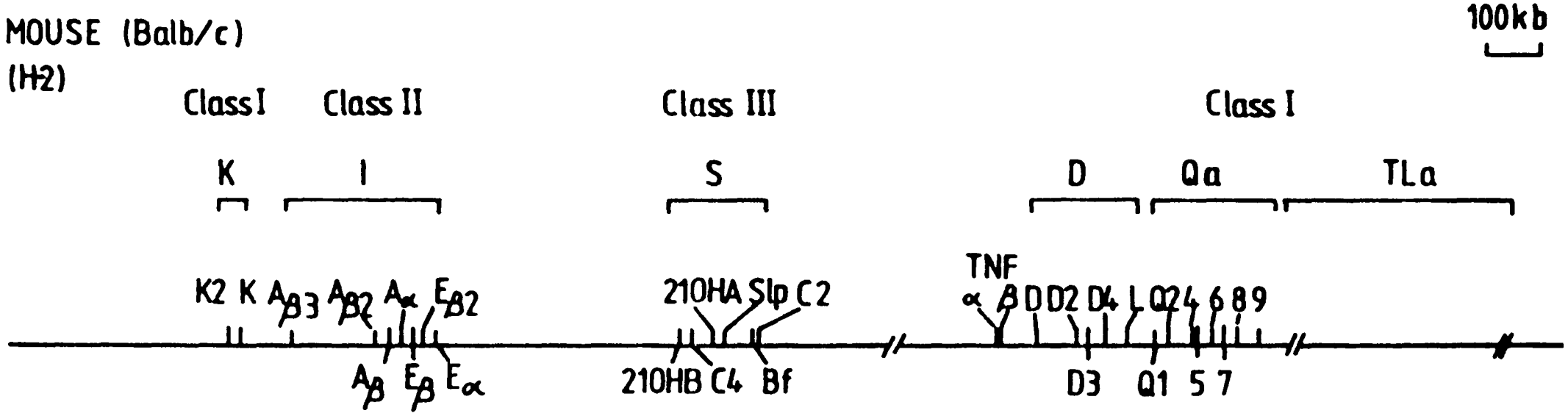


Figure 1.1 Map of the mouse *H-2* complex on chromosome 17.
The centromere is to the left and the telomere is to the right (from Campbell *et al.* 1988).

and DQ subregions of the HLA complex.

In addition to the class I alloantigen loci and the class II *Ir* genes, another group of loci were also mapped to the mouse *H-2* complex. Demant *et al.* (1973) observed that haemolytic complement levels among inbred mouse strains were controlled by a gene or genes linked to the *H-2* complex. These genes mapped to a region designated *H-2S*, between *H-2I* and *H-2D*, and were subsequently identified as the *C2*, *Slp*, *Ss* (C4) and *Bf* loci (Shreffler 1976). The corresponding human genes C2, C4A, C4B and factor B were mapped as a tightly linked cluster to the same region of the human MHC (Carroll *et al.* 1984), now known as the class III region.

1.2 THE ORGANISATION OF THE MHC

The histocompatibility antigens are encoded by a single gene complex in mammalian species. In man the MHC is located on the short arm of chromosome 6 (Olaisen *et al.* 1987) and the mouse *H-2* complex is located on chromosome 17 (Figueroa and Klein 1986). The corresponding rat, *RT1*, and rabbit, *RLA*, complexes have also been identified (Klein and Figueroa 1986) and MHC class I and class II loci have been identified in nonhuman primate species, birds (chicken), reptiles, amphibians (axolotl) and fish (carp) (Accolla *et al.* 1991).

1.2.1 Structure of the mouse MHC

The mouse *H-2* complex on chromosome 17 spans ~2.3Mb (Muller *et al.* 1987a; Lafuse *et al.* 1992) and is divided into several gene clusters (Figure 1.1). The class I genes are located in the *K*, *D Qa* and *Tla* regions, but only the highly polymorphic *H–2K* and *H-2D* loci encode the major class I antigens (Klein *et al.* 1983). The class II genes are located in the *I-A* and *I-E* subregions, between the *K* and *S* regions, which express the Aα, Aβ, Eα and Eβ class II antigens (Hood *et al.* 1983). The *S* region encodes the complement gene cluster (*C2*, *C4*, *Slp* and *Bf*) and lies between the *I* and *D* regions.

Recent analysis of the *H-2* complex has led to the identification of many novel genes in the region [section 7.3].

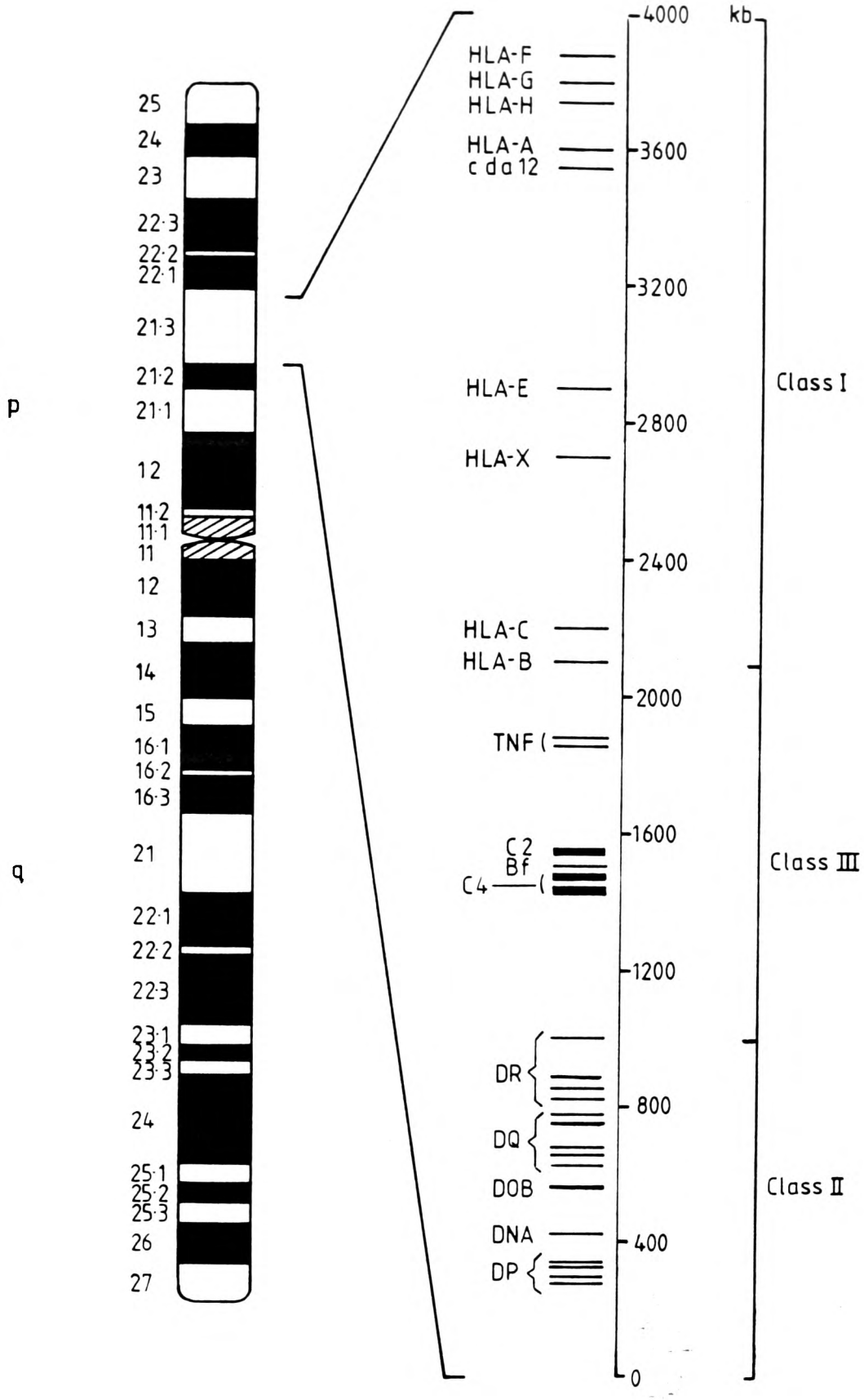


Figure 1.2 Schematic representation of human chromosome 6 and the MHC. The light and dark band regions on the left side correspond to the G-light and G-dark bands of chromosome 6.The region of the chromosome corresponding to the MHC, in band 6p21.3 is enlarged on the right. The extents of the class I, class II and class III regions are marked and genes are represented by bars. In the class III region the complement (C4, C2 and factor B) and the TNF genes are marked (from Milner and Campbell 1992b).

1.2.2 Structure of the human MHC

The human MHC is located on the short arm of chromosome 6 at 6p21.3 (Figure 1.2; Olaisen *et al.* 1987). Unlike the mouse *H-2* complex, the human class I genes all lie telomeric of the class III region. As a result the class II, HLA-D, region is located at the centromeric end and the class I, HLA-B, -C and -A, loci are located at the telomeric end of the MHC. The class III region lies between the most telomeric class II locus, DRA, and the most centromeric class I locus, HLA-B. The organisation of the genes in the human MHC has been determined by PFGE (Hardy *et al.* 1986; Carroll *et al.* 1987; Dunham *et al.* 1987; Lawrance *et al.* 1987; Shukla *et al.* 1991) and by cosmid and YAC cloning (Ragoussis *et al.* 1991; Sargent *et al.* 1989a; Spies *et al.* 1989a,b 1990; Kendall *et al.* 1990; Chimini *et al.* 1990; Geraghty *et al.* 1992a). The locations of the genes are shown in figure 1.3 (Trowsdale *et al.* 1991).

1.3 THE MHC CLASS I REGION

1.3.1 The class I genes

The human MHC class I region comprises ~2Mb of DNA (Lawrance *et al.* 1987) and contains >25 related genes, pseudogenes and gene fragments (Koller *et al.* 1989; Geraghty *et al.* 1992a,b). Seven genes have been characterised (Shukla *et al.* 1991; Chimini *et al.* 1990) and mapped in the order:

CEN - HLA-B - HLA-C - HLA-X - HLA-E - HLA-A - HLA-G - HLA-F - TEL

The pseudogenes HLA-H (Chorney *et al.* 1990) and HLA-J, cda 12, (Ragoussis *et al.* 1989) are located telomeric and centromeric of HLA-A, respectively (Figure 1.3). Only the classical class I genes HLA-A and HLA-B are known to encode polypeptides that are functional in the presentation of antigens to T cells. HLA-C is closely linked to HLA–B and these loci are believed to have been generated by the duplication of an ancestral gene. However, HLA-C expression is very low and antigen presentation by HLA-C antigens has not been observed (Gussow *et al.* 1987). HLA-E, -F and -G all express mRNA transcripts (Houlihan *et al.* 1992) and HLA-F has been detected in the cytoplasm (Geraghty *et al.* 1990). HLA-E and HLA-G form heterodimers with β_2

Figure 1.3 Map of the human MHC (Trowsdale *et al*. 1991).
Maps of the class II (top), class III (middle) and class I (bottom) regions are shown. Genes are marked by coloured boxes and the direction of transcription of the genes, where known, is shown by arrows beneath the boxes. The insets show haplotype dependent variations in the number of DRB genes and the size and number of C4/CYP21 genes.

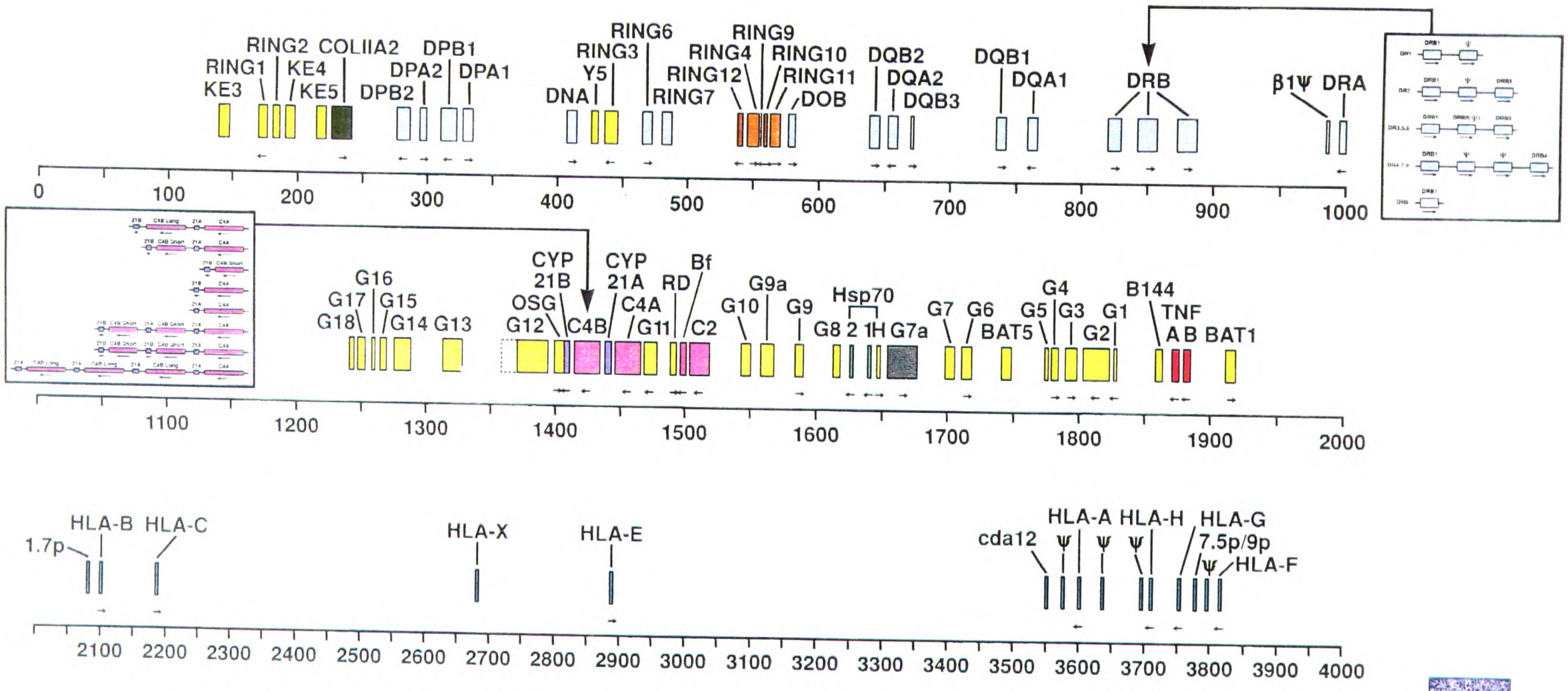
IMMUNOLOGY*today*

December 1991

Map of the human major histocompatibility complex

John Trowsdale, Jiannis Ragoussis and
R. Duncan Campbell

Sponsored by
Amersham International
Amersham



Ψ: pseudogene; DRβ box: some of the alternative gene arrangements at the DRβ loci;
Complement box: some of the alternative gene arrangements at the CYP21 and C4 loci.

ELSEVIER
TRENDS
JOURNALS

microglobulin (β_2m), but only HLA-G is expressed on the cell surface and its role there is unclear (Koller *et al.* 1988; Geraghty *et al.* 1990). The pseudogenes HLA-H (Chorney *et al.* 1990) and HLA-J (Messer *et al.* 1992) contain deleterious mutations and the class I related gene HLA-X has not been shown to be expressed. The DNA content of the class I region can vary as a result of ~50kb deletions between the HLA–A and HLA-G loci. These deletions include the HLA-H pseudogene and are associated with the HLA-A23 or HLA-A24 haplotypes (Venditti and Chorney 1992).

Several cDNA clones not related to the MHC class I genes have been isolated by cDNA selection, using a yeast artificial chromosome (YAC) encoding part of the MHC class I region including the HLA-A locus (Parimoo *et al.* 1991), but no sequence data from these clones has been reported. However, another cDNA clone, mapped ~5kb from the 5′ end of the HLA-E gene, encodes a putative G-protein (Denizot *et al.* 1992).

1.3.2 Expression and function of the classical class I antigens

The class I antigens are expressed on the surface of all nucleated cells and are upregulated in response to viral infection and cytokines *e.g.* TNFα and IFN-γ (David-Watine *et al.* 1990). The regulatory sequences that control MHC class I gene transcription are also present in the β_2m gene promoter, located on chromosome 15 in man (Goodfellow *et al.* 1975), indicating that there may be a common mechanism for the control of class I and β_2m expression. Nascent class I α chains and β_2m are targeted to the endoplasmic reticulum (ER) where the presence of β_2m is essential for correct folding of the α chain (Krangle *et al.* 1979), possibly because part of the α chain structure is stabilised by interactions with β_2m (Saper *et al.* 1991). Short peptides are also essential for complete assembly of the class I-β_2m complex and its transport to the cell surface (Townsend *et al.* 1989).

The class I antigens are responsible for the MHC restriction of T cell mediated cytotoxicity to allogeneic or virally infected cells (Zinkernagel and Doherty 1979). The cytotoxic T cell response requires the presentation of foreign antigen in the context of self MHC molecules at the cell surface. Class I molecules bind peptides derived from endogenous proteins (Jardetzky *et al.* 1991) in the ER and are transported to the cell

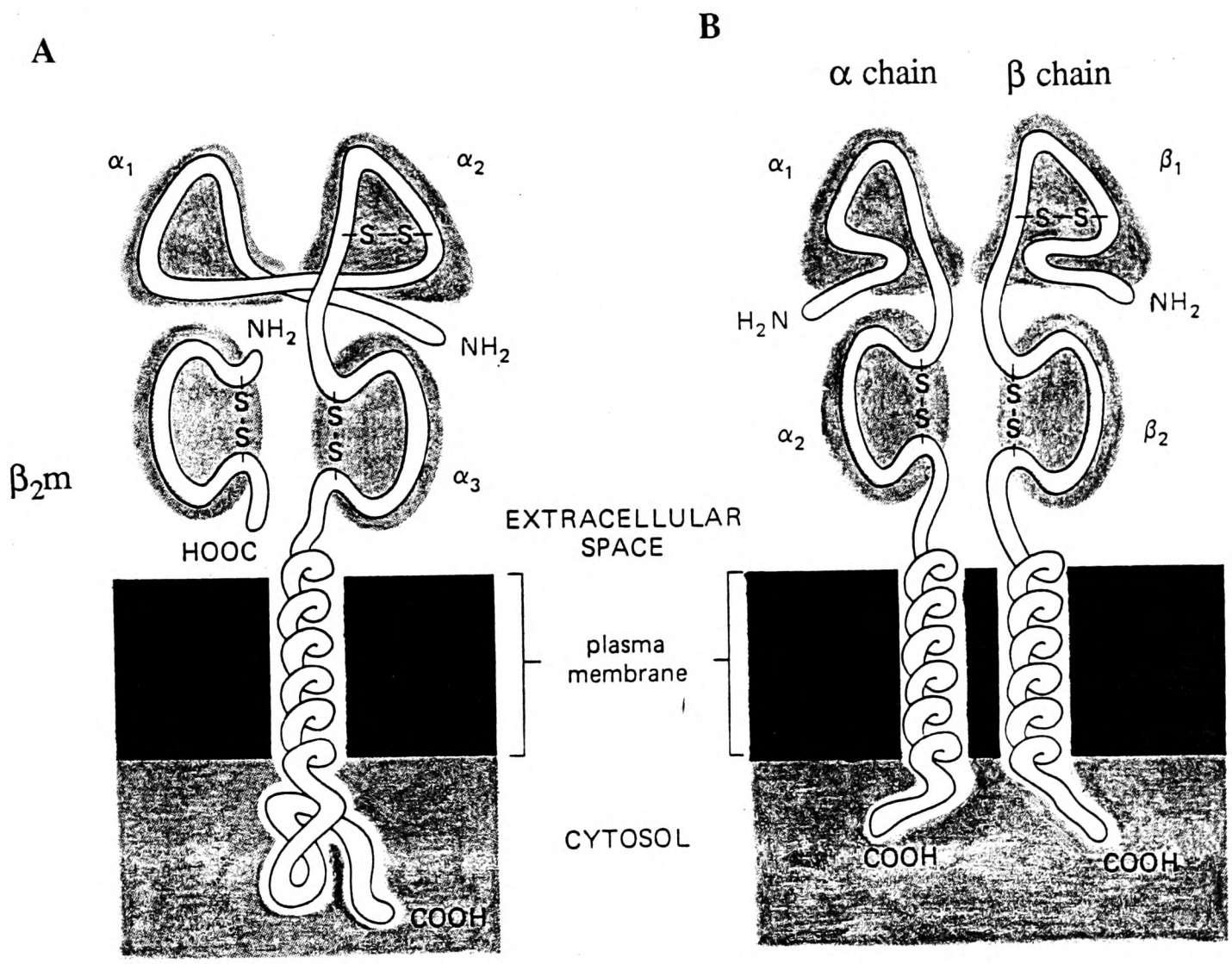


Figure 1.4 Primary structures of the MHC class I and class II antigens.

A Class I antigen. The extracellular domains of the class I α chain and β_2m are indicated.

B Class II antigen. The extracellular domains of the class II α and β chains are indicated.

surface where virally encoded antigenic peptides are recognised by T cell receptors. The MHC class I molecules are recognised by T cell receptors on cytotoxic T cells because the cell adhesion molecule CD8 on their cell surface specifically recognises the relatively invariant α_3 domain of the class I molecule (Salter *et al.* 1990).

1.3.3 Structure of the classical class I antigens

The HLA-A and HLA-B loci of the MHC class I region encode the glycosylated, highly polymorphic α chain (45kDa) of the class I antigens which become noncovalently linked to the invariant β_2m (12kDa) chain (Cresswell *et al.* 1973). Amino acid and nucleotide sequence analysis has shown that the class I α chain has three extracellular domains, α_1, α_2 and α_3 (Figure 1.4), of 90, 92 and 92 amino acids, respectively, each of which is encoded by a separate exon (Srivastava *et al.* 1985). The membrane proximal, α_3, domain and β_2m have strong similarity to the immunoglobulin domain and belong to the Ig superfamily (Williams and Barclay 1988) whilst the membrane distal, α_1 and α_2, domains contain the most polymorphic regions of the α chain (Parham *et al.* 1988). A short transmembrane domain and a cytoplasmic tail of ~30 amino acids anchor the extracellular domains to the cell membrane.

The crystallisation of soluble class I HLA-A2 molecules has allowed the three-dimensional structure of the α_1, α_2 and α_3 domains, complexed with β_2m, to be determined by X-ray crystallography (Bjorkman *et al.* 1987a). Subsequently the structures of the HLA-A68 and HLA-B27 molecules have been determined (Garrett *et al.* 1989; Madden *et al.* 1991) and found to have the same structure (Figure 1.5). The molecule is composed of two sets of structurally homologous domains. As expected from the sequence homology, the α_3 and β_2m domains have the Ig constant domain structure of two antiparallel β sheets connected by an internal disulphide bond. The two Ig domains are paired together and the CD8 α chain binding site has been identified on an exposed loop connecting two of the β strands of the α_3 domain (Salter *et al.* 1990).

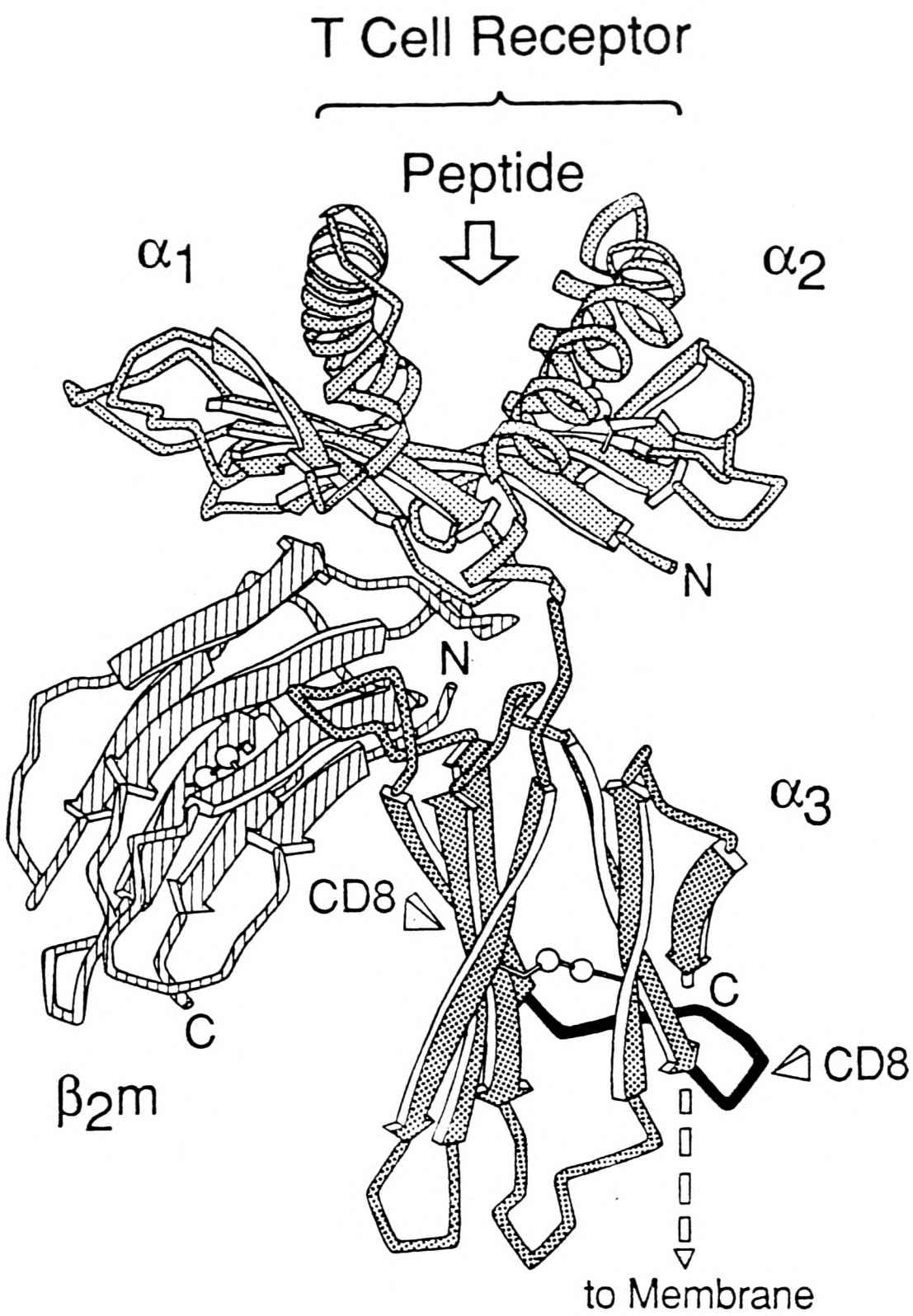


Figure 1.5 A ribbon diagram of the human MHC class I molecule structure. β sheets are represented by thick arrows and α helices by helical ribbons. The CD8 binding site on the α_3 domain is shown in black and the peptide binding groove and hypothesised area of interaction with the T cell receptor are indicated (from Lawlor *et al.* 1990).

The α_1 and α_2 domains both consist of a four-stranded antiparallel β sheet and a long α helical region. These two domains interact to form a platform of 8 antiparallel β strands above the α_3 and β_2m Ig domains and the two α helical regions lie antiparallel on top (Figure 1.5; Bjorkman *et al.* 1987a). The two α helical regions generate a deep groove along the top surface of the class I molecule, the base of which is formed by the central strands of the β sheet. This groove contained unaccounted electron density that has subsequently been shown to be peptides in an extended conformation (Garrett *et al.* 1989; Saper *et al.* 1991). The polymorphic amino acid residues of the HLA class I molecules, concentrated in the α_1 and α_2 domains (Parham *et al.* 1988), occur mainly at positions that orientate their side chains into the peptide-binding groove, either from the sides of the α helical regions or at the base from residues in the β sheet pointing upwards (Bjorkman *et al.* 1987b).

Refinement and comparison of the class I molecule structures has revealed that the peptide-binding groove contains pockets, the shape and structure of which varies according to the polymorphic residues present (Garrett *et al.* 1989; Madden *et al.* 1991). The peptides bound in the groove have been isolated and some have been sequenced (Jardetzky *et al.* 1991; Guo *et al.* 1992). Most of the peptides are 8-10 residues in length and contain conserved 'anchor' residues located towards the ends of the peptides that vary between different class I molecules (Jardetzky *et al.* 1991; Guo *et al.* 1992). The structure of several mouse and human class I structures, complexed with single peptides, have been determined to reveal the conformation of the peptides in the groove (Fremont *et al.* 1992; Matsumura *et al.* 1992; Zhang *et al.* 1992; Silver *et al.* 1992). The peptides adopt an extended β conformation, bound and oriented in the groove by their termini and the side chains of the anchor residues which make specific interactions with residues in pockets at the ends of the groove. Heterogeneous peptide populations can be accommodated within the peptide-binding groove of a single class I molecule because intermolecular interactions in the middle of the groove are less specific, thereby allowing peptides of variable length to fit by bulging out of the groove in the middle (Guo *et al.* 1992). In addition, the side chains of residues pointing out of the groove are not restricted (Jardetzky *et al.* 1991; Guo *et al.* 1992). Therefore, the structure of the class I molecule-

antigen complex allows presentation of antigens in a form that can be recognised by the T cell receptor (Figure 1.5).

1.4 THE MHC CLASS II REGION

1.4.1 The class II genes

The human class II region comprises ~900kb of DNA (Hardy *et al.* 1986) at the centromeric end of the MHC (Figure 1.2), although the DNA content has been shown to vary from ~750kb up to ~860kb between haplotypes (Dunham *et al.* 1989). This region is divided into subregions of related loci, HLA-DP, -DQ and -DR, (Trowsdale *et al.* 1985), each containing at least one A and one B gene encoding the class II α and β chains, respectively. Additional loci, DOB, DMA, DMB and DNA, have been identified and their arrangement in the class II region is:

CEN - DP - DNA - DMA - DMB - DOB - DQ - DR - TEL

The DP, DQ and DR subregions contain multiple genes and pseudogenes (Figure 1.3; Trowsdale *et al.* 1991). In the DP subregion there are two functional genes, DPA1 and DPB1, encoding the DP α and β chains and two pseudogenes DPA2 and DPB2. The DQ subregion contains two functional genes, DQA1 and DQB1, encoding the DQ α and β chains and three related sequences, DQA2, DQB2 and DQB3, which are not known to be expressed. In the DR subregion there is one functional DRA locus encoding the DR α chain and up to 5 DRB genes and pseudogenes, the number depending upon the DR haplotype. These include four expressed genes, DRB1, DRB3, DRB4 and DRB5, that can all associate with the DR α chain, and five pseudogenes, DRB2, DRB6, DRB7, DRB8 and DRB9 (WHO Nomenclature Committee 1992).

The variation in DNA content observed between haplotypes (Dunham *et al.* 1989) occurs between the DQ and DR loci (Kendall *et al.* 1991). The DR3, DR5 and DR6 haplotypes contain ~750kb, DR2 haplotypes have 30kb of extra DNA within the DR subregion and DR4, DR7 and DR9 haplotypes have an extra 110kb between HLA-DQ

and HLA-DR (Dunham *et al.* 1989; Kendall *et al.* 1991).

A large number of non-classical class II genes have also been mapped to the class II region and these are described in section 1.4.4.

1.4.2 Expression and function of the class II antigens

The class II HLA antigens are expressed on a limited range of cell types such as B lymphocytes, macrophages, dendritic cells, thymic epithelium, Kupffer cells Langerhans cells and activated T lymphocytes. Expression of the class II antigen genes is regulated by cytokines (Benoist and Mathis 1990) via transcription regulatory elements that include X, Y and S boxes which have also been identified in the promoter of the invariant chain, a polypeptide which associates with the class II antigen in the ER, the gene for which is located on human chromosome 5.

Nascent class II α and β chains are directed into the ER where they complex with the invariant chain (Kvist *et al.* 1982) that can exist in various forms due to post-translational modifications (Teyton and Peterson 1992). The class II/invariant chain complex is transported towards the cell surface and in the Golgi complex the invariant chain is proteolytically processed, resulting in transport of the class II molecules to endosomes (Lotteau *et al.* 1990) where the invariant chain is removed.

Class II antigens are responsible for the MHC restriction of T helper cell activity (Benacerraf 1981). These polymorphic glycoproteins present peptides derived from exogenous proteins (Rudensky *et al.* 1991). Extracellular proteins are internalised by antigen presenting cells and, after proteolytic degradation, peptides bind to class II molecules in the endosomes (Harding *et al.* 1991). The acidic environment within the endosomes facilitates the dissociation of the invariant chain and allows association of peptides with the class II $\alpha\beta$ dimers, which is essential for correct, stable folding of the class II molecule (Sadegh-Nasseri and Germain 1992). The MHC class II/peptide complex is then transported to the cell surface. T helper cells bearing the CD4 cell adhesion glycoprotein specifically recognise MHC class II molecules and the identification of foreign antigenic peptides by T cell receptors results in the activation of an immune response via the production of lymphokines.

1.4.3 Structure of the class II antigens

The MHC class II molecules are noncovalent heterodimers consisting of a highly polymorphic β chain (27-29kDa) and a less variable α chain (33-35kDa). Sequence analysis (Kappes and Strominger 1988) has shown that the α and β chains are similar in structure. Both chains have two extracellular domains, α_1, α_2 and β_1, β_2, of ~90 amino acid residues, which are encoded by separate exons, and a C-terminal hydrophobic transmembrane sequence ~25 amino acids long, followed by a short cytoplasmic tail of 10-20 amino acid residues (Figure 1.4).

The crystal structure of the class II molecule is not yet available, but the α_2 and β_2 domains bear strong sequence homology to the Ig constant domain (Kaufman *et al.* 1984) and are members of the Ig superfamily (Williams and Barclay 1988). These two membrane proximal domains are therefore believed to correspond to a pair of domains in a similar way to the β_2 m and α_3 domains in the class I structure of Bjorkman *et al.* (1987a).

The N-terminal α_1 and β_1 domains of the class II molecule appear to be responsible for binding peptides (Babbitt *et al.* 1985) and are highly polymorphic (Kaufman *et al.* 1984). As a result, these two domains have been modelled upon the class I α_1 and α_2 domain peptide-binding groove (Brown *et al.* 1988). The class II α_1 and β_1 domains show very little pairwise homology with the class I α_1 and α_2 domains. However, a disulphide bond in the class I α_2 domain is present in the class II β_1 domain, the N-glycosylation site in the class I α_1 domain is present in the class II α_1 domain, there are several structural residues conserved between the sequences and the polymorphic residues in the class I molecule occur in the corresponding positions in the class II α_1 and β_1 domains.

Using this information, Brown *et al.* (1988) modelled the predicted class II α_1 and β_1 domain secondary structures onto the class I α_1 and α_2 domain tertiary structure of Bjorkman *et al.* (1987a) to form a peptide-binding groove of α helical regions and a β sheet. Additional salt bridges within the proposed structure are also conserved and help to stabilise the tertiary structure. Many of the polymorphic residues in the class

II molecule occur at positions in which they are likely to interact with peptide in the groove of the modelled structure and may thus determine the peptides bound to the class II molecule for presentation to the T cell receptor.

Peptides have been eluted from class II molecules and sequenced by Rudensky *et al.* (1991). The sequences obtained revealed several differences with those of peptides eluted from class I molecules. The peptides isolated from class II molecules, 13-17 residues in length, are longer than the 8-10 residue peptides isolated from the class I molecules and are more heterogeneous in length, even amongst overlapping peptides with the same antigenic determinant. In addition there do not appear to be any conserved 'anchor' residues in the peptides isolated from class II molecules. As a result, Rudensky *et al.* (1991) proposed that one end of the peptide-binding groove may be open, thereby allowing peptides to protrude from the molecule.

1.4.4 Novel non-HLA class II genes

The MHC class II region has been cloned in overlapping YAC clones (Ragoussis *et al.* 1991) and mapped in detail by PFGE (Hanson *et al.* 1991). Using HTF-islands (Bird 1986, 1987) mapped within the region as markers for genes (Hanson *et al.* 1991) and cDNA probes mapped to the mouse *H-2K* region (Janatipour *et al.* 1992), fifteen new genes have been mapped within the ~450kb of DNA located centromeric of the HLA–DOB locus (Figure 1.3; Trowsdale *et al.* 1991).

Six of the novel genes are located centromeric of DPB2 and may not be part of the class II region although it has not been determined whether or not there are additional class II A and B genes centromeric of DPB2 (Trowsdale *et al.* 1991). HKE2-5 and HSET are human homologues of the mouse genes KE2-5 and SET, respectively, (Janatipour *et al.* 1992) for which no function has been determined. COL 11A2 is the α2 (XI) collagen gene (Hanson *et al.* 1989), and RING1 and RING2 are two novel genes (Hanson *et al.* 1991), one of which, RING1, contains a novel cysteine-rich sequence motif (Freemont *et al.* 1991).

The nine novel genes mapped within the class II region between DNA and DOB (Figure 1.3) encode a diverse mixture of protein products. The DMA (RING6) and DMB (RING7) genes encode the α and β chains of the HLA class II-related locus DM (Kelly *et al.* 1991a). The RING3 (Y4) gene encodes a homologue of the *Drosophila*

female sterile homeotic (*fsh*) gene and also has partial sequence homology with other proteins involved in cell cycle control, cell division and the regulation of cell growth (Beck *et al.* 1992).

The RING4, 9, 10-12 genes are a very tightly linked gene cluster ~10kb centromeric of the DOB locus (Figure 1.3) that include two pairs of related proteins that are upregulated by IFN-γ (Kelly *et al.* 1991b). RING12 and RING10 (Y2) encode polypeptides from a large intracellular protease complex, called the proteasome (Kelly *et al.* 1991b; Glynne *et al.* 1991), which have been named LMP2 and LMP7, respectively (WHO Nomenclature Committee 1992). The RING4 (Y3, PSF1) and RING11 (Y1, PSF2) genes encode polypeptides that contain the ATP-binding cassette (ABC) and share homology with the ABC superfamily of transporter proteins (Trowsdale *et al.* 1990; Spies *et al.* 1990; Bahram *et al.* 1991; Powis *et al.* 1992a). These genes have been named TAP1 and TAP2, for - 'transporter associated with antigen processing' (WHO Nomenclature Committee 1992). No further data on Y5 (Spies *et al.* 1990) or RING9 (Glynne *et al.* 1991) has been reported.

1.4.5 Functions of the novel HLA class II gene products

The functions of the TAP1, TAP2, LMP2 and LMP7 gene products and their mouse and rat homologues have been studied, in particular with reference to their possible roles in antigen processing and presentation as well as their involvement in susceptibility to autoimmune diseases.

Mutant human and mouse cell lines deficient for cell surface expression of MHC class I molecules have been analysed to determine the cause of this defect. The addition of exogenous peptides to cells was found to upregulate cell surface expression of the class I molecules and it was proposed that a defect in intracellular peptide transport was responsible for the deficiency (Townsend *et al.* 1989). Some of the mutant cell lines contained deletions within the MHC (Cerundolo *et al.* 1990) that were mapped to the class II region (Spies *et al.* 1990) and have subsequently been shown to include the LMP and TAP genes.

The identification of LMP2 and LMP7 in the human MHC (Brown *et al.* 1991; Glynne *et al.* 1991) and mouse *H-2* complex (Ortiz-Navarrete *et al.* 1991; Martinez and Monaco 1991), closely linked to the TAP genes or their homologues, suggested that the LMP protease components play a role in the degradation of proteins in the

cytoplasm, possibly for antigen presentation (Brown *et al.* 1991; Ortiz-Navarrete *et al.* 1991; Martinez and Monaco 1991). However, Arnold *et al.* (1992) and Momburg *et al.* (1992) have subsequently shown that LMP2 and LMP7 are not essential for peptide processing and subsequent presentation by MHC class I molecules. Thus, the functions of the LMP2 and LMP7 gene products are not known.

The pair of TAP genes in the human MHC (Bahram *et al.* 1991; Powis *et al.* 1992a) have homologues in the mouse *H-2* complex and the rat *RT1* complex named *Ham*, histocompatibility antigen modifier (Monaco *et al.* 1990), and *mtp*, MHC-linked transporter protein (Deverson *et al.* 1990), respectively. The presence of these genes within the regions deleted in mutants defective for assembly and transport of MHC class I molecules (Deverson *et al.* 1990; Spies *et al.* 1990) has suggested that these genes encode proteins that transport peptides from the cytoplasm to the ER for association with MHC class I molecules. These proteins share homology with the ABC superfamily of transporter proteins that are symmetrical molecules consisting of two ATP-binding cassettes and up to twelve transmembrane helices in two hydrophobic domains (Higgins *et al.* 1990). In the cystic fibrosis transmembrane conductance regulator protein (CFTR) and the multidrug resistance (MDR) proteins these domains are encoded by a single polypeptide chain. However, TAP1, TAP2 and their homologues each encode a polypeptide with a single ATP-binding cassette and six putative transmembrane helices (Trowsdale *et al.* 1990; Spies *et al.* 1990; Deverson *et al.* 1990; Monaco *et al.* 1990; Bahram *et al.* 1991; Powis *et al.* 1992a). It has been shown that the TAP1 and TAP2 gene products form a complex, possibly a heterodimer, and both proteins are required for MHC class I molecule assembly and antigen presentation (Kelly *et al.* 1992; Spies *et al.* 1992). In addition, Kleijmeer *et al.* (1992) have shown that the putative transporter complex is located in the ER and is probably oriented with its ATP-binding domains in the cytosol. Functional data supporting the role of the TAP and *Ham* gene products in intracellular peptide transport has come from the restoration of class I antigen presentation in mutant cell lines transfected with transporter gene cDNA clones (Spies and DeMars 1991; Powis *et al.* 1991; Attaya *et al.* 1992).

Evidence conflicting with the model that the TAP gene products are ATP–dependent peptide transporters has been presented by Levy *et al.* (1991) who have demonstrated that peptide transport across the ER membrane is ATP

independent and Gaskins *et al.* (1992) who discovered that the TAP gene products do not show affinity for ATP-agarose beads. However, more direct evidence for the association of peptide transporters with antigen presentation has been provided by Powis *et al.* (1992b) who have shown that polymorphism in the rat *mtp* genes is associated with variation in the spectrum of peptides bound to the rat MHC class I molecules. Polymorphic analysis of four *mtp2* alleles revealed 29 variable positions, of which 25 correlated with the variation in peptide binding and most of these occur in the N-terminal region of the polypeptide which contains the hydrophobic, membrane spanning regions (Powis *et al.* 1992b). In the mouse a point mutation resulting in a truncated *Ham2* gene protein product in the RMA-S cell line appears to be responsible for deficient class I assembly and antigen presentation (Yang *et al.* 1992). In man 7 different polymorphic nucleotides in 6 alleles, resulting in six amino acid changes, have been identified in the TAP1 and TAP2 genes in both the ATP-binding cassette and the hydrophobic domain (Powis *et al.* 1992a; Spies *et al.* 1992; Colonna *et al.* 1992a). However, none of these polymorphisms appear to be associated with susceptibility to ankylosing spondylitis, insulin-dependent diabetes or coeliac disease (Colonna *et al.* 1992a). This level of polymorphism in man appears to be lower than in the rat and has not been shown to be functional.

1.5 THE MHC CLASS III REGION

1.5.1 Organisation of the MHC class III region

The ~1,100kb of DNA located in the human MHC between the class II DRA locus and the class I HLA-B locus is referred to as the class III region. The gene products encoded within this region have a diverse range of functions that differ from the classical class I and class II antigens. Originally identified as a result of variations in haemolytic complement levels associated with the mouse *H-2* complex (Demant *et al.* 1973), the genes for complement components C2, C4A, C4B and factor B were the first to be mapped within this region. Subsequently the gene for cytochrome P-450 steroid 21–hydroxylase, CYP21, responsible for congenital adrenal hyperplasia, and the TNF–A and B genes were mapped within the class III region (White *et al.* 1985; Carroll *et al.* 1985, 1987; Dunham *et al.* 1987). Following the isolation of cosmids

from this region (Sargent *et al.* 1989a; Spies *et al.* 1989a,b, 1990; Kendall *et al.* 1990) a large number of novel unrelated genes have been identified (Figure 1.3). These include the genes for the major heat shock proteins HSP70-1 and HSP70-2 and the homologous gene HSP70-Hom (Sargent *et al.* 1989b; Milner and Campbell 1990) as well as valyl-tRNA synthetase (Hsieh and Campbell 1991).

The MHC class III region contains three duplicated loci. The CYP21, C4 and opposite strand gene (OSG or X) are duplicated (Morel *et al.* 1989; Gitelman *et al.* 1992) and show haplotype dependent variation in gene copy numbers (Campbell *et al.* 1990). The TNF locus contains the TNF-A and TNF-B genes whilst the HSP70 locus contains three genes HSP70-1, HSP70-2 and HSP70-Hom (Milner and Campbell 1990). Analysis of the DNA content within the MHC class III region in seven HLA haplotypes by PFGE (Dunham *et al.* 1990) has revealed that there is no variation other than the deletions or duplications observed at the C4 and CYP21 loci, and the length polymorphism of C4 (Campbell *et al.* 1990).

1.5.2 Characterisation of the MHC class III region

Detailed characterisation of the human MHC class III region has been facilitated by isolating overlapping cosmid clones in a single contig from HLA-B to within ~200kb of the DRA locus (Sargent *et al.* 1989a; Spies *et al.* 1989a,b, 1990; Kendall *et al.* 1990). The complement genes, the TNF genes and the duplicated CYP21 genes, previously mapped to this region, were located within the contig and comparative mapping also identified human homologues of the mouse genes B144 (Tsuge *et al.* 1987) and RD (Levi-Strauss *et al.* 1988) within the class III region. The remaining ~30 genes identified within the human MHC class III region were detected by searching for genes within the isolated cosmids.

Sargent *et al.* (1989a) cloned a 540kb DNA segment and used CpG-rich HTF-islands, which are often associated with the 5′ end of genes (Bird 1986, 1987), to predict the locations of new genes within this region. These islands were identified by the presence of multiple, clustered, restriction sites for methylation-sensitive, rare-cutting restriction enzymes in uncloned genomic DNA. Probes flanking the HTF-islands were used in Northern blot analysis to detect novel transcripts. This approach confirmed the localisation of B144 close to TNF-A, and RD within the complement gene cluster, and in addition identified eleven novel genes G1–11

(Sargent *et al.* 1989a), plus the duplicated HSP70 genes (Sargent *et al.* 1989b).

Spies *et al.* (1989a,b, 1990) cloned ~900 kb extending from HLA-B towards DRA and used cosmid inserts to probe Northern blots of RNA derived from a selection of cultured cell lines (Spies *et al.* 1989a). Following the detection of transcripts, the cosmid inserts were used to screen cDNA libraries constructed from RNA derived from B and T lymphocytic cell lines. cDNA clones from the 'HLA-B associated transcripts' (BAT) were isolated and have been named BAT1-5 (Spies *et al.* 1989a) and BAT6-9 (Spies *et al.* 1989b). The genes identified centromeric of the complement cluster have been named X1-6 (Spies *et al.* 1990).

The cosmid contig isolated by Sargent *et al.* (1989a) was extended by Kendall *et al.* (1990) with an additional ~160kb being isolated in overlapping cosmids centromeric of CYP21B. Cosmid inserts spanning this region were used to screen a cDNA library constructed from RNA derived from a monocytic cell line and seven new genes G12-18 were identified. The cosmids isolated by Sargent *et al.* (1989a) have also been used to isolate new probes and to screen directly for additional cDNA clones encoded with cosmid inserts. Two new genes, G7a (Hsieh and Campbell 1991) and G9a (Milner and Campbell 1993), were isolated in this manner. In addition another gene, G3a, has been identified in the G3-G4 interval. Two other genes, designated XA and XB, have been identified immediately centromeric of, and overlapping the CYP21A and CYP21B genes, respectively (Morel *et al.* 1989; Gitelman *et al.* 1992).

Many of the BAT and X genes are the same as the G genes and the corresponding names have been listed by Trowsdale *et al.* (1991). The locations of these genes are shown in figure 1.3.

1.5.3 The complement gene cluster

The genes encoding the complement components C2, C4A, C4B and factor B are clustered within an ~120kb region located ~400kb telomeric of the HLA-DRA locus (Carroll *et al.* 1984, 1987; Dunham *et al.* 1987).

The complement system is a cascade of at least 30 plasma and membrane bound proteins that perform many of the major effector functions of humoral immune and inflammatory responses (Campbell *et al.* 1988; Law and Reid 1988). The main functions of this proteolytic cascade are: complement mediated cytolysis of target cells; opsonisation of microorganisms and antigens to promote phagocytosis by

leukocytes; activation and recruitment of cells involved in the immune response to the site of complement activation and solubilisation of large antibody-antigen complexes for immune clearance.

The central component of the complement system is C3 which is activated by the C3 convertases which cleave C3 to yield C3a and C3b. C3a is a potent anaphylatoxin in the inflammatory response and C3b can be deposited on foreign cells where it covalently binds, via a reactive acyl group, to promote adherence to phagocytic cells or, in association with the C3 convertase, activate the complement lytic pathway, via cleavage of C5, resulting in cytolysis.

There are two different C3 convertases which are generated by the classical and alternative pathways of complement activation. The classical pathway is activated primarily by the interaction of C1, via C1q, with immune complexes containing IgM and subclasses of IgG. The C1s component of the activated C1 complex then cleaves C4 to yield C4a, an anaphylatoxin, and C4b which covalently binds to the target surface via a reactive acyl group. The serine protease C2 then binds to C4b and is split by C1s to yield the C3 convertase C4b2a.

The alternative pathway is activated on antigenic surfaces where C3b, continuously present at low levels in the fluid phase, can bind. Surface bound C3b can bind the serine protease factor B which is cleaved by factor D to produce the alternative pathway C3 convertase C3bBb. Thus, C2, C4 and factor B play essential roles in the complement system. C2 and factor B are highly related serine proteases that form the catalytic components of the C3 and C5 convertases. These proteins share ~35% sequence identity and probably evolved from a duplication of an ancestral gene (Campbell 1987). The two forms of C4, C4A and C4B, both form covalent linkages with hydroxyl and amino groups, but C4A has a greater affinity for amino groups (Law *et al.* 1984).

1.5.4 The CYP21A and CYP21B genes

The duplicated steroid 21-hydroxylase genes CYP21A and CYP21B lie ~3 kb from the 3′ ends of the C4A and C4B genes, respectively (White *et al.* 1985; Carroll *et al.* 1987). The CYP21B gene, adjacent to C4B, is expressed as the microsomal enzyme cytochrome P450 steroid 21-hydroxylase which plays an important role in the synthesis of cortisol in the adrenal cortex (Nerbert and Gonzalez 1987). CYP21A is a

pseudogene containing numerous mutations in the coding sequence. Deficiency or dysfunction of steroid 21-hydroxylase is the most common cause of congenital adrenal hyperplasia (Miller 1988).

1.5.5 The TNF genes

The TNF-A and TNF-B genes are closely linked and lie ~220kb centromeric of HLA–B (Carroll *et al.* 1987; Dunham *et al.* 1987). The protein products, TNF-α (cachectin) and TNF-β (lymphotoxin) share ~35% sequence identity. TNF-α is mainly secreted by macrophages, following activation by lipopolysaccharides, whilst TNF-β is mainly produced by activated T cells. These cytokines interact with a common cell surface receptor and upregulate the adhesion of leukocytes to the endothelium as well as activating macrophages, granulocytes and cytotoxic T cells resulting in the destruction of virally infected and tumour cells. TNF-α and β therefore play an important role in the regulation of an immune response (Beutler and Cerami 1989).

1.5.6 The HSP70 gene cluster

The genes encoding three members of the HSP70 family of proteins have been located in the MHC class III region, HSP70-1, HSP70-2 and HSP70-Hom (Sargent *et al.* 1989b; Milner and Campbell 1990). The HSP70-1 and HSP70-2 genes are ~12kb apart and lie ~92kb telomeric of the C2 gene (Sargent *et al.* 1989b). These encode identical protein products (Milner and Campbell 1990) while HSP70-Hom, which is located ~4kb telomeric of HSP70-1, encodes a homologous protein of the same size which shares 90% identity with them (Milner and Campbell 1990). Expression of the genes for the major HSP70 proteins, HSP70-1 and HSP70-2, is induced by heat shock and other stress stimuli, but HSP70-Hom has not been shown to be heat inducible (Milner and Campbell 1990).

HSP70 proteins associate with ribonucleoproteins and play a role in recovery of heat shocked cells. In addition, HSP70 proteins interact with other proteins and peptides in an ATP-dependent manner and act as chaperones in the secretion of unfolded proteins or the folding of proteins in the ER, mitochondria and cytosol (Deshaies *et al.* 1988).

The protein and peptide binding activities of HSP70 proteins has resulted in the suggestion that they may play a role in antigen presentation (Pierce *et al.* 1989).

1.5.7 The B144 and RD genes

The B144 and RD genes were mapped within the mouse *H-2S* region (Tsuge *et al.* 1987; Levi-Strauss *et al.* 1988) and human homologues were identified at corresponding positions in the human MHC class III region (Sargent *et al.* 1989a; Levi-Strauss *et al.* 1988).

B144 lies ~10kb centromeric of the TNF-A gene and is expressed in B lymphocytes and macrophages (Sargent *et al.* 1989a). The predicted mouse protein sequence has been determined, but does not show any significant sequence identity with any known protein (Tsuge *et al.* 1987).

RD is located within the complement gene cluster, ~10kb centromeric of the factor B gene in both man and mouse, and is expressed in all the tissues examined (Levi-Strauss *et al.* 1988). The predicted protein sequence of the 41kDa human RD protein contains 24 dipeptide repeats of alternating basic and acidic residues, similar to those present in the U1 small nuclear ribonucleoprotein (snRNP) 70kDa subunit (Speiser and White 1989).

1.5.8 Other novel genes in the MHC class III region

During the identification of the novel genes in the human MHC class III region, G1–18, G3a, G7a, G9a, BAT1 and BAT5, the distribution of their expression was examined by Northern blot analysis (Sargent *et al.* 1989a; Spies *et al.* 1989a,b, 1990; Kendall *et al.* 1990; W Thomson personal communication). All of the genes except G1 and G7 were shown to be ubiquitously expressed in the cell lines examined. The nucleotide sequences of many of the novel genes identified in the MHC class III region have now been identified, thereby allowing putative protein sequences to be predicted. No nucleotide sequence data is currently available for the BAT1, G5/BAT4, BAT5 or G10 genes. Expression of G1, G3a, G4, G6, G9a, G11, G13 and G16 at the protein level has been shown by Western blot analysis (Cheng 1991, Milner and Campbell 1993, N Gomez Escobar and A Khanna personal communications).

G1 is a single copy 1kb gene that expresses an ~0.6kb mRNA detected in non-granular leukocytic cell lines. The mRNA transcript encodes a novel ~10kDa

intracellular protein that contains two potential 'EF hand' calcium binding domains in the N-terminal region (Olavesen *et al.* 1993). The function of G1 is unknown, but the activation of non-granular leukocytes involves an increase in intracellular calcium levels and G1 may play a role in intracellular signalling following this event.

G2/BAT2 encodes a putative novel ~228kDa proline rich polypeptide (Banjeri *et al.* 1990). The predicted protein sequence contains several related regions plus four RGD motifs typical of the integrin receptor family, however, the absence of a hydrophobic leader peptide indicates that the putative BAT2 protein is an intracellular protein.

G3/BAT3 encodes a putative novel ~120kDa proline rich protein containing sequence homology in an N-terminal domain with ubiquitin (Banjeri *et al.* 1990).

G3a encodes two different mRNA transcripts in liver and monocytes as a result of alternative splicing at the 5′ end. The putative G3a novel peptides contain two N–glycosylation sites, but only the transcript isolated from liver contained a hydrophobic leader peptide that allows secretion of the protein (W Thomson personal communication). The ~23kDa G3a protein has been detected in plasma (Cheng 1991).

G4 is a single copy gene that encodes a novel ~50kDa protein with homology to glucose transporter proteins (W Thomson personal communication).

G6 encodes mRNA transcripts of 1.4 and 1.5kb and is related to a homologous gene outside the MHC (Cheng 1991). Both mRNA transcripts encode a putative ~26kDa novel protein with regions of sequence homologous to the *E.coli* stringent starvation protein (SSP), which may be a transcription factor, and the tubulin α chain (Cheng 1991).

G7a/BAT6 is a single copy gene that encodes a putative ~140kDa polypeptide, the sequence of which shares 48% identity with the valyl-tRNA synthetase of *Saccharomyces cerevisiae* over 1043 amino acids and at the N-terminus shares 28% identity over 150 amino acids with the elongation factor 1 γ-chain from the brine shrimp, *Artemia salina* (Hsieh and Campbell 1991). In addition, the predicted molecular weight of the G7a gene product is similar in size to other mammalian tRNA synthetases, strongly suggesting that G7a encodes the human valyl-tRNA

synthetase.

G8 is a single copy gene that encodes a novel, putative ~13kDa protein that shares no significant homology with any known protein sequence (Partanen and Campbell 1992).

G9 encodes a putative novel ~38kDa polypeptide, the sequence of which shows 22% sequence identity with the sialidase, neuraminidase, enzyme of *Clostridium perfringens* over its entire length (R D Campbell, S Smith and C Milner personal communication). Neuraminidases remove sialic acid moieties from oligosaccharides on glycoproteins or glycolipids. In the mouse the *Neu-1* locus controlling liver neuraminidase activity has been mapped to the S region of the *H-2* complex (Figueroa *et al.* 1982). G9 therefore appears to be the human equivalent of the mouse *Neu-1* locus.

G9a is a single copy gene that encodes a novel ~110kDa polypeptide (Milner and Campbell 1993). The G9a protein contains six contiguous copies of a 33 amino acid motif known as the ankyrin or 'type C' notch repeat, a polyglutamic acid region, a cysteine rich region and a C-terminal region that shares homology with the *Drosophila* trithorax protein that binds zinc (Milner and Campbell 1993).

G11 encodes a number of alternatively spliced transcripts that encode a variety of putative polypeptide products (Kendall 1992; Hsieh 1992). One of the putative novel products shows 23% sequence identity over 125 amino acids with cytochrome P450 aldosterone synthase (Kendall 1992). Although this level of homology is low, it is interesting to observe that the cytochrome P450 steroid 21-hydroxylase gene CYP21B and the CYP21A pseudogene lie ~50 and ~25kb away, respectively. The major putative G11 product contains a bipartite nuclear localisation signal and shares protein sequence homology with Fujinamic sarcoma virus tyrosine kinase transforming protein (Hsieh 1992). At the C-terminus the shared homology corresponds to the tyrosine kinase domain, whilst at the N-terminus the homologous region is believed to be responsible for transformation activity.

G13 is a single copy gene that encodes a novel ~77kDa protein that shares protein sequence similarity to the cAMP response element binding (CREB) protein, proto-oncogene *c-jun* and activation of transcription factor 6 (ATF-6) in an N–terminal region that contains a leucine zipper DNA binding domain and to fumarate hydratase at the C-terminus (Khanna and Campbell 1993).

G14 and G15 are single copy genes that express mRNA species of 2.0 and 2.1kb, respectively. Partial cDNA sequences reveal the presence of a putative hydrophobic leader peptides at the N-termini of both sequences (Kendall 1992).

G16 contains sequences that cross hybridise to DNA fragments from outside the MHC and is therefore related to a homologous gene or pseudogene elsewhere in the human genome. The G16 protein is an ~20kDa polypeptide that shares 25% sequence identity over 85 amino acids with the transforming protein *ret* (A Khanna personal communication).

G17 is another gene that cross hybridises with nucleotide sequences elsewhere in the human genome and the original probe used to detect the gene hybridised to 3 mRNA species of ~1.6, 2.4 and 3.3kb in Northern blot analysis (Kendall *et al.* 1990).

G18 is a single copy gene that is transcribed to produce mRNA transcripts of ~1.4 and 1.5kb. cDNA sequence analysis has revealed two possible ORFs (Kendall 1992), one of which encodes a putative polypeptide that shares 21% sequence identity over the whole sequence with Mullerian inhibitory substance (MIS), a glycoprotein that causes regression of the Mullerian duct in both the male and female reproductive systems (Kendall 1992).

The characterisation of novel genes G7 and G7b is described in chapters 3 and 4.

1.5.9 G12, OSG/XB and Tenascin

G12 was identified by Kendall *et al.* (1990) when two ~1kb cDNA clones, isolated from a library constructed from RNA derived from a monocytic cell line, were shown to hybridise to a 23-34kb region of DNA located ~9kb centromeric of CYP21B. However, no mRNA species was detected by Northern blot analysis.

Morel *et al.* (1989) identified a pair of novel, duplicated genes, designated XA and XB, transcribed on the opposite strand from the C4A and C4B/CYP21B genes, respectively. XA expresses an ~2.6 kb mRNA species in adrenal tissue and XB expresses an mRNA species of at least 3.5kb, primarily in steroidogenic tissues (Morel *et al.* 1989). XA contains a deletion and an insertion, relative to XB, within the coding sequence that result in a frameshift and premature termination of the putative protein product (Gitelman *et al.* 1992). The putative XB gene product contains a C-terminal domain with homology to a fibrinogen domain and three type III fibronectin repeats. Potential coding sequence for three additional type III

fibronectin repeats was also identified upstream of the 5′ end of the longest, 2.7kb, cDNA clone (Morel *et al.* 1989).

By sequencing genomic DNA fragments located centromeric of XB and by hybridisation with a partial XB cDNA sequence encoding two of the type III fibronectin repeats, Matsumoto *et al.* (1992a,b) identified four potential type III fibronectin repeats oriented in the same direction as XB and hybridisation data for the existence of a further four repeats. Additional 5′ sequence data has also revealed 18.5 potential EGF repeats (Matsumoto *et al.* 1992b). This arrangement of tandem EGF repeats, tandem type III fibronectin repeats and a C-terminal fibrinogen domain is homologous to tenascin, a multidomain and multifunctional extracellular matrix glycoprotein, which contains 15 EGF repeats, 15 type III fibronectin repeats and a fibrinogen domain at the C-terminus (Siri *et al.* 1991). In addition, partial sequence of the putative G12 protein (Kendall 1992) shows significant homology to the sequences determined by Morel *et al.* (1989) and Matsumoto *et al.* (1992a,b) and to that of tenascin. Matsumoto *et al.* (1992a,b) have proposed that G12 and XB are parts of the same tenascin-like gene. However, there is no evidence for the presence of coding sequence for the heptad repeats, present at the N-terminus of tenascin, in the MHC class III region and no explanation has been offered for the independent expression of XA, homologous to the 3′ end of XB.

Tenascin is transiently expressed in many organs during development and can mediate intercellular interactions, including those of the immune system.

Although homologous proteins and homologous domains have been identified for many of the putative protein products encoded by the novel genes in the MHC class III region, the significance of the sequence homologies has yet to be determined. The only functional data obtained is the zinc-binding property of the C-terminal region of G9a that is homologous to the *Drosophila* trithorax protein (Milner and Campbell 1993). The relationships between other sequence homologies and structural motifs with respect to protein function remain to be established.

1.6 POLYMORPHISM IN THE MHC

The HLA antigens are highly polymorphic proteins and the loci encoding them constitute one of the most highly polymorphic gene clusters identified (Bjorkman and Parham 1990). Originally identified using serological and cell typing methods, the alleles can now be identified by nucleic acid sequencing (Bell *et al.* 1987) and typed using PCR and allele specific oligonucleotides (Todd *et al.* 1987).

1.6.1 The class I antigens

The mixed lymphocyte reaction (MLR) and primed lymphocyte typing (PLT) have currently been used to define 28 HLA-A, 59 HLA-B and 10 HLA-C serological specificities (WHO Nomenclature Committee 1992). Additional genetic polymorphisms have also been observed and, as a result, at the class I polymorphic loci, HLA-A, B, C and E, 41 HLA-A alleles, 61 HLA-B alleles, 18 HLA-C alleles and 4 HLA-E alleles have currently been identified (WHO Nomenclature Committee 1992).

1.6.2 The class II antigens

The class II α and β chains encoded by the A and B genes, respectively, in the DP, DQ and DR subregions encode the following number of recognised HLA specificities: Dw26, DR 24, DQ 9 and DP 6 (WHO Nomenclature Committee 1992). At the DRA locus only two alleles have been identified whereas at the DRB loci 72 alleles have been identified that are responsible for virtually all the DR polymorphisms. Both the DQA and DQB loci are polymorphic and 14 DQA and 19 DQB alleles have been identified, while at the DP locus 8 DPA and 8 DPB alleles have been identified (WHO Nomenclature Committee 1992).

1.6.3 The MHC class III region

1.6.3.1 *The complement genes*

The complement components C2 C4 and factor B are polymorphic at both the DNA and protein levels. The C4 isotypes, C4A and C4B, exhibit extensive structural polymorphism (Campbell *et al.* 1990), but there are only 15 amino acid differences between C4A and C4B (Belt *et al.* 1985; Yu *et al.* 1986) which are responsible for their

different haemolytic activities, covalent binding activities and antigenic determinants. There are at least 33 polymorphic C4 variants (Mauff *et al.* 1983), but there is no obvious variation in their function, except for the C4A6 allele product which is haemolytically inactive (Dodds *et al.* 1985).

Recombination has been shown to occur in the C4/CYP21 region and deletion or duplication of the null, C4AQ0 and C4BQ0, alleles has been observed. Identical alleles can occur at the C4A and C4B loci and deletion or duplication of the C4 loci can result in haplotypes containing between one and four C4 genes (Campbell *et al.* 1990).

Factor B is polymorphic with two major variants, F and S, and two minor variants, F1 and S1, as well as 14 very rare variants having been identified (Dykes *et al.* 1983). C2 polymorphism is very limited as the major variant C2C has a frequency of 0.97 whilst rare allotypes C2A and C2B have frequencies of 0.01 and 0.02, respectively (Alper 1976).

1.6.3.2 *Other genes in the MHC class III region*

Polymorphic analysis of other genes in the MHC class III region and their protein products has not been extensive. Two variants of the TNF-B gene have been identified, on the basis of an *Nco*I RFLP (Abraham *et al.* 1991; Messer *et al.* 1991) that differ by a single amino acid residue.

Analysis of the HSP70 genes for polymorphism in a limited number of haplotypes revealed very limited sequence variation (Milner and Campbell 1992a). Six base changes were observed between HSP70-1 and HSP70-2 which occur in the 5′ flanking region of HSP70-1 and the 5′ untranslated regions and coding regions of both genes. However, none of these is associated with any amino acid polymorphism. In the HSP70-Hom gene two base changes were observed in the coding region, one of which is associated with an amino acid change.

G4 has been sequenced from three different haplotypes and nucleotide variations indicate that this gene may be polymorphic (Cheng 1991).

1.6.3.3 *RFLPs and microsatellite repeats*

RFLP analysis of an ~100kb region extending centromeric of TNF-A has been carried out by Cross *et al.* (1991) and six RFLPs were detected using cosmid inserts as hybridisation probes. Four of these are located within the region of DNA that contains the G2 and G3 genes, one is located within the region that contains the G1

gene and the other is located within the region that contains the G4 and G5 genes.

Twelve $(CA)_n$ dinucleotide repeats have been mapped within the MHC class III region (Hsieh 1992; Colonna *et al.* 1993). Ten of these microsatellite repeats are highly polymorphic and may be used to study disease associations across the region. A variable number of tandem repeats (VNTR) has also been identified at a locus within the third intron of the C2 gene (Zhu *et al.* 1992) and five different alleles have been observed by RFLP analysis (Cross *et al.* 1985; Ratanachaiyavong *et al.* 1989).

1.6.4 Linkage disequilibrium and extended haplotypes

Although the MHC class I and class II loci are highly polymorphic, they are very closely linked. As a result, particular alleles exhibit a tendency to remain associated rather than separate by recombination so that the frequency of a pair of alleles occurring together is greater than the product of their individual allelic frequencies. This phenomenon is known as linkage disequilibrium. The resulting combinations of alleles that have been observed are termed extended haplotypes and approximately 30 have been identified which account for most linkage disequilibria observed. The combinations of alleles may confer advantages upon an individual *e.g.* particular class II TAP alleles and class I alleles, or may remain linked due to infrequent recombination within the MHC.

Many extended haplotypes, or supratypes, have specific disease associations (Dawkins *et al.* 1983) and the linkage disequilibrium, or low recombination frequency, within the MHC has made recombination mapping of the associated diseases more difficult than in regions of the human genome where linkage disequilibrium does not occur.

1.7 MHC AND DISEASE

The human MHC has been shown to be associated with genetic susceptibility to a large number of diseases (Tiwari and Terasaki 1985; Batchelor and McMichael 1987; Thomson 1988). Patients with these diseases show an increased frequency of certain MHC alleles compared with normal individuals. However, as a result of linkage disequilibrium, some diseases are associated with loci encoding several different

antigens *i.e.* a particular extended haplotype (Table 1). Sometimes this can be overcome by analysing patients from several different ethnic groups which may contain different haplotype combinations. However, most MHC associated diseases are both polygenic and multifactorial such although the genetic factors may cause susceptibility to a disease, environmental factors are also important in the pathogenesis of the disease.

The MHC-associated diseases can be divided into three categories: diseases that are not due to chronic autoimmunity, diseases that involve autoimmunity and diseases of unknown pathogenesis.

1.7.1 Diseases in which immunological mechanisms are not involved

Several diseases of non-immunological pathogenesis are known to be associated with genes in the MHC. Congenital adrenal hyperplasia is a metabolic disorder caused by deficiency of cytochrome P450 steroid 21-hydroxylase activity as a result of deletion or mutation of the CYP21B gene (Rodrigues *et al.* 1987; Collier *et al.* 1989).

Infantile sialidosis type II, neuraminidase deficiency, was found to be associated with congenital adrenal hyperplasia in a Japanese family (Oohira *et al.* 1985) and one of the novel genes in the MHC, G9, shares sequence homology with a bacterial sialidase.

Idiopathic haemochromatosis is characterised by abnormal iron metabolism and is associated with the HLA-A3 antigen, and to a lesser extent with HLA-B7 and B14, although the disease pathogenesis is unknown.

1.7.2 Autoimmune diseases

This is the largest group of diseases associated with the MHC, but the direct involvement of the MHC in disease pathogenicity is not known, even though some disease associations are very strong.

Ankylosing spondylitis, or rheumatoid spondylitis, is characterised by inflammation of the spinal joints, but the disease aetiology is unknown. Incidence of the disease is very strongly associated with the HLA-B27 antigen which is present in 95% of patients, but only 7% of normal controls. However, only ~2% of individuals with HLA-B27 suffer from ankylosing spondylitis, implicating the importance of

Disease	Race	Patients (% positive)	Controls (% positive)	Relative Risk
Ankylosing Spondylitis (AS)				
B27	C	89	9	69.1
B27	O	85	15	207.9
B27	N	58	4	54.4
Idiopathic Haemochromatosis				
A3	C	72	28	6.7
B7	C	48	26	2.9
B14	C	19	6	2.7
Coeliac Disease				
B8	C	68	22	7.6
DR3	C	79	22	11.6
DR7	C	60	15	7.7
Insulin Dependent Diabetes Mellitus (IDDM)				
B8	C	40	21	2.5
B15	C	22	14	2.1
DR3	C	52	22	3.8
DR4	C	74	24	9.0
Rheumatoid Arthritis (RA)				
B27	C	16	9	2.0
DR4	C	68	25	3.8
B54	O	24	11	2.5
DR4	O	66	39	2.8
DR4	N	40	10	5.4
Systemic Lupus Erythematosus (SLE)				
B8	C	40	20	2.7
DR3	C	42	21	2.6
Narcolepsy				
DR2	C	100	22	129.8
DR2	O	100	34	358.1

C = Caucasian O = Oriental N = Negro

Table 1 Some HLA disease associations (data from Tiwari and Terasaki 1985; Thomson 1988)

environmental factors in the pathogenesis of this disease. The *Klebsiella pneumoniae* nitrogenase and the B27 antigen share six consecutive amino acids (Oldstone 1987) and this molecular mimicry may play a role in ankylosing spondylitis pathogenesis.

Rheumatoid arthritis involves inflammatory changes in the synovial membrane and is associated with the DR4 specificity. This association is stronger with the Dw4 and Dw14 subtypes and a ten amino acid epitope, present in the DRβ chain membrane distal domain of several of the associated haplotypes, is present in 83% of patients and 46% of controls (Wordsworth *et al.* 1989).

Insulin dependent diabetes mellitus (IDDM), or type I diabetes, is characterised by selective and irreversible destruction of insulin producing pancreatic β cells. This is the best characterised of the class II associated diseases. Sixty per cent of IDDM patients possess B8 or B15 specificities and 95% possess DR3 or DR4, with DR3/DR4 heterozygotes being at greatest risk (Nepom and Erlich 1991). Sequence analysis of HLA-DQ and DR alleles, however, has revealed that the DQβ chains of IDDM patients contain Ala, Val and Ser at position 57, correlating with susceptibility to the disease, and the presence of Asp at this position correlates with reduced susceptibility to IDDM in some haplotypes and dominant protection in others (Todd 1990). In the proposed MHC class II antigen structure modelled upon the class I structure (Brown *et al.* 1988), position 57 is located at one end of the peptide binding groove and forms a salt bridge with a conserved Arg residue at position 79 in the DQα chain (Todd *et al.* 1987). The absence of this salt bridge could conceivably allow the structure to become more permissive and allow presentation of peptides that result in IDDM pathogenesis. However, there are low levels of concordance of IDDM in monozygotic twins and HLA identical siblings, indicating that other genetic or environmental factors are also required for IDDM pathogenesis.

The possible role of an environmental factor in an autoimmune disease has been proposed for coeliac disease. Associated with the B8/DR3 haplotype and the DR7 serological specificity, this syndrome is characterised by malabsorption following ingestion of gluten-containing foods. The A-gliadin peptide of gluten shares significant sequence similarity with the E1B protein of the intestinal adenovirus Ad–12 and an immune response against the virus could result in the induction of T

cells cross reactive with gluten (Shanahan and Weinstein 1988).

One disease shown to be associated with the MHC class III region is systemic lupus erythematosus (SLE). This is a chronic inflammatory disorder of the connective tissue and affected individuals contain autoantibodies against nuclear and cytoplasmic cellular components. Deficiency of the classical complement pathway components, including C2 and C4, and the terminal components, C5-8, is associated with idiopathic SLE, but C2 and C4 null alleles in particular occur with significantly increased frequencies in SLE patients (Batchelor and McMichael 1987). One of the functions of the complement system is the clearance of immune complexes and deficiency of C2 or C4 may result in the accumulation of immune complexes leading to inflammation, cell destruction and release of autoantigens characteristic of SLE patients.

The advent of sequence analysis of the peptides bound to MHC molecules (Jardetzky *et al.* 1991; Rudensky *et al.* 1991) may assist in the identification of peptides, the specific antigens to which they are bound and the environmental factors that play a role in the pathogenesis of many of the MHC-linked autoimmune diseases.

1.7.3 Diseases of uncertain pathogenesis

Narcolepsy, a neurological disorder characterised by recurrent, uncontrollable brief episodes of sleep, is ~100% associated with haplotypes containing the DR2 specificity.

Psoriasis is a common, chronic squamous dermatosis known to have a polygenic inheritance pattern although the B13, B17, B37, Cw6 and DR7 haplotypes do show association with the condition in both caucasian and oriental populations.

1.8 GENETIC TRAITS AND LOCI MAPPED WITHIN THE MHC

Many diseases have been shown to be associated with the MHC, however, even in cases where the association is strong it is not clear whether it is the associated allele or a closely linked allele, in linkage disequilibrium with the associated allele, that is responsible for disease susceptibility. In both the human and mouse MHCs several loci have been mapped to specific regions of DNA.

1.8.1 Loci mapped within the human MHC

In man, the *EC1* locus, controlling natural killer cell alloreactivity, was mapped between the complement factor B gene and the HLA-A locus (Ciccone *et al.* 1990) and subsequent analysis has shown association of the *EC1* locus with HLA-C or a closely linked gene (Colonna *et al.* 1992b).

Common variable immunodeficiency (CVID) and IgA deficiency (IgA-D) have been mapped to the MHC class III region by association with C4A and CYP21A deletions and C2 RFLPs (Schaffer *et al.* 1989). This supports the hypothesis that the two conditions are related disorders, susceptibility to which is determined by a single locus. Subsequent familial analysis of CVID and IgA-D suggests that a locus controlling these conditions is located between the C4B and C2 genes (Volonakis *et al.* 1992).

1.8.2 Loci mapped in the mouse *H-2* complex

In the mouse *H-2* complex several loci have been mapped, in particular between the *Hsp70* gene cluster and the *Tnfa* gene, where recombinations have been mapped in congenic mouse strains. The major haemopoietic histocompatibility antigen regulatory locus, *Hh-1*, has been mapped centromeric of the *H-2D* region (Rembecki *et al.* 1988), within a region that corresponds to the position within which the *EC1* locus has been mapped in man, supporting the hypothesis that the *EC1* and *Hh-1* genes may encode related products.

Control of the antibody response to TNP-ficoll has been mapped between *H-2S* and *H-2D* (Hillstrom-Shapiro *et al.* 1991) to the interval between *C2* and *Bat5* (Snoek *et al.* 1991). The locus encoding corticosteroid induced cleft palate susceptibility, *Cps-1*, was mapped to the *H-2D* region (Demant 1985) and subsequently to the *Hsp70.1-Bat5* interval (Snoek *et al.* 1991).

One locus of particular interest is the *Orch-1* locus which plays a role in controlling susceptibility to the induction of experimental autoimmune orchitis (EAO). Originally mapped towards the *H-2D* region (Teuscher *et al.* 1985), this locus has since been mapped to the *Hsp70.1-G7* interval (Snoek *et al.* 1991, 1993).

1.9 EXPERIMENTAL AUTOIMMUNE ORCHITIS AND THE *Orch-1* LOCUS

1.9.1 Infertility and orchitis

Idiopathic infertility is believed to occur as a result of the presence of anti-sperm antibodies, aspermatogenesis or stenosis of the epididymis or vas deferens arising from orchitis or epididymovasitis. Orchitis is characterised by inflammation of the testis marked by pain and swelling and, in man, may occur idiopathically or as a result of infection e.g. mumps, tuberculosis or syphilis, often resulting in infertility. To determine the immune mechanisms responsible for the pathogenesis of idiopathic infertility, both clinical and experimental analyses have been carried out. Focal monocytic orchitis in infertile men has been associated with anti-sperm autoantibodies which are also produced in ~50% of men following vasectomy. The phenotypic expression of infertility is characterised by multiple testicular abnormalities associated with an immune response to sperm or other testis specific antigens, including deposition of IgG and C3 around the seminiferous tubules.

Animal models of spontaneous autoimmune infertility have been seen in partially inbred 'A' line beagle dogs, T/t^{w18} backcross mice and black mink (Fritz *et al.* 1976; Dooher *et al.* 1981; Tung *et al.* 1981). The phenotypic expression of infertility in these models resembles idiopathic azoospermia or oligozoospermia and is characterised by multiple testicular abnormalities accompanied by autoimmune responses to sperm and testis-specific antigens. A similar testicular autoimmune disease occurs in guinea pigs, rabbits and rhesus macaques following vasectomy (Tung 1978; Bigazzi *et al.* 1976; Tung and Alexander 1980).

The best characterised animal model is murine experimental autoimmune orchitis (EAO) which can be reproducibly induced in BALB/c mice. In this model, dialysed, lyophilised, mouse testicular homogenate, emulsified in complete Freund's adjuvant, is used to immunise mice which simultaneously receive 10μg of *Bordetella pertussis* extract, rich in pertussigen (Kohno *et al.* 1983). Active EAO is then induced in the mice 20 days later, manifested as inflammation of the testes, aspermatogenesis, epididymovasitis and vasitis. Passive EAO can be induced by adoptive transfer of testis-sensitised lymphocytes (Mahi-Brown *et al.* 1987).

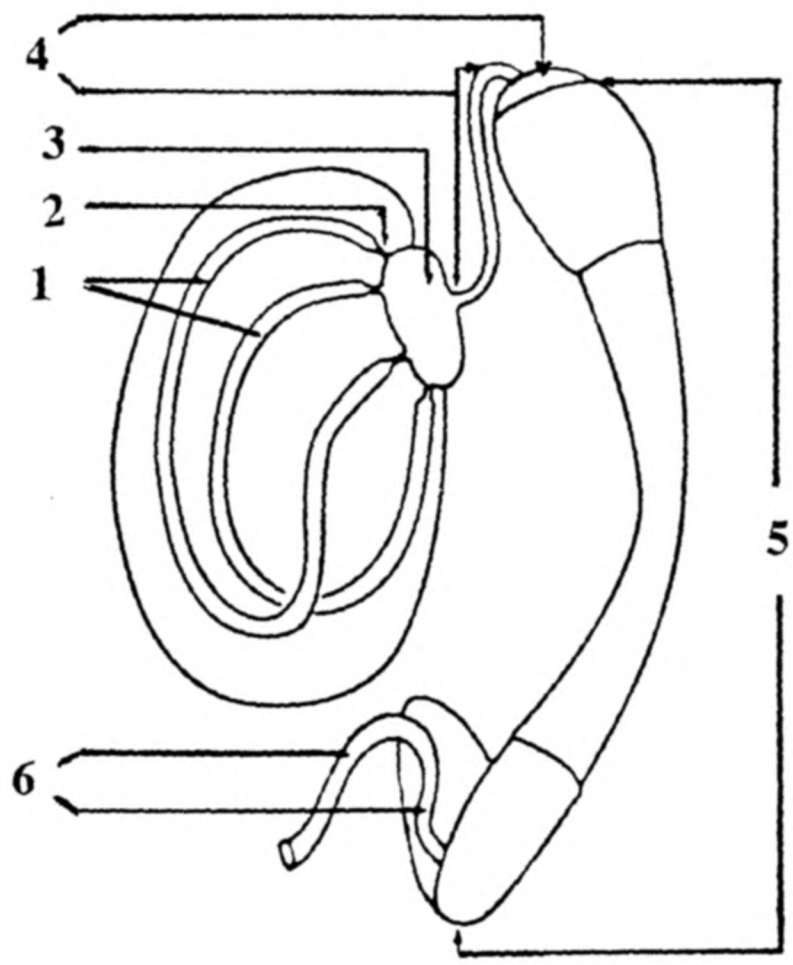


Figure 1.6 Morphology of the mouse testis.
Seminiferous tubules (**1**), straight tubules (**2**), rete testis (**3**), ductus efferentes (**4**), epididymis (**5**) and vas deferens (**6**)

1.9.2 Immunopathology of orchitis

Active EAO is characterised by inflammatory lesions located in the testis (100%), rete testis (37%), cauda epididymis (21%) and vas deferens (37%) (Figure 1.6) and autoantibodies to testicular antigens in the serum (Kohno *et al* 1983). However, there is no correlation between the histopathological changes and antibody-dependent cytotoxicity or antibody titres despite the deposition of IgG1 and IgG3 on germ cells outside the cells of the blood-testis barrier (Person *et al.* 1993).

In passive EAO, maximum orchitis is located in the straight tubules, rete testis and ductus efferentes (Figure 1.6) which produce maximal periductal and epithelial MHC class II antigen expression. Therefore, active and passive EAO are believed to arise by different immunopathogenic mechanisms.

The early lesions formed during active EAO contain eosinophils, neutrophils lymphocytes and macrophages both in and around the seminiferous tubules. This is followed by aspermatogenesis and desquamation of spermatocytes and spermatids. The late lesions are characterised by necrosis and fibrosis of the seminiferous tubules and disruption of the blood-testis barrier (Kohno *et al.* 1983).

The role of *Bordetella pertussis* extract in EAO is unknown, but in addition to lymphocytosis and a possible role as an adjuvant, it is known to cause histamine sensitivity of the blood-brain barrier and increase tissue barrier sensitivity to vasoactive substances (Bergman *et al.* 1978). Therefore, the *Bordetella pertussis* extract may be required to sensitise the testis-blood barrier for cellular infiltration rather than in the inflammatory stage of EAO induction, even though germ cells located outside the blood-testis barrier express autoantigens that are accessible to the vascular and lymphoid systems.

1.9.3 Immunogenetics of EAO

The role of genetic factors in infertility and orchitis was originally suggested following the observation that the development of anti-sperm antibodies after vasectomy is strongly associated with the MHC. HLA-A28 and B40 frequencies are increased among azoospermic patients and HLA-B7 is associated with anti-sperm antibodies in infertile couples. These associations imply that a common locus closely linked to HLA-A and HLA-B may control the two conditions (Person *et al.* 1993). In addition, such immune responses in animals have been mapped to MHC and non-

MHC loci (Teuscher *et al.* 1985, 1990), suggesting that control of autoimmune infertility is polygenic.

In order to study orchitis and generate a better understanding of human fertility, the loci controlling EAO in mouse have been studied. Using inbred mouse strains it has been observed that susceptibility to EAO is controlled by a locus or loci in the *H-2* complex. *H-2*^{d} genetic backgrounds correlate with susceptibility to EAO and *H-2*^{k} or *H-2*^{b} correlate with low susceptibility or resistance to EAO (Teuscher *et al.* 1985). Subsequent analysis of a large panel of intra *H-2* recombinant congenic mouse strains that possess various defined crossovers within the *H-2* region mapped the orchitis susceptibility gene-1 (*Orch-1*) with the *H-2*^{d} alleles as a single dominant autosomal trait between the *H-2S* and *H-2D* regions (Teuscher *et al.* 1990). Further analysis of intra *H-2* recombinant mouse strains by RFLP analysis, using cDNA probes corresponding to novel genes located within the human MHC class III region, has reduced the location of the *Orch-1* locus to the region between *Hsp70.1* and *G7* (Snoek *et al.* 1991, 1993).

Not all mouse strains exhibiting the susceptible *Orch-1* genotype develop EAO. Some mouse strains are resistant, indicating that other non-MHC genes play a role in controlling disease susceptibility. The orchitis susceptibility gene-2 (*Orch-2*) has been implicated in controlling EAO in backcross experiments where the *Orch-2* locus is inherited as a dominant trait (Person *et al.* 1993). A third gene, *Orch-3*, controlling resistance to EAO has been identified by segregation analysis of mouse strains that vary in the severity of orchitis. This is a recessive locus and in mice with the wild-type *Orch-3* locus, susceptibility to EAO is controlled by *Orch-1* (Person *et al.* 1993). However, this locus is also implicated in passive EAO following adoptive transfer. *Orch-2* and *Orch-3* both appear to play a role in immune suppression, *Orch-2* is a non-*H-2* complex dominant gene that does not behave as a T cell suppressor (Person *et al.* 1993) and *Orch-3* is a recessive locus controlled by the *Orch-1* locus. The *Orch-1* locus itself has been postulated to be a second order Ir gene regulating the level of proliferation or development of the effector mechanisms involved in testicular lesions (Person *et al.* 1993).

1.10 AIMS OF THE PROJECT

A large number of diseases have been shown to be associated with the human MHC (Tiwari and Terasaki 1985). In addition, several genetic traits have been mapped within the corresponding mouse H-2 complex. These include the *Orch-1* (Teuscher *et al.* 1990) and *Cps-1* (Demant 1985) loci which have both been mapped to the same ~100kb region between the *Hsp70* gene cluster and the *Bat5* gene in the mouse *H2-S* region (Snoek *et al.* 1991). The corresponding region of the human MHC class III region contains the G6, G7 and G7a genes. G6 and G7a have been characterised and G7a has been shown to encode valyl-tRNA synthetase (Hsieh and Campbell 1991). In order to search for the human homologues of the *Orch-1* and *Cps-1* loci, and to investigate the potential roles of additional genes between HSP70-Hom and BAT5 in autoimmune disease pathogenesis, the following approaches were adopted:

1. The novel gene G7 was characterised at the cDNA and genomic levels in order to determine the protein product. The derived amino acid sequence was analysed to investigate the potential functions of the G7 protein product and its possible role in MHC-associated diseases.

2. The intergenic gaps within the HSP70-Hom to BAT5 interval were investigated to search for additional genes in the region and two new genes, G7b and G7c, were discovered. G7b was characterised at the cDNA and genomic levels to determine the protein product. The derived amino acid sequence was analysed in order to investigate the potential functions of the G7b protein product and its possible role in MHC-associated diseases.

3. The mouse homologue of a candidate gene for the *Orch-1* locus was analysed for polymorphism in mouse strains shown to be susceptible or resistant to EAO.

CHAPTER 2

MATERIALS & METHODS

2.1 MATERIALS

2.1.1 Chemicals and Reagents

All chemicals and reagents were obtained from BDH Chemicals Ltd. (AnalaR grade) or Sigma Chemical Company Ltd. except those listed below or specified in the text.

Acrylamide solutions; Accugel 40 Protogel	National Diagnostics
Agarose; Seakem HGT SeaPlaque LGT	FMC Laboratories
BiTek Agar BactoTryptone	Difco Laboratories
dATP, dCTP, dGTP, dTTP (100mM Li salts) BCIG Tris	Boehringer Mannheim
Dextran sulphate Ficoll 400 Sephadex G-100	Pharmacia
Guanidine isothiocyanate 1kb DNA ladder	Gibco BRL
Phenol (saturated solution pH8)	Camlab
Radioactive nucleotides; [α-^{32}P]dATP, [α-^{32}P]dCTP, [α-^{32}P]CTP specific activity >3000 Ci/mmol, [γ-^{32}P]ATP specific activity >5000 Ci/mmol, [α-^{35}S]-thio dATP specific activity >1000 Ci/mmol, all at 10mCi/ml.	Amersham

PBS (Dulbecco's 'A') pH7.3 tablets	Oxoid
Yeast extract	

2.1.2 Enzymes

Restriction enzymes were obtained from Amersham, Boehringer Mannheim, Gibco BRL, Pharmacia, Promega and New England Biolabs. Other modifying enzymes were obtained from the following suppliers.

Klenow fragment (cloned)	Amersham
T4 DNA ligase	
Reverse Transcriptase RAV2	
Calf intestinal phosphatase	Boehringer Mannheim
T4 polynucleotide kinase	
Proteinase K	Sigma
RNase A	
Sequenase (modified T7 DNA polymerase)	USB Corporation
Taq DNA polymerase	Promega

2.1.3 Reagent Kits

Reagent kits for DNA and RNA manipulation were obtained from the following sources.

cDNA Synthesis System Plus	Amersham
Multiprime DNA Labelling System	
dsDNA Cycle Sequencing system	Gibco BRL

Magic Minipreps DNA Purification System	Promega
Magic PCR Preps DNA Purification System	
Riboprobe Gemini System II	
Sequenase Version 2.0 DNA Sequencing Kit	USB Corporation

2.1.4 *E.coli* Cloning Vectors

pBluescript	Stratagene (Short *et al.* 1988)
pBC	Stratagene (catalogue 1991/2)
CDM8	gift from D Simmons (Seed 1987)
pDVcos	(Knott *et al.* 1988)
pGEM-3Zf	Promega (Sambrook *et al.* 1989)
pGEX-2T	Amrad Corporation (Smith and Johnson 1988)
M13mp10	Amersham (Messing and Vieira 1982)
pSPL1	gift from A.J. Buckler (Buckler *et al.* 1991)

2.1.5 *E. coli* Strains

MC1061/p3	gift from D. Simmons (Sambrook *et al.* 1989)
NM554	(Raleigh *et al.* 1988)
TG1	(Sambrook *et al.* 1989)

2.1.6 Standard buffers and solutions

4% native acrylamide gel mix	4% acrylamide (Accugel 40) in 1x TBE polymerised with 0.08% (w/v) APS and 0.25% (v/v) TEMED
6% denaturing acrylamide gel mix	6% acrylamide (Accugel 40) 7M urea in 1x TBE polymerised with 0.08% (w/v) APS and 0.25% (v/v) TEMED
Formamide	deionised using Duolite MB 6113 mixed resin and filtered through Whatman 50 paper
Formamide loading dyes	10mM EDTA pH7.4 0.1% (w/v) bromophenol blue 0.1% (w/v) xylene cyanol in formamide

Filter hybridisation buffer	50% (v/v) formamide 1M NaCl 50mM Tris-HCl pH7.4 10% (w/v) dextran sulphate 0.2% (w/v) BSA 0.2% (w/v) Ficoll 400 0.2% (w/v) polyvinylpyrrolidone 0.1% (w/v) sodium pyrophosphate 0.1% (w/v) SDS
Glycerol loading dyes	30% (v/v) glycerol 5mM EDTA pH7.4 0.1% (w/v) bromophenol blue 0.1% (w/v) xylene cyanol
Phenol, (for genomic DNA preparations)	solid phenol was dissolved in 1M Tris-HCl pH8 and equilibrated in 10mM Tris-HCl/1mM EDTA until pH8
Prewash buffer	1M NaCl 50mM Tris-HCl 1mM EDTA pH8.0
TE	10mM Tris-HCl 0.1mM EDTA pH7.4
10x TBE	0.9M Tris 0.9M Boric acid 20mM EDTA
20x SSC	3M NaCl 0.3M tri-sodium citrate
MTPBS	150mM NaCl 16mM Na_2HPO_4 4mM NaH_2PO_4 pH7.3
SDS-polyacrylamide gels:	
10% separating gel	10% (w/v) acrylamide (Protogel) 0.375M Tris-HCl pH8.8 0.025% (w/v) SDS polymerised with 0.1% (w/v) APS and 0.1% (v/v) TEMED

5% stacking gel	5% (w/v) acrylamide (Protogel) 0.125M Tris-HCl pH6.8 0.025% (w/v) SDS polymerised with 0.1% (w/v) APS and 0.1% (v/v) TEMED
10x running buffer	0.25M Tris 1.9M glycine pH8.2 0.1% (w/v) SDS added to 1x buffer prior to use
2x sample loading buffer	6% (w/v) SDS 20% (v/v) glycerol 1.4M β-mercaptoethanol 0.12M Tris-HCl pH6.8 0.025% (w/v) bromophenol blue
Coomassie blue stain	0.05% (w/v) Coomassie brilliant blue R (BioRad) 40% (v/v) methanol 10% (v/v) acetic acid
Destain	7% (v/v) methanol 5% (v/v) acetic acid
Western blot transfer buffer	50mM Tris 40mM glycine 0.05% (w/v) SDS

All other buffers and solutions are specified in the text.

2.1.7 Bacterial culture media

L broth	1% (w/v) BactoTryptone 1% (w/v) NaCl 0.5% (w/v) yeast extract
L plates	As for L broth with 1.5% (w/v) agar and 10mM $MgSO_4.7H_2O$
Top agar	As for L broth with 0.7% (w/v) agar
2x TY broth	1.6% (w/v) BactoTryptone 1% (w/v) yeast extract 0.5% (w/v) NaCl

2x TY plates	As for 2x TY broth with 1.5% (w/v) agar
HMF-1 plates	0.5% (w/v) yeast extract 1.0% (w/v) BactoTryptone 1.0% (w/v) NaCl 1.5% (w/v) agar 10mM $MgSO_4.7H_2O$ 3.6mM K_2HPO_4 1.3mM KH_2PO_4 2mM tri-sodium citrate 5% (w/v) glycerol
HMF-2 plates	As for HMF-1 plates with 25% (v/v) glycerol
10x Hogness buffer	40% (v/v) glycerol 10mM $MgSO_4.7H_2O$ 36mM K_2HPO_4 13mM KH_2PO_4 pH7.4 20mM tri-sodium citrate
Ampicillin stock	20mg/ml ampicillin in 1M $NaHCO_3$ (filter sterilised)
Chloroamphenicol	30mg/ml in ethanol (filter sterilised)
Kanamycin	10mg/ml in water (filter sterilised)
Tetracycline	10mg/ml in 50% ethanol (filter sterilised)

$MgSO_4$ was excluded from media or plates when tetracycline was used.

LA plates contained ampicillin at 100µg/ml.

LAT plates contained ampicillin at 20µg/ml and tetracycline at 10µg/ml.

LCm plates contained chloroamphenicol at 30µg/ml.

In liquid cultures ampicillin was added at 50µg/ml, chloroamphenicol at 30µg/ml, kanamycin at 75µg/ml and tetracycline at 10µg/ml.

2.2 STORAGE AND GROWTH OF *ESCHERICHIA COLI* BACTERIAL STRAINS

2.2.1 Storage of bacterial strains and clones

Frozen stocks were prepared from 500µl of an overnight culture by adding 55 µl 10x Hogness buffer and then stored at -70°C.

2.2.2 Overnight cultures

L broth or 2x TY broth was inoculated with a single colony or 10µl of glycerol stock and, for transformed strains, the appropriate amount of antibiotic was added. The culture was grown overnight in a shaking incubator, at 37°C.

2.2.3 Preparation of competent bacterial cells for immediate use

1ml of an overnight culture of the required strain was added to 10ml of 2x TY broth and incubated at 37°C for ~3 h. The cells were pelleted by centrifugation at 625g for 5 min at 4°C and gently resuspended in 2ml ice cold $CaCl_2$ solution (50mM $CaCl_2$, 10mM Tris-HCl pH7.4). After incubation on ice for 30 min, the cells were repelleted and gently resuspended in 2ml ice cold $CaCl_2$ solution and stored on ice for >30 min before use.

2.2.4 Preparation of frozen competent bacterial cells

A single colony from the required bacterial strain, streaked out and grown on an L plate, was used to inoculate 5ml 2x TY broth containing 10mM $MgSO_4$ and grown at 37°C for 2 h. 100ml of 2x TY broth containing 10mM $MgSO_4$ was added to the culture and grown at 37°C until the OD_{600} was 0.5-0.9. The cells were pelleted by centrifugation for 15 min at 625g at 4°C, resuspended in 30ml cold TfbI buffer (30mM potassium acetate, 50mM $MnCl_2$, 100mM KCl, 10mM $CaCl_2$, 15%(v/v) glycerol, filter sterilised) and incubated on ice for 10 min. The cells were repelleted by centrifugation at 625g for 10 min at 4°C and resuspended in 4ml cold TfbII buffer (10mM Na-MOPS, pH7.0, 75mM $CaCl_2$, 10mM KCl, 15%(v/v) glycerol, filter sterilised). The cells were aliquoted into prechilled 1.5ml tubes, frozen on dry ice and stored at -70°C.

2.3 PREPARATION OF NUCLEIC ACIDS

2.3.1 Recovery of nucleic acids from solution

2.3.1.1 Organic extraction

To remove proteins from nucleic acid solutions an equal volume of phenol or phenol/chloroform/isopropanol (25:24:1) was added and vortexed together. After centrifugation at 12,000 rpm for 5 min in a microfuge (or 625g in a Beckman GPR centrifuge) the aqueous, upper layer was removed to a fresh tube and subsequently extracted with an equal volume of chloroform/ isopropanol (24:1).

2.3.1.2 Ethanol precipitation of nucleic acids

To concentrate nucleic acids, 0.1 volume of 3M sodium acetate pH7 and 2.5 volumes of cold ethanol were added to one volume of aqueous solution. The nucleic acids were precipitated at -20°C overnight or at -70°C for 10-30 min before recovery by centrifugation at 12,000 rpm for 5-10 min in a microfuge. Nucleic acid pellets were washed with 70% (v/v) ethanol and air dried before resuspending in the appropriate buffer.

2.3.2 Small scale preparation of plasmid and cosmid DNA

2.3.2.1 Alkali lysis method (Birnboim and Doly 1979)

A 10ml overnight culture [section 2.2.2] was centrifuged at 625g for 15 min and the cell pellet was resuspended in 200µl lysis buffer (50mM glucose, 10mM EDTA, 25mM Tris-HCl pH8) before transferring to a 1.5ml microfuge tube. The cell suspension was mixed, by inversion, with 400µl of freshly prepared 0.2M NaOH/1% (w/v) SDS until the solution had cleared. 300µl of 3M sodium acetate pH4.8 was then added and mixed to precipitate cell debris and genomic DNA. The debris was pelleted by centrifugation at 12,000 rpm for 10 min in a microfuge and the plasmid/cosmid DNA was precipitated from the supernatant, using 600µl isopropanol, at -70°C for 10 min. This was recovered by centrifugation and resuspended in 200µl 0.3M sodium acetate pH7 before phenol/chloroform extracting and reprecipitating with isopropanol, as above. After washing with 70% (v/v) ethanol, the pellet was resuspended in 50µl TE pH7.4 before RNase digestion with 20µg/µl heat inactivated

RNase A (Sigma Type III-A) at 37°C for 30 min. The DNA, 0.5–1µg/µl, was stored at –20°C.

2.3.2.2 Magic minipreps DNA purification (Promega)

For reliable, rapid purification of plasmid DNA for double-stranded sequencing, plasmid minipreps were carried out from 5ml overnight cultures [section 2.2.2] using the Magic miniprep DNA purification system, as described in Promega Technical Bulletin No.117, using disposable syringes.

2.3.3 Large scale preparation of plasmid DNA (Sambrook *et al.* 1989)

A 10ml culture [section 2.2.2], grown for ~5 h at 37°C in L broth containing the appropriate antibiotic, was used to inoculate 500ml of L broth containing antibiotic which was then grown overnight at 37°C in a shaking incubator. The cells were pelleted by centrifugation at 4,225g for 10 min at 4°C in a Sorvall GS-3 rotor (RC-5b centrifuge DuPont instruments) and resuspended in 10ml lysis buffer [section 2.3.2.1] containing 2mg/ml lysozyme. After incubation on ice for 30 min, 40ml of 0.2M NaOH/1% (w/v) SDS was added, mixed and incubated on ice for a further 5 min. 30ml of 3M sodium acetate pH4.8 was added and after 30 min incubation on ice the bacterial debris and genomic DNA was pelleted by centrifugation at 16,755g at 4°C for 10 min. Plasmid DNA was precipitated from the supernatant, using 2 volumes of cold ethanol, at -70°C for 1 h and pelleted by centrifugation for 10 min at 16,755g and 4°C.

The pellet was resuspended in 6ml of SET buffer (150mM NaCl, 5mM EDTA, 50mM Tris-HCl pH8) containing 0.015% (w/v) ethidium bromide. CsCl, 7g of Optical Grade (Sigma), was dissolved in the DNA solution in order to adjust the density to ~1.6g/ml. The solution was centrifuged in sealed polyallomer tubes overlaid with paraffin at 206,000g and 18°C for >12 h with acceleration=5 and deceleration=5 (Ti70.1 rotor, Beckman L8 ultracentrifuge) and the gradient was relaxed at 138,000g for 1 h before stopping with no brake.

The lower, closed circular, DNA band was removed from the gradient using a 19G gauge syringe needle. The ethidium bromide was removed by repeated extractions with an equal volume of water-saturated butan-1-ol. Three volumes of water were added and the sample precipitated in Corex tubes (DuPont), using 2 volumes of ethanol, for 1 h at -70°C. The DNA was recovered by centrifugation for 30 min at

12,000g in a Sorvall SS-34 rotor at 4°C, resuspended in 400µl 0.3M sodium acetate pH7 and reprecipitated using 1ml ethanol overnight at -20°C. The recovered DNA was resuspended in 100µl TE pH7.4 and stored at 4°C.

2.3.4 Preparation of genomic DNA (Bell *et al.* 1981)

10ml of whole blood, containing 10mM EDTA, was centrifuged at 625g for 15 min and 4°C and the upper, plasma layer removed to leave the buffy coat and red blood cells. The cells were lysed by resuspending in 40ml of genomic lysis buffer (10mM Tris-HCl pH7.4, 5mM $MgCl_2$, 0.32M sucrose, 1% (v/v) Triton X-100) and the nuclei were pelleted by centrifugation for 15 min at 625g. The process was repeated until the pellet was off-white in colour. Cells grown in tissue culture (~$2x10^7$) were washed three times in PBS (Dulbecco's, Gibco BRL) before lysing once, as above. The final pellet was resuspended in 10ml of SET buffer [section 2.3.3] containing 200µg/ml freshly added proteinase K (Sigma Type XI) and 0.5% (w/v) SDS and incubated at 37°C for 12-24 h until the solution was clear. Two phenol extractions and one phenol/chloroform/isopropanol (25:24:1) extraction were carried out [section 2.3.1.1] with removal of the phenol layer using a syringe. The final aqueous layer was transferred to a fresh tube using a syringe to avoid shearing the DNA. Two volumes of cold ethanol were added and gently mixed by inversion to precipitate the DNA, which was removed from solution using a sealed, glass capillary tube. The DNA precipitate was washed with 70% ethanol, dried briefly and resuspended in 200µl TE pH7.4 at 4°C overnight.

2.3.5 Preparation of total RNA (Chirgwin *et al.* 1979)

To avoid RNA degradation all the plasticware used was sterile and stock solutions were autoclaved before use.

Tissue culture cells (10^7-10^8) were washed three times in ice cold PBS and the washed pellet resuspended in 3ml of freshly mixed 4M guanidine isothiocyanate, 0.1M β-mercaptoethanol. A 3ml cushion of 5.7M CsCl, 25mM tri-sodium citrate, 100mM EDTA pH7 was placed in the bottom of an ultraclear centrifuge tube and an equal volume was mixed with the cell suspension which was then layered on top of the cushion. The sample was overlaid with paraffin, heat sealed and centrifuged at

83,000g and 18°C for 16 h (Ti70.1 rotor, Beckman L8 ultracentrifuge) with no brake.

After centrifugation the tube was cut open and the supernatant discarded. The pellet was washed with 70% (v/v) ethanol and resuspended in 400µl TES (25mM Tris-HCl pH7.4, 1mM EDTA, 0.2% (w/v) SDS). The RNA was stored as an ethanol precipitate at -20°C.

2.3.6 Preparation of oligonucleotide primers

Oligonucleotides for use in sequencing, reverse transcription or in PCR reactions were synthesised on an Applied Biosystems 381A automated DNA synthesiser by Dr A J Day and T Gascoyne. β-cyanoethyl phosphoramidite nucleoside derivatives were added to the 5′ end of the oligonucleotide attached to a solid support on a column. The dimethoxytrityl protecting group of the last nucleotide was removed prior to completion of synthesis.

The oligonucleotides were released from the column using 2ml concentrated ammonia solution (~35%), in 0.5ml stages, at room temperature, over 1 h. Oligonucleotides were then deprotected by incubation at 55°C in 35% ammonia solution for ~14 h. 400µl aliquots of the deprotected oligonucleotide were precipitated using 1ml ethanol and 40µl 3M sodium acetate pH7, at -20°C. The DNA pellets were washed, air dried and resuspended in 100µl of water for storage at -20°C.

Oligonucleotides were selected from regions of sequence that avoided runs of a single nucleotide greater than three and that avoided G/C or A/T rich sequences. Most oligonucleotides were 20–30 bases long with $T_d \geq 60°C$. For duplexes <25bp;

$$T_d = \{4x(G+C)\} + \{2x(A+T)\}$$

2.3.7 Reverse transcription of RNA to cDNA

First strand cDNA synthesis was carried out using ~10µg total RNA using the cDNA Synthesis System Plus kit (Amersham). Synthesis was primed using an oligo dT primer (supplied in the kit) or with a sequence specific primer. Reactions were carried out as directed by the manufacturer, followed by incubation with an additional 20U of reverse transcriptase for 45 min. The reaction was stopped by heating the sample at 70°C for 10 min and then stored at 4°C.

2.3.8 Polymerase chain reaction (PCR) protocols (Saiki *et al.* 1988, Erlich *et al.* 1991)

PCR amplification was carried out in a DNA Thermal Cycler (Perkin Elmer Cetus). Reaction mixtures contained 1x *Taq* polymerase buffer (50mM KCl, 10mM Tris-HCl pH8.8, 1.5mM $MgCl_2$, 0.1% (v/v) Triton X-100) 100-200μM dNTPs, 1μM each primer, 1-4 units *Taq* polymerase and 10–200ng template in a 50 or 100μl final volume which was overlaid with paraffin. Uncloned templates (genomic DNA, first strand cDNA) and bacterial cultures were used directly whilst cloned templates were linearised by digestion prior to amplification.

PCR conditions were determined by the size and concentration of the target sequence in the template DNA and the primers used. All denaturation steps were carried out at 94°C for 30 seconds, annealing was carried out ~2°C below the lower T_d of the two primers for 30 seconds and extension at 72°C was for 1–3 min, dependent upon the length of the target sequence. 15-35 cycles were carried out followed by incubation at 72°C for 10 min to ensure all extensions had been completed.

2.3.9 Estimation of nucleic acid concentration

The concentrations of DNA and RNA in solution were determined by measuring the OD_{260} of a dilution (1:200 or 1:1,000) of the sample, in water, in a quartz cuvette. At 260nm one absorbance unit corresponds to 50μg/ml of double-stranded DNA, 40μg/ml of RNA or 37μg/ml of single-stranded DNA.

2.4 GENERAL NUCLEIC ACID MANIPULATIONS

2.4.1 Restriction enzyme digests

Restriction digests were carried out according to the suppliers recommended conditions. Spermidine was added to 5mM for all digests containing ≥50mM NaCl and those carried out >37°C were overlaid with paraffin. Cosmid, plasmid and PCR samples (0.1–1μg) were digested to completion in 1–2 h using a fivefold excess of enzyme, ensuring that ≤5% (v/v) glycerol was present.

Genomic digests were carried out with 20-30U of enzyme per 5μg DNA, in the presence of 2mM spermidine, for 24 h.

2.4.2 Fractionation of DNA by agarose gel electrophoresis (Southern 1975)

Digested DNA samples were fractionated on submerged horizontal slab gels (10 x 10 x 0.5cm) prepared by melting 0.8-2.0% (w/v) HGT agarose in 60ml TBE and ethidium bromide was added to a final concentration of 0.01% (w/v). Samples were loaded in 1x glycerol dyes and gels were electrophoresed at 75mA for 1–2 h.

Genomic DNA digests were fractionated on 0.8% (w/v) HGT agarose 24x 20x 0.5cm gels and stained by soaking in 0.001% (w/v) ethidium bromide for 30 min following electrophoresis.

Size markers were electrophoresed on gels to facilitate size determination of DNA fragments which were visualised on a UV TransIlluminator TM-20 (UVP Inc.) and photographed.

2.4.3 Isolation of DNA probes using LGT agarose

Preparative DNA digests were fractionated on 0.8–1.0% (w/v) agarose minigels in 1x TBE. Samples were loaded in 1x glycerol dyes and electrophoresed at 5mA and room temperature, or 30mA and 4°C, in 1x TBE. Agarose slices, containing DNA fragments corresponding to probes, were excised and heated at 70°C for 10 min with two volumes of deionised water. The samples were thoroughly vortexed and stored at –20°C.

2.4.4 End-filling of digested DNA with recessed 3′ termini

Following restriction enzyme digestion, 3′ recessed termini of DNA fragments were end-filled using the Klenow fragment of *E.coli* DNA polymerase I. The DNA was resuspended in 1x Klenow buffer (50mM NaCl, 10mM Tris-HCl pH7.4, 10mM $MgCl_2$, 1mM DTT, 50µg/ml BSA) with 250µM dNTPs and 2U of enzyme and incubated at room temperature for 30 min. The Klenow enzyme was heat-inactivated at 70°C for 5 min.

If the digested DNA was to be separated by polyacrylamide gel electrophoresis then 10% of the DNA was end-labelled using 5µCi [α-^{32}P]-dATP in place of dATP or [α-^{32}P]-dCTP in place of dCTP in the reaction mixture. The two reactions were pooled 10 min before the end of the reaction.

2.4.5 Fractionation of DNA by polyacrylamide gel electrophoresis

End-filled, radiolabelled DNA [section 2.4.4] was fractionated on 0.4mm thick, 3.5–6.0% (w/v) native polyacrylamide gels. The gels were cast between glass plates (35 x 20cm) one of which was treated with dimethyldichlorosilane 2% (v/v) in trichloroethane. Samples were loaded in 2x glycerol dyes and run at 30mA in 1x TBE for 1.5–3 h. After electrophoresis, gels were covered with Saran Wrap (Dow Chemical Company) and exposed for autoradiography [section 2.6.3] at room temperature, without sensitising the film, to allow detection of the DNA fragments.

2.4.6 Recovery of DNA fragments from polyacrylamide gel slices

Using the autoradiograph to locate the required DNA fragments, polyacrylamide gel slices were excised from the gel and transferred to 1.5ml microfuge tubes containing 500µl 0.3M sodium acetate pH7. The DNA was eluted overnight in a shaking incubator at 37°C. The gel slice was removed and the DNA recovered from the supernatant using three sequential ethanol precipitations. The final pellet was washed with 70% (v/v) ethanol, air dried and resuspended in TE pH7.4.

2.4.7 Fractionation of RNA by agarose gel electrophoresis (Fourney *et al.* 1988)

RNA was fractionated on submerged horizontal slab gels (24 x 20 x 0.5cm) prepared by melting 1% (w/v) HGT agarose in 255ml dH_2O, 30ml 10x MOPS buffer (0.2M MOPS, 50mM sodium acetate, 10mM EDTA pH7) and then adding 15.3ml of 37% (v/v) formaldehyde to the gel at 50°C to give a final concentration of 1.9% (v/v) formaldehyde.

10µg of total RNA samples were recovered from ethanol precipitates and resuspended in 5µl RNase-free water. The samples were denatured by heating at 65°C in 25µl RNA sample buffer (50% v/v) formamide, 16% (v/v) of 37% formaldehyde, 4% (v/v) glycerol, 1x MOPS buffer, 0.005% (w/v) bromophenol blue) with 0.03% (w/v) ethidium bromide. Samples were loaded immediately and electrophoresed at 40mA in 1x MOPS buffer. Samples were visualised and photographed as in section 2.4.2.

2.5 TRANSFER OF NUCLEIC ACIDS TO NITROCELLULOSE MEMBRANES

2.5.1 Transfer of DNA from agarose gels (Southern blotting) (Southern 1975)

Agarose gels containing fractionated genomic DNA were washed at room temperature for 30 min each in;

a) 0.25M HCl, to depurinate the DNA

b) 0.5M NaOH, 1M NaCl, to denature the DNA

c) 0.5M Tris-HCl pH7.4, 3M NaCl, to neutralise the gel

The gel was placed on sponges soaked in 20x SSC and the DNA transferred onto nylon-backed nitrocellulose (Hybond-C extra, Amersham) membrane, prewetted in 2x SSC, by the capillary action of a weighted stack of paper towels for ~24 h.

DNA fragments in minigels were blotted in a similar manner except that the gel washes were reduced to 10 min each and the blotting time was reduced to 30 min. Depurination was omitted when DNA fragments <2kb were blotted.

DNA was fixed to the nitrocellulose by baking at 80°C for 2 h or by crosslinking damp membranes in an XL-1500 UV Crosslinker (Spectronics Corporation) set on *optimal crosslink* (120mJ/cm^2).

2.5.2 Transfer of RNA from agarose gels (Northern blotting) (Fourney *et al.* 1988)

RNA gels were washed at room temperature for 10 min in 0.05M NaOH, 1x SSC then twice for 20 min in 10x SSC. The RNA was transferred onto nylon-backed nitrocellulose (Hybond-C extra, Amersham) membrane, prewetted in 10x SSC, as described in section 2.5.1, using 10x SSC transfer buffer for ~24 h. The ethidium bromide-stained RNA could be visualised on the membrane using a uv transilluminator. RNA was fixed to the nitrocellulose by baking at 80°C for 2 h.

2.5.3 Transfer of bacterial DNA to nitrocellulose membranes

To screen bacterial clones by hybridisation [section 2.6.2], transformants were either transferred individually or replica plated onto asymmetrically marked Hybond-C gridded nitrocellulose membranes (Amersham) on L plates, containing appropriate antibiotics, and regrown at 37°C for 2–4 h. The colonies were lysed by placing the membranes, colony side up, sequentially on 3MM Whatman paper soaked with a) 10% SDS (w/v) b) 0.5M NaOH and c) 1M Tris-HCl pH7.4 for 3 min each. The filters

were subsequently washed in 2x SSC for 10 min and the DNA fixed to the membrane as in section 2.5.1.

2.5.4 Transfer of bacteriophage M13 DNA to nitrocellulose membranes

To screen plates of M13 plaques by hybridisation [section 2.6.2], Hybond-C gridded nitrocellulose membranes (Amersham) were placed on the plates for 1 min and marked asymmetrically. The phage were lysed by placing the membranes, phage side up, on 3MM Whatman paper soaked with 0.5M NaOH, 1.5M NaCl for 3 min and neutralised with 0.5M Tris-HCl pH7.4, 1.5M NaCl twice for 3 min each. The filters were washed in 6x SSC for 10 min and the DNA fixed to the membrane as in section 2.5.1.

2.6 HYBRIDISATION TECHNIQUES

2.6.1 Radiolabelling of double-stranded DNA probes (Feinberg and Vogelstein 1984)

Restriction fragments isolated from LGT agarose [section 2.4.3] and fragments isolated from polyacrylamide gels [section 2.4.6] were labelled using the random hexanucleotide priming method. 20–50ng of DNA was labelled overnight at room temperature or for 1 h at 37°C using 30μCi [α-^{32}P]dCTP with the Multiprime Labelling System (Amersham). Probes in LGT agarose were melted at 70°C and thoroughly vortexed before labelling.

Unincorporated nucleotides were removed by desalting the probe in a 2ml Sephadex G-100 column using 200mM NaCl, 1mM EDTA, 10mM Tris-HCl pH7.4

2.6.2 Hybridisation of radiolabelled DNA

Nitrocellulose filters were prewetted in prewash buffer, inserted into Hybaid bottles and prewashed for 30 min at 42°C in a rotating Hybaid hybridisation oven. Filters were then prehybridised at 42°C for 2–14 h in hybridisation buffer containing 100μg/ml sonicated salmon sperm DNA, which had been boiled for 5 min and cooled to ~42°C immediately before use. Probes were boiled in hybridisation buffer (at 10^5cpm/ml for cloned DNA or 10^6cpm/ml for genomic DNA or RNA blots) and allowed to cool to ~42°C immediately before use. Filters were hybridised for 16–24 h

at 42°C.

DNA probes containing repetitive sequences were boiled for 5 min in hybridisation buffer containing 100μg/ml sonicated salmon sperm DNA and 125μg/ml sonicated human placental DNA (gift from S J Cross) and then incubated at 42°C for 2 h to allow reassociation of repetitive sequences prior to hybridisation with the filters (Sealey *et al.* 1985).

2.6.3 Washing of nitrocellulose membranes and autoradiography

Following hybridisation, the filters were washed twice in 2x SSC/0.1% (w/v) SDS for 15 min at 42°C in Hybaid bottles followed by two high stringency washes for 30 min each at 65°C in 0.2x SSC/0.1% (w/v) SDS. Filters were air dried, wrapped in Saran wrap and exposed to sensitised Kodak X-OMAT X-ray film between two tungstate intensifying screens at -70°C (Laskey and Mills 1977). Films were developed using a Kodak ME-1 automatic processor. To facilitate orientation of the developed films, asymmetrically placed Glogo phosphorescent markers (Stratagene) were exposed with filters.

2.7 SUBCLONING METHODS

2.7.1 Preparation of plasmid vectors

5–10μg of plasmid DNA, purified from a CsCl gradient [section 2.3.3], was digested to completion using 20U of enzyme and incubating for 2 h [section 2.4.1]. Digested DNA was end-filled [section 2.4.4], if necessary, and then dephosphorylated using 5U calf intestinal phosphatase, incubated at 37°C for 30 min, in the buffer provided. The reaction was stopped by heating at 70°C for 10 min.

The blunt-ended vectors used were; pBluescript KS^+ *Eco*RV cut, pBC SK^- *Eco*RV cut, pBC SK^- *Hin*cII cut, pGEM 3Zf(+) *Hin*cII cut, pGEX-2T *Sma*I cut (supplied by W Thomson), M13mp10 *Sma*I cut (Amersham) and pSPL1 *Bam*HI cut and filled-in.

Vectors were tested for a) complete digestion and b) efficient dephosphorylation, by transformation into competent cells [section 2.7.5]. All vectors contained <1% DNA undigested or not dephosphorylated.

2.7.2 Preparation of insert DNA

Following restriction enzyme digestion [section 2.4.1], specific fragments of plasmid or cosmid DNA were end-filled [section 2.4.4] and isolated from polyacrylamide gels [section 2.4.6]. Fragments generated by PCR were purified away from dNTPs, PCR primers and short DNA fragments, following digestion and end-filling, using the Magic PCR preps purification system (Promega) as described in Promega Technical Bulletin No.118 using disposable syringes. Some PCR products were ligated directly after kinase treatment with 500µM rATP in 1x kinase buffer (10mM $MgCl_2$, 7mM Tris-HCl pH7.4, 5mM DTT, 1.5mM spermidine) using 5U polynucleotide kinase in 20µl reaction volume for 30 min at 37°C, followed by 10 min at 80°C. To subclone multiple DNA fragments from cosmid or plasmid inserts, the constructs were either cut within the antibiotic resistance gene, in addition to the required restriction enzyme digestion, or the fragments were subcloned into a vector containing the gene giving resistance to a different antibiotic.

2.7.3 Ligation reactions

Ligations were carried out using 10ng of vector and 1–20ng of insert DNA in 1x ligase buffer (50mM Tris pH7.4, 20mM DTT, 10mM $MgCl_2$, 0.5mM spermidine, 2.5mg/ml BSA) containing 1mM rATP and 2.5U T4 DNA ligase in a total volume of 10µl. The reactions were incubated at room temperature for ~5 h.

2.7.4 Religation reactions

In order to delete specific fragments from plasmid or cosmid inserts, religation reactions were carried out. Approximately 100ng of construct DNA was digested with appropriate restriction enzymes to excise part of the insert. Following end-filling [section 2.4.4], if necessary, ~0.5ng of the vector fragment was religated as in section 2.7.3 for ~2 h.

2.7.5 Transformation procedures

Frozen TG1 or NM554 competent cells [section 2.2.4] were thawed on ice and 50µl added to each ligation reaction. Samples were incubated on ice for 10 min and then heat-shocked at 37°C for 5 min. After addition of 150µl of L broth, the samples were incubated at 37°C for ~30 min, spread onto L plates containing appropriate antibiotic

and incubated at 37°C overnight.

To allow colour selection [section 2.7.7.1] of pBluescript, pBC and pGEM recombinant clones to be performed, TG1 competent cells carrying the ΔM15 *lac* gene on the F′ episome, were used and 100µl 0.1M IPTG and 40µl 4% (w/v in dimethylformamide) BCIG were spread onto the plates immediately before use. Cells containing the *lac Z* gene disrupted by the subcloned insert grew as white colonies. Recombinant pGEX clones were transformed into NM554 competent cells.

2.7.6 Transfection procedures

200µl of TG1 competent cells [section 2.2.3] were added to each M13 ligation reaction and incubated on ice for 40 min. The cells were heat-shocked at 42°C for 5 min and transferred to 10ml Sterilin tubes. A mixture containing 3ml of melted L broth top agar (at 50°C), 25µl 0.1M IPTG, 25µl 2% (w/v in dimethylformamide) BCIG and 500µl of a TG1 daytime culture (1:1,000 dilution of an overnight culture in 2x TY broth grown at 37°C for ~6 h) were added to the cells, poured onto prewarmed 2x TY plates and incubated at 37°C overnight. Bacteriophage containing DNA inserts that disrupt the *lac Z* gene [section 2.7.5] grew as white plaques.

2.7.7 Identification of recombinant clones

2.7.7.1 *Colour selection*

Recombinant clones in *E.coli* carrying the ΔM15 *lac* gene on the F′ episome could be colour selected by complementation with plasmids carrying the *lac Z* gene. Wild-type clones contain the functional α peptide of β-galactosidase, expression of which is induced by IPTG, allowing hydrolysis of the colourless substrate, BCIG, to give a blue precipitate which is visible on the plate. Subcloning of an insert into a polylinker site, within the coding sequence of the α peptide, can either alter the translation reading frame or disrupt the protein structure to produce inactive β-galactosidase and produce white clones on IPTG/BCIG plates.

Care must be taken because subcloning of small inserts that do not disrupt the protein structure can produce false 'negative' results and substitution or deletion of 1 or 2 bases from the vector cloning site can produce mutants that cannot hydrolyse BCIG and appear as false 'positive' clones.

2.7.7.2 *PCR amplification of inserts*

Using oligonucleotide primers flanking the cloning site of the vector, PCR can be used to amplify cloned inserts. 5µl from a single bacterial colony diluted in 100µl L broth was used in a standard 50µl PCR reaction [section 2.3.8]. 25 cycles were carried out and the PCR products were analysed by electrophoresis on agarose minigels [section 2.4.2].

2.7.7.3 *Hybridisation using a radiolabelled probe*

Bacterial colonies or M13 plaques were transferred to nitrocellulose membranes [sections 2.5.3 and 2.5.4] and screened using radiolabelled probes as described in sections 2.6.2 and 2.6.3.

2.7.7.4 *Restriction enzyme digestion of plasmid DNA*

To determine the size and orientation of inserts subcloned into plasmid vectors, individual colonies were picked into L broth and plasmid DNA was isolated as described in section 2.3.2.1. The plasmid DNA was digested with restriction enzymes and the restriction pattern was analysed by agarose gel electrophoresis on minigels [section 2.4.2].

2.8 ISOLATION OF CLONES FROM cDNA LIBRARIES

cDNA libraries, constructed from mRNA derived from U937 (monocyte) tissue cultured cells and foetal brain tissue, in the vector CDM8, were provided by D Simmons (Seed 1987, Simmons and Seed 1988) and transformed into MC1061/p3 competent cells. Approximately $5x10^5$ clones were screened with genomic DNA probes to identify cDNAs encoded by genes in those regions.

2.8.1 Screening and isolation of positive clones

DNA, from the library plated on duplicate replica filters, was transferred onto nitrocellulose [section 2.5.3] and hybridised under competition conditions [section 2.6.2]. Autoradiographs of washed filters were aligned with the library filters to identify duplicating positive clones.

Frozen pairs of master filters, stored as a sandwich at -70°C, were thawed at room temperature, separated, air dried on 3MM Whatman paper and transferred to fresh

agar plates containing the relevant antibiotics. The bacteria were grown for ~2 h at 37°C and the master filter copies were aligned with the autoradiographs on a light box. Bacteria within the area containing the positive clone was scraped from the filter into 500µl L broth using a sterile pipette tip. The scraped bacteria were allowed to recover for 2 h at 37°C. These master filter copies were transferred to HMF-1 plates and stored for up to 1 month at 4°C.

The original master filters were transferred to HMF-2 plates, recovered for ~3 h at 37°C and then new copy filters, dampened on L plates, were compressed against the master filters and marked with alignment dots before refreezing at -70°C.

2.8.2 Rescreening of positive clones

Dilutions (1:1,000 and 1:10,000) were made from the stocks of scraped bacteria and 100µl of each, plated onto nitrocellulose filters on LAT plates, was grown overnight at 37°C. Replica filters were generated and rescreened [section 2.7.7.3] with the original probe. Single, duplicating positive colonies were identified and grown as overnight cultures to prepare glycerol stocks and miniprep DNA [section 2.3.2]. cDNA insert fragments were excised from the vector by restriction enzyme digestion and analysed by electrophoresis on agarose minigels.

To rapidly screen large numbers of potential positive clones, overnight cultures of original mixed positive clones were grown up and DNA prepared. Inserts were excised, separated from the vector DNA on agarose minigels, transferred onto nitrocellulose membranes and probed with the original probe. This allows identification of positive clones and their insert sizes so that fewer cDNA clones need to be rescreened as above, to obtain single colonies.

2.9 DNA SEQUENCING

2.9.1 Preparation of single-stranded DNA from bacteriophage M13 (Sanger *et al.* 1980)

Each M13 plaque, or 10µl of phage stock, was picked into a 10ml Sterilin tube containing 1.5ml of a 1:100 dilution of TG1 overnight cultured cells in 2x TY broth and incubated at 37°C, with vigorous shaking, for ~5 h. The culture was centrifuged twice, for 5 min at 12,000rpm in a microfuge, and at each stage the supernatant was

transferred to a fresh 1.5ml microfuge tube. 100µl was retained after the second spin and stored at 4°C as phage stock.

200µl 20% (w/v) PEG 4000 in 2.5M NaCl was added to the supernatant, thoroughly mixed and incubated for 20 min at room temperature. The precipitated phage were pelleted by centrifugation at 12,000rpm for 10 min and resuspended in 100µl TE pH7.4. The aqueous samples were extracted using phenol and the single-stranded DNA recovered by ethanol precipitation. The washed and dried DNA pellet was resuspended in 25µl TE pH7.4.

Cloned inserts were characterised by sequencing only the 'T' lane [section 2.9.3] *i.e.* T-tracking.

2.9.2 Preparation of single-stranded DNA from phagemid clones (Short *et al.* 1988)

Single-stranded DNA can be prepared from the pBluescript, pBC and pGEM phagemids using a helper phage. 1.5ml of 2x TY broth, containing 10^7pfu/ml helper phage M13K07 (Promega), was inoculated with 10µl of phagemid-transformed TG1 glycerol stock in a 10ml Sterilin tube. Following incubation at 37°C, with shaking, for 1 h, kanamycin was added to 75µg/ml and the cultures incubated for a further 12–14 h. Single-stranded DNA was recovered from the cultures as described in section 2.9.1.

2.9.3 Dideoxy sequencing of single-stranded DNA templates

Sequencing reactions were performed on single-stranded DNA templates using the USB Corporation, Sequenase Version 2.0, DNA sequencing kit according to the manufacturers instructions (Tabor and Richardson 1987). The M13 primer supplied with the kit, and specific oligonucleotide primers [section 2.3.6], were used to prime extension, catalysed by Sequenase enzyme, incorporating [α-^{35}S]-thio dATP, using ~0.5µg of single-stranded DNA as the template, in the presence of 2',3'-dideoxy-nucleoside triphosphate chain-terminating inhibitors.

2.9.4 Dideoxy sequencing of double-stranded DNA templates

Double-stranded DNA, prepared using the Magic Minipreps DNA Purification System (Promega) [section 2.3.2.2], was used as template in sequencing reactions performed using the USB Corporation, Sequenase Version 2.0, DNA sequencing kit.

Approximately 5μg DNA was denatured by incubating in 0.2M NaOH, 0.2mM EDTA for 30 min at 37°C (Lim and Pene 1988), neutralised using 0.1 volume of 3M sodium acetate pH4.8 and precipitated using ethanol. This method allows sequence to be determined from both ends of DNA inserts in double-stranded plasmid vectors and from large inserts which are unstable in single-stranded bacteriophage.

2.9.5 Cycle sequencing (Murray 1989)

Cycle sequencing was carried out using the dsDNA Cycle Sequencing System (Gibco BRL) according to the manufacturers instructions. This method permits direct sequencing of double-stranded DNA templates using successive rounds of heat denaturation, annealing of [γ-^{32}P]ATP end-labelled primer and extension with dideoxynucleoside chain terminators. The use of repeated cycles allows <50fmol template to be used, and extension using *Taq* DNA polymerase at elevated temperature eliminated compressions that could not be resolved using reagents in the Sequenase kit.

2.9.6 Denaturing polyacrylamide gel electrophoresis

Sequencing reaction samples were electrophoresed on 0.4mm thick 6% (w/v) denaturing acrylamide gels in 1x TBE buffer at 30mA. Samples were loaded in a 0.4mm doublefine (3mm) sharkstooth comb (Gibco BRL). 0.3M sodium acetate was included in the lower reservoir of gels electrophoresed for 3 h to compress the shorter DNA fragments at the bottom of the gel. The buffer was changed after 2.5–3 h during long electrophoresis runs of 5 h.

The back gel plate was pretreated with 2.5ml ethanol containing 75μl of 10% (v/v) acetic acid and 2.5μl Bindsilane (LKB). The front plates were treated with dimethyldichlorosilane (2% v/v in trichloroethane). Following electrophoresis, gels were fixed in 10% (v/v) methanol 10% (v/v) acetic acid and dried onto the back plate by baking at 110°C. Direct autoradiography was carried out at room temperature for 1–7 days without intensifying screens [section 2.6.3].

2.9.7 Analysis of sequence data

Sequence data was analysed using the Oxford University Molecular Biology Software and Databases on the VAX/VMS cluster. Specific DNA sequences were entered into

individual files and data from shotgun analyses were entered into databases using the 1991 Staden Package program SAP (Sequence Assembly Program). Sequences were analysed using the Staden program NIP (Nucleotide Interpretation Program) and the Genetics Computer Group (GCG) Sequence Analysis Software Package Version 7.1 March 1992 (Devereux *et al.* 1984). To search for similarity with other DNA sequences, the GENEMBL database was screened using the GCG package FASTA program (Pearson and Lipman 1988).

2.10 DETERMINATION OF TRANSCRIPTION START SITES

2.10.1 Primer extension analysis (Ausubel *et al.* 1989, McKnight and Kingsbury 1982)
Primer extension analysis was used to determine the 5′ end of a mRNA.

2.10.1.1 *Preparation of radiolabelled primer*

Oligonucleotide primers [section 2.3.6] were reprecipitated with ethanol to remove all traces of ammonium ions which inhibit T4 polynucleotide kinase. 100ng of primer was end-labelled using 200μCi [γ-^{32}P]ATP in 30μl 1x kinase buffer [section 2.7.2] with 5U T4 polynucleotide kinase by incubating at 37°C for 1 h. The DNA was ethanol precipitated with 2μg of tRNA as a carrier, resuspended in formamide loading dyes and heated at 85°C for 5 min before loading on a 10% (w/v) denaturing acrylamide gel. Following electrophoresis at 30mA, the gel was exposed for autoradiography [section 2.6.3], at room temperature for 1 min, to visualise the position of the labelled oligonucleotide for excision. The DNA was eluted from the gel slice by incubating, overnight at 37°C, in 400μl 5M ammonium acetate and precipitated using 1ml ethanol in a bath of dry ice and industrial methylated spirits for 30 min. The oligonucleotide was pelleted at 12,000rpm for 30 min, at 4°C, in a microfuge and resuspended in 50μl 0.3M sodium acetate pH7.

2.10.1.2 *Hybridisation and reverse transcription*

Radiolabelled primer (10^6cpm) [section 2.10.1.1] was coprecipitated with 40μg total RNA from the U937 cell line, or with 40μg tRNA as a control. Each pellet was resuspended in 30μl R-loop hybridisation buffer (80% v/v formamide, 40mM PIPES pH6.4, 400mM NaCl, 1mM EDTA), heated to 75°C for 5 min and then cooled to 40°C over a period of three hours, enabling the primer to anneal to the RNA template.

The DNA/RNA hybrids were precipitated using ethanol.

The hybrid pellet was resuspended in 25µl of reverse transcriptase mix (50mM Tris-HCl pH8.2, 50mM KCl, 6mM $MgCl_2$, 0.5mM dNTPs) containing 50U RNasin (Promega) and 40U AMV reverse transcriptase (Northumbria Biologicals Ltd) and incubated at 42°C for 1 h, then 50°C for 30 min. EDTA at 20mM and 40µg/ml RNase A was added at 37°C for 30 min to stop the reaction. 100µl 2.5M ammonium acetate was then added, the sample was extracted with phenol/chloroform and ethanol precipitated with 20µg tRNA as a carrier.

The DNA pellet was resuspended in 5µl formamide loading dyes and heated at 85°C before loading on a 6% (w/v) denaturing acrylamide gel together with M13mp18 sequencing reactions from the Sequenase kit. The gel was electrophoresed, fixed and dried [section 2.9.6], then exposed for autoradiography at -70°C [section 2.6.3] for 1 week.

2.10.2 RNase protection analysis (Ausubel *et al.* 1989, Sambrook *et al.* 1989)

Single-stranded, antisense RNA probes were used to study 5′ mRNA and gene structure. Following annealing to mRNA, the riboprobe was digested with RNase to remove regions of the probe that do not form hybrids with the corresponding mRNA.

2.10.2.1 *Preparation of radiolabelled riboprobe*

Genomic DNA inserts containing 5′ exons and upstream sequence were subcloned into pGEM-3Zf(+) and the constructs were linearised at the 5′ end of the insert by restriction enzyme digestion. Linearised template was treated with 100µg/ml proteinase K in 50mM NaCl, 5mM EDTA, 10mM Tris-HCl, 0.5% (w/v) SDS for 1 h at 37°C to remove RNase, then phenol/chloroform extracted and ethanol precipitated. The DNA pellet was resuspended in RNase-free water.

Transcription reactions were set up using 1µg template DNA with the Riboprobe Gemini System II (Promega) using 20U T7 RNA polymerase (Promega) and 50µCi [α-^{32}P]CTP (Amersham) in a 20µl reaction volume. The reactions were incubated for 1 h at 40°C and then 1U RQ1 DNase (Promega) was added at 37°C for 15 min to digest the DNA template. The antisense RNA probes were diluted to 50µl, phenol/chloroform extracted and precipitated using 200µl 2.5M ammonium acetate and 750µl ethanol with 20µg tRNA as a carrier. The RNA pellets were resuspended

in 10µl formamide loading dyes, heated at 85°C and loaded on a 6% (w/v) denaturing acrylamide gel. Following electrophoresis at 30mA, the gel was exposed for autoradiography [section 2.6.3] for 1 min at room temperature to locate the fragment. A gel slice including the riboprobe was excised from the gel. The RNA was eluted by shaking at 37°C in 400µl RNA elution buffer (0.5M ammonium acetate, 10mM magnesium acetate, 1mM EDTA, 0.1% (w/v) SDS) for 4 h. The eluted riboprobe was precipitated with 1ml ethanol and resuspended in 20µl R-loop hybridisation buffer [section 2.10.1.2].

2.10.2.2 *Hybridisation and RNase digestion*

20µg of total RNA from U937 cells, and 20µg tRNA as a control, from ethanol precipitates were resuspended in 30µl of R-loop hybridisation buffer containing $5x10^5$cpm of riboprobe, heated at 85°C for 5 min and then hybridised at 45°C for ~8 h. RNase digestion was carried out by adding 350µl RNase digestion buffer (300mM NaCl, 10mM Tris-HCl pH7.5, 5mM EDTA) containing 40µg/ml DNase-free RNase A and 2µg/ml DNase-free RNase T1 (Grade V, Sigma) and incubating at 30°C for 45 min. The reactions were stopped by adding 20µl 10% (w/v) SDS and 10µl 10mg/ml proteinase K in water and incubating at 37°C for 15 min. The samples were phenol/chloroform extracted and the RNA was precipitated using 1ml ethanol with 10mg tRNA carrier.

The RNA pellets and 10^4cpm riboprobe control were resuspended in 5µl formamide loading dyes and heated at 85°C before loading on a 6% denaturing acrylamide gel, together with M13mp18 sequencing reactions from the Sequenase kit. The gel was electrophoresed, fixed [section 2.9.6] and then exposed for autoradiography at -70°C [section 2.6.3] for 1 week.

2.11 EXON AMPLIFICATION

This method utilises the splice sites that flank exons in primary RNA transcripts to isolate coding sequences from genomic DNA. Fragments of genomic DNA up to 4.5kb in length were subcloned into the *Bam*HI cloning site of the pSPL1 vector (Buckler *et al.* 1991).

2.11.1 Cell culture and electroporation

COS-7 cells were grown in DMEM (Flow Labs) supplemented with 10% (v/v) heat-inactivated FCS, 100U/ml penicillin, 100µg/ml streptomycin and 50µg/ml kanamycin to ~80% confluence, in 5% CO_2, at 37°C. The cells were harvested by trypsin digestion and washed twice with PBS (Flow Labs) before resuspending in 1x HEPES buffered saline containing 1% (w/v) PEG 8000 (275mM NaCl, 10mM KCl, 2mM Na_2HPO_4, 0.2% (w/v) dextrose, 1% (w/v) HEPES pH7.05 with NaOH, 1% (w/v) PEG 8000) at $1.25x10^6$ cells/ml. 20µg of plasmid construct DNA, purified by CsCl gradient centrifugation [section 2.3.3], was pipetted onto the electrode of a 4mm Gene Pulser Cuvette (BioRad). 800µl cell suspension was added to the cuvette and the cells were incubated for 10 min on ice. The cells were electroporated, using a BioRad Gene Pulser, with 250V (field strength=625V/cm) and the Capacitance Extender to obtain a capacitance of 950µFD. The time constants for the electroporations were 26–28ms. Following electroporation, the cells were incubated on ice for 5 min and then grown in fresh medium for 48 h.

2.11.2 RNA isolation and RT-PCR

Approximately 25% of the cells survived electroporation [section 2.11.1] and were subsequently harvested and washed [section 2.11.1]. To isolate the RNA the cell pellet was solubilised in 100µl RNazol (Biogenesis). 10µl of chloroform was added, mixed with the homogenate and incubated for 5 min on ice. The suspension was centrifuged at 12,000rpm in a microfuge, at 4°C for 15 min, the aqueous layer was removed and an equal volume of isopropanol was added to precipitate the RNA. The pellet was washed with 75% (v/v) ethanol and the total amount of RNA present estimated, as described in section 2.3.9. ~5µg total RNA was reverse transcribed using the cDNA Synthesis System Plus Kit (Amersham) with oligo dT as the primer [section 2.3.7].

Approximately 0.5µg reverse transcribed RNA was used for a first round of PCR [section 2.3.8] using 5µM primers SD2 and SA2 (Buckler *et al.* 1991), from the rabbit β globin sequence, for 35 cycles in a final volume of 100µl. 2µl of the first round PCR product was used as template for a second, more specific, round of 20 PCR cycles using primers SD1 and SA1 (Buckler *et al.* 1991), from the HIV *tat* sequence. The

amplified products of the second round of PCR were then digested with *Bam*HI and *Sal*I and subcloned for sequence analysis [section 2.9].

2.12 GENERAL PROTEIN TECHNIQUES

2.12.1 Computer analysis of protein sequences

Predicted protein sequences were obtained from translation of cDNA ORFs. These were analysed using the Staden programs PIP (Protein Interpretation Program) and SIP (Similarity Investigation Program), from the 1991 Staden Package, and the GCG 1992 Package (Devereux *et al.* 1984). To determine sequence similarity with other proteins the Genetics Computer Group (GCG) Sequence Analysis Software Package program FASTA (Pearson and Lipman 1988) was used to screen the National Biomedical Research Foundation (NBRF) and SwissProt databases. Initial protein alignments were carried out using the CLUSTAL V program (Higgins and Sharp 1988) and subsequent alignments were carried out using the AMPS package (Barton 1990). The AMPS multiple sequence alignment was converted into a Postscript file for annotation and laser printed output using the ALSCRIPT (Barton 1993) program.

2.12.2 Analysis of proteins by SDS-PAGE

Proteins were separated on 10% (w/v) and 15% (w/v) SDS-PAGE gels on PROTEAN II *xi* or Mini-PROTEAN II (BioRad) vertical electrophoresis cells. Samples were boiled for 5 min in 1x sample buffer before loading on gels. Electrophoresis was carried out at 100 or 200V in 1x running buffer (Laemmli 1970) and prestained molecular weight markers (BioRad) were loaded on gels to allow estimation of protein sizes. Proteins were stained with Coomassie blue stain for 15–30 min at room temperature and destained at 55°C overnight.

2.12.3 Transfer of proteins onto nitrocellulose membranes (Western blotting)

Proteins fractionated on SDS-PAGE gels [section 2.12.2] were transferred by Western blotting onto Hybond-C extra nitrocellulose membranes (Amersham). Gels were soaked in Western blot transfer buffer for 5 min, placed on top of 3 layers of Whatman 3MM paper, covered first with the nitrocellulose membrane and then a further 3 layers of Whatman 3MM paper, prewetted in transfer buffer. This

'sandwich' was placed between the electrodes of a LKB 2117 Multiphor electrophoresis unit. The proteins were 'dry blotted' by electrophoresis at 0.8mA/cm^2 for 2 h at room temperature. Following transfer of the proteins, the membranes were gently washed in PBS, wrapped in Saran wrap and stored overnight at 4°C or at –20°C.

In order to prepare total cultured cell lysates for Western blot transfer ~10^7 cells were washed three times with PBS (Flow Labs), resuspended in 900µl water and incubated on ice for 20 min. The cells were lysed by sonication (2x30 seconds), adjusted to 0.1M $NaHCO_3$ and stored at -20°C. Approximately 1.5mg protein was boiled with an equal volume of 2x protein sample buffer and loaded into a single large well. Gels were electrophoresed as in section 2.12.2 and protein transferred as above.

To check the efficiency of transfer, the gel was stained with Coomassie blue [section 2.12.2] after blotting.

2.12.4 Determination of protein concentration

The concentration of purified proteins bound to agarose beads was estimated by comparing the intensity of Coomassie blue stained proteins on SDS-PAGE gels [section 2.12.2] with a series of dilutions of a standard solution of protein of a similar molecular weight.

For protein mixtures in solution the total concentration was determined by measuring the OD_{280} of the sample, assuming 1mg/ml corresponds 1 absorbance unit. With purified antibodies, it was assumed that 1 absorbance unit corresponds to 0.75mg/ml, at 280nm.

2.12.5 Expression of polypeptides in *E. coli* (Smith and Johnson 1988)

cDNA sequences encoding polypeptides to be expressed were subcloned into the pGEX-2T vector and transformed into NM554 competent cells [section 2.7]. This vector contains the *Schistosoma japonicum* glutathione S-transferase gene (Sj26), modified at its 3′ end to encode a thrombin-cleavable linker and cloning sites to allow in-frame insertion of the cDNA sequences.

To detect induction of fusion proteins, *E.coli* transformed with the pGEX-2T recombinants were grown in 10ml of L broth, containing 50µg/ml ampicillin,

overnight at 37°C [section 2.2.2]. 1:10 dilutions of the overnight cultures were incubated at 37°C for 1 h and then IPTG was added to 100μM to induce expression of the fusion protein. The cultures were grown for a further 3-4 h and the cells were pelleted at 625g for 5 min. The pellets were resuspended in 10μl of water and 20μl 2x sample buffer was added. The samples were boiled for 5 min, briefly centrifuged and 5μl of supernatant loaded on 10% (w/v) SDS-PAGE minigels [section 2.12.2]. Induction of fusion protein expression could be detected by comparison of induced samples with controls not induced with IPTG.

2.12.6 Isolation of fusion proteins from *E.coli*

2.12.6.1 *Small scale preparations*

To determine the efficiency of fusion protein purification from *E.coli* cell lysates 10ml cultures were induced with 100μM IPTG [section 2.12.5] and the cells pelleted after ~4 h. The cells were resuspended in 250μl MTPBS containing 1% (v/v) Triton X-100 and lysed by mild sonication (10–20 seconds) on ice. Cell debris was removed by centrifuging twice at 156g for 1 min and then one fifth volume of 50% (w/v in MTPBS) pre-swollen glutathione agarose beads (Sigma) was added. The samples were mixed on a rotator for 10 min, to allow absorption of the fusion protein to the beads via the Sj26 region of the protein. The beads were recovered by centrifuging at 156g for 5 min. Unbound proteins were removed by washing the beads three times each with 1ml MTPBS. The beads were diluted to 50% suspension in MTPBS and samples were analysed by SDS-PAGE [section 2.12.2] to estimate the efficiency of fusion protein isolation.

2.12.6.2 *Large scale preparations*

For large scale preparations of fusion proteins, a 20ml culture in L broth containing 50μg/ml ampicillin [section 2.2.2] was grown overnight at 37°C and used to inoculate 800ml of L broth/ampicillin. The large cultures were grown for 2 h at 37°C before 100μM IPTG was added, and the cells then grown for a further 4 h. The cells were pelleted by centrifugation for 10 min at 4224g and 4°C, in a Sorvall GS-3 rotor (RC-5b centrifuge DuPont instruments), resuspended in 10ml MTPBS/1% (v/v) Triton X-100 and lysed by sonication on ice. Cell debris was removed by centrifuging twice at 625g for 5 min each. 2ml of 50% (w/v in MTPBS) pre-swollen glutathione agarose beads were added to the supernatant and mixed on a rotator for 10 min. The beads

were pelleted at 156g for 5 min and washed three times, each with 10ml MTPBS, before being diluted to 50% suspension in MTPBS. 5µl beads were boiled with 5µl 2x sample buffer and analysed by SDS-PAGE [section 2.12.2].

2.12.7 Direct thrombin cleavage of purified fusion protein

To separate the expressed polypeptide away from the Sj26 protein the fusion protein was cleaved with thrombin whilst attached to the glutathione agarose beads (Cheng 1991).

Purified fusion protein absorbed onto beads [section 2.12.6] was resuspended in an equal volume of 0.3M NaCl, 5mM $CaCl_2$. Human thrombin (Sigma) was added at 1U per 50µl beads and digestion was carried out at 25°C for 1–2 h. Release of the expressed polypeptide was determined by analysing the supernatant and beads from the digest separately by SDS-PAGE. Soluble peptides can be released from the Sj26 protein into the supernatant (Cheng 1991).

2.12.8 Elution of fusion proteins from glutathione agarose beads

The addition of reduced glutathione can be used to elute the intact fusion protein from the glutathione agarose beads (Smith and Johnson 1988).

Purified fusion protein, absorbed onto beads, was washed twice for 10 min at room temperature with two bead volumes of elution buffer (50mM Tris-HCl pH8, 5mM reduced glutathione) containing 0.1% (v/v) Triton X-100, 0.1% (v/v) Tween 20, 0.003% (w/v) SDS and 1mM DTT to solubilise the eluted protein. The washes were concentrated, if necessary, by spinning through a Centricon-30 column (Amicon) at 12,000g. The reduced glutathione was removed by diluting the concentrate 1:100 with MTPBS containing 0.1% (v/v) Triton X-100, 0.1% (v/v) Tween 20, 0.003% (w/v) SDS, 1mM DTT, reconcentrating in the Centricon-30 column and then repeating the process. The concentration of the recovered fusion protein was estimated as in section 2.12.4.

2.13 PREPARATION AND PURIFICATION OF POLYCLONAL ANTIBODIES

2.13.1 Immunisation of rabbits with antigen (Harlow and Lane 1988)

Whole fusion protein, prepared as described in section 2.12.8, was used as antigen for the immunisation of rabbits.

Prior to immunisation with antigen, a 10ml prebleed was taken from each rabbit. Two rabbits were each immunised by multiple intradermal injections containing 100μg of antigen as a 50% (v/v) emulsion with Freund's complete adjuvant (Difco). 10ml blood samples were taken from the rabbits 7–14 days after each immunisation.

2.13.2 Preparation and storage of rabbit serum

10ml blood samples were allowed to clot at room temperature for ~2 h in 50ml polypropylene tubes (Falcon). The clot was detached from the sides of the tube with a sealed glass Pasteur pipette and allowed to contract at 4°C overnight. The serum was removed and cleared of insoluble material by centrifugation at 1,000g for 10 min at 4°C. Aliquots of serum were stored at -20°C.

2.13.3 Purification of antibodies by affinity chromatography

Anti-G7b antibodies were purified by affinity chromatography. Two columns were prepared by covalently binding total *E.coli* cell lysates to CNBr-activated Sepharose 4B (Pharmacia) columns. The first column contained G9a-Sj26 fusion protein (G9a is an unrelated novel gene in the MHC class III region) and was used to remove antibodies against Sj26. The second column contained G7b-Sj26 fusion protein and was used to purify anti-G7b antibodies.

2.13.3.1 *Preparation of affinity chromatography columns*

E.coli cell lysates containing G7b and G9a fusion proteins were prepared from 200ml cultures [section 2.12.6.2]. 250–300mg total protein from the cell lysates was dialysed against the coupling buffer (50mM NaH_2PO_4, 100mM NaCl pH8) at 4°C overnight to remove low molecular weight amides.

Each cell lysate was made up to 10ml with coupling buffer and added to 2g CNBr-activated Sepharose, pretreated according to the manufacturers instructions. The coupling reaction was carried out for 2 h at room temperature on a rotator and then the Sepharose was washed successively with 2 volumes 1M NaCl, 2 volumes of

water and then 2 volumes of coupling buffer. Any remaining reactive groups were blocked by incubating with ethanolamine pH8.5 at room temperature for 2 h after which 5mM EDTA pH7.5 was added as a preservative.

To estimate the amount of protein bound, the amount of free protein present after the coupling reaction and in the washing solutions was measured [section 2.12.4]. The G7b-Sj26 column contained ~10mg protein/ml resin and the G9a-Sj26 column contained ~13mg protein/ml resin.

2.13.3.2 *Affinity chromatography*

Both affinity columns were prepared at 4°C with 3.5ml bed volumes, washed with 1M NaCl and equilibrated with PBS/1mM EDTA pH7.4. 2.5ml anti-G7b antiserum was diluted to 10ml with PBS/1mM EDTA pH7.4 and passed through the G9a-Sj26 column 4 times using a peristaltic pump with a flow rate of ~20ml/h. Using measurements before and after passage through the column, the proportion of protein bound to the column was determined. The serum was then passed through the G7b-Sj26 column, twice, using a peristaltic pump with a flow rate of ~10ml/h and the amount of protein bound was again determined.

The column was washed with ~10 volumes of PBS/1mM EDTA pH7.4 and ~5 volumes 0.4M NaCl to remove weakly bound material. The antibody was eluted with 2 volumes 3M $MgCl_2$. 1ml fractions of 0.4M NaCl and 3M $MgCl_2$ washes were collected and OD_{280} measured to produce an elution profile. The column was washed with 10ml 0.4M NaCl before equilibrating with PBS/1mM EDTA pH7.4 for storage.

The fractions containing the purified antibodies, eluted with 3M $MgCl_2$, were pooled and dialysed against water for 3 h and then PBS/1mM EDTA pH7.4 overnight at 4°C.

2.14 CHARACTERISATION OF ANTIBODIES

2.14.1 ELISA assays (Harlow and Lane 1988)

ELISA assays were used to determine the titre of antisera and purified antibodies. Microtitre wells were incubated with 2µg of purified antigen [section 2.12.8] in 100µl 15mM $NaHCO_3$, 35mM Na_2CO_3 pH9.6 at room temperature overnight. The wells

were washed 3 times with 200µl PBS and then serial dilutions (10^{-2}–10^{-5}) of antisera, or purified antibodies, and pre-immune serum as a control, were added to the wells in 100µl PBS and incubated at room temperature for 2 h. The free antibodies were washed out using four washes with 200µl PBS, 0.05% (v/v) Tween 20. To enable detection of the bound antibodies, 100µl of alkaline phosphatase conjugated goat anti-rabbit IgG (Sigma) diluted 1:5,000 in PBS, 0.05% (v/v) Tween 20 was added to each well and incubated at room temperature for 2 h. The wells were then washed four times with 200µl PBS, 0.05% (v/v) Tween 20 before 100µl of substrate solution (1% (w/v) p-Nitrophenyl phosphate, 5mM $MgCl_2$, 5mM $CaCl_2$, 50mM Tris pH9.6) was added to each well.

The yellow colour was allowed to develop at room temperature and the OD_{405} was measured at 10 min intervals in a Titertek Multiskan PLUS MkII plate reader for up to 1 h. The titre of antisera was determined by comparison of the OD_{405} of the antisera with that of the pre-immune serum.

2.14.2 Detection of antigens on Western blots (Harlow and Lane 1988)

Purified and thrombin-cleaved fusion proteins [section 2.12.7] and total cell lysates were transferred onto nitrocellulose by Western blotting as described in section 2.12.3. Blots of total cell lysates were cut into strips for use with different serum samples.

The nitrocellulose membranes were blocked with PBS, 0.1% (v/v) Tween 20 overnight at 4°C or for 1 h shaking at room temperature and then incubated with a 1:200 dilution of the antiserum, in PBS, containing 0.1% (v/v) Tween 20 for two hours at room temperature. The membranes were then washed three times, each for 10 min, in PBS. To detect bound antibodies, the blots were incubated for 2 h with a 1:2,000 dilution, of alkaline phosphatase conjugated goat anti-rabbit IgG (Sigma) in PBS, 0.1% (v/v) Tween 20. After washing in PBS three times for 10 min each, the blots were developed using freshly mixed alkaline phosphatase substrate solution (70mM Tris-acetate pH8.6, 0.05% (w/v) Nitrotetrazolium blue (NBT) chloride, 0.0025% (w/v) BCIP, 4mM $MgCl_2$) and incubated at room temperature for 1–5 h before development was stopped by transferring the blots to PBS.

With some antisera, blots were blocked with 5% (w/v) fat-free milk solution in PBS, 0.1% (v/v) Tween 20 solution to reduce non-specific binding.

CHAPTER 3

CHARACTERISATION OF G7

3.1 INTRODUCTION

The cloning and subsequent characterisation of the human MHC class III region (Sargent *et al.* 1989a; Spies *et al.* 1989a,b, 1990; Kendall *et al.* 1990) has shown that there are at least nineteen genes in the ~350kb interval between C2 and TNF-A (Figure 3.1). The initial approach adopted by Sargent *et al.* (1989a) to locate genes in this region was to identify HTF-islands, which are undermethylated, CpG-rich sequences that have been shown to be frequently associated with the 5′ ends of housekeeping, and some tissue-specific, genes (Bird 1986, 1987; Gardiner-Garden and Frommer 1987). HTF-islands are characterised by the presence of clustered sites for methylation-sensitive, rare-cutting restriction enzymes (*e.g. Bss*HII, *Eag*I, *Not*I, *Sac*II), which are not methylated in uncloned genomic DNA. Between C2 and TNF-A eleven potential HTF-islands were mapped (Sargent *et al.* 1989a) and genomic DNA probes derived from the sequences flanking these islands were isolated from cosmids covering this region. The use of these probes led to the identification of 14 novel genes; G1–G10, B144 and the HSP70 gene cluster (Sargent *et al.* 1989a; Milner and Campbell 1990).

One of these genes, G7, designated D6S56E (Spence *et al.* 1989), was identified using walking probe F, a 1.3kb, single copy, *Bgl*II fragment, located ~60kb telomeric of the HSP70 gene cluster (Figure 3.1; Sargent *et al.* 1989a). Probe F lies ~3kb centromeric of an HTF-island which is characterised by two *Bss*HII sites, 400bp apart. Northern blot analysis using probe F detected a mRNA transcript of ~3kb in total RNA isolated from cell lines representing monocyte (U937) and macrophage (U937 cells stimulated with Phorbol 12-myristate 13-acetate [PMA]), but not liver (HepG2), T lymphocyte (Molt4) or B lymphocyte (Raji) cell lineages (Sargent *et al.* 1989a).

The G7 gene has been sequenced at both the cDNA and genomic DNA levels in order to derive the sequence of the G7 protein product, and to search for similarities of this sequence with existing sequences in DNA and protein databases. This strategy

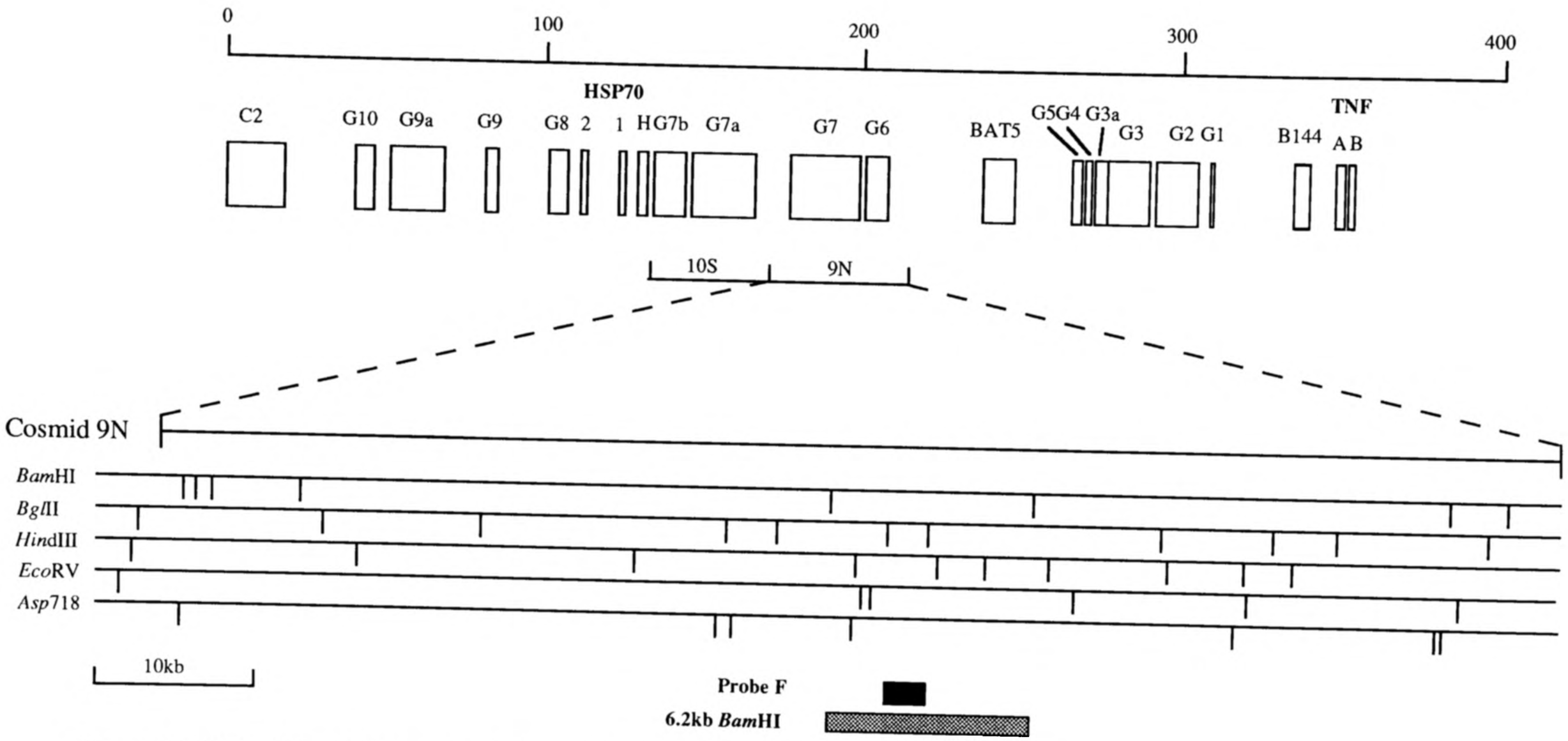


Figure 3.1 Molecular map illustrating the location of the G7 gene in the human MHC class III region.
The upper half of this figure shows the locations of the genes mapped between the C2 and TNF-B genes. The genes are shown as open boxes and the genomic DNA inserts of cosmid 9N and 10S are shown as a horizontal bar. The cosmid 9N genomic DNA insert is expanded in the lower half of the figure and the positions of restriction sites in this region are shown. The position of probe F is depicted by a filled box and the position of the 6.2kb *Bam*HI fragment is depicted by a shaded box.

has previously been successful in identifying the HSP70 genes (Sargent *et al.* 1989b; Milner and Campbell 1990) and the gene for valyl-tRNA synthetase (Hsieh and Campbell 1991) which are both located in the MHC class III region. In addition, regions showing similarity to other, known, proteins can be used to predict a possible function for the G7 protein or suggest a role in disease aetiology.

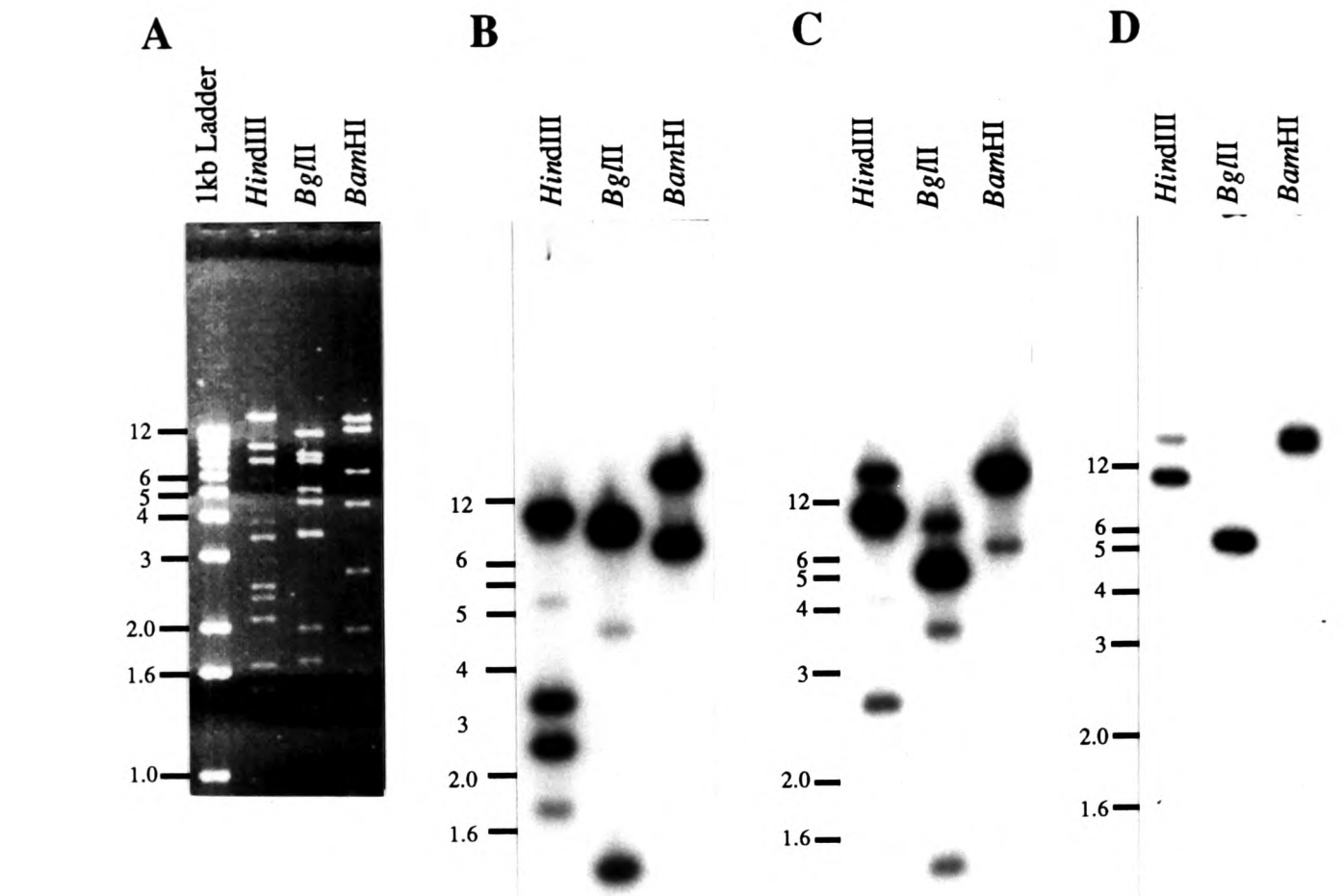
3.2 ISOLATION AND CHARACTERISATION OF G7 cDNA CLONES

3.2.1 Isolation of G7 cDNA clones

A 6.2kb *Bam*HI single copy fragment, which includes probe F (Figure 3.1), was isolated from cosmid 9N in order to screen for cDNA clones. This larger probe was isolated in order to detect partial cDNA clones that may not have been hybridised by probe F. Since G7 expression had only been detected in U937 cells (Sargent *et al.* 1989a), a cDNA library derived from that cell line was screened using the 6.2kb *Bam*HI fragment. The library (a gift from D Simmons, Institute of Molecular Medicine, Oxford) was constructed from RNA prepared from unstimulated U937 cells and cloned in the expression vector CDM8 (Seed 1987; Simmons and Seed 1988).

A total of $5x10^5$ colonies were screened using the 6.2kb *Bam*HI fragment from cosmid 9N, competed with 125µg/ml placental DNA [section 2.8.1], and two duplicating positive clones were identified. The cDNA inserts excised from these clones, G7-1 and G7-2, using *Xba*I, were 1.8kb and 2.2kb, respectively.

To confirm that these clones contained sequences encoded in cosmid 9N, the cDNA inserts were isolated and used to probe Southern blots of cosmids 9N and 10S, as a negative control, digested with *Hin*dIII, *Bgl*II and *Bam*HI. Clone G7-1 hybridised strongly to the 6.2kb *Bam*HI fragment, but it also hybridised weakly to multiple fragments derived from cosmid 10S, extending up to 60kb from probe F and lying centromeric of G7a, suggesting that the G7-1 clone contains some form of repetitive sequence element. Clone G7-2 hybridised to the region of cosmid 9N from which the original probe was derived, in addition to sequences up to 15kb away (Figure 3.2 C), indicating that G7 extends centromeric of probe F. As a result, the cDNA library was screened again using the G7-2 cDNA insert. A single, duplicating positive clone, G7–3, was identified with an insert size of 2.0kb.



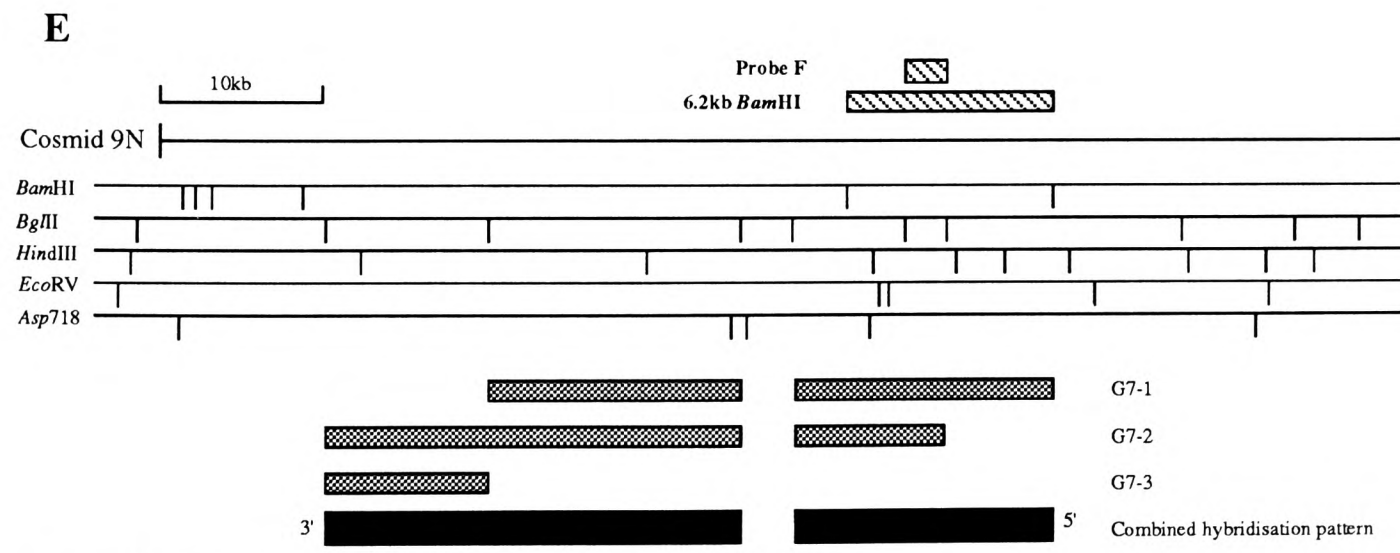


Figure 3.2 Mapping of the G7 cDNA clones within cosmid 9N.
A Cosmid 9N DNA digestedwith restriction enzymes and fractionated on a 0.8% (w/v) agarose gel. Southern blots of digested cosmid 9N DNA hybridised with the G7-1 (**B**), G7-2 (**C**), and G7-3 (**D**) cDNA clones. Sizes are in kb. **E** The extent of the cosmid 9N insert and the positions of restriction sites in this region are shown. The positions of probe F and the 6.2kb *Bam*HI fragment are depicted by hatched boxes and the regions of DNA hybridised by the G7 cDNA inserts are depicted by shaded boxes. The combined regions hybridised by the cDNA inserts are depicted by filled boxes.

The G7-3 cDNA insert was subsequently used to probe the Southern blot of cosmid 9N DNA (Figure 3.2 D). The blot was also reprobed with the G7-1 cDNA insert after absorption of repetitive sequence elements with placental DNA (Figure 3.2 B). From the resulting hybridisation patterns it was possible to map the three cDNA inserts to a region of ~20kb in cosmid 9N (Figure 3.2 E).

3.2.2 Characterisation of cDNA inserts

The three G7 cDNA inserts [section 3.2.1] were digested with restriction enzymes to produce the maps shown in figure 3.3 F. From these maps, clones G7-2 and G7-3 could be aligned, except for a single *Nco*I site, present in G7-2 and absent in G7-3. G7-1 contained no restriction sites that allowed it to be aligned with G7-2 or G7-3.

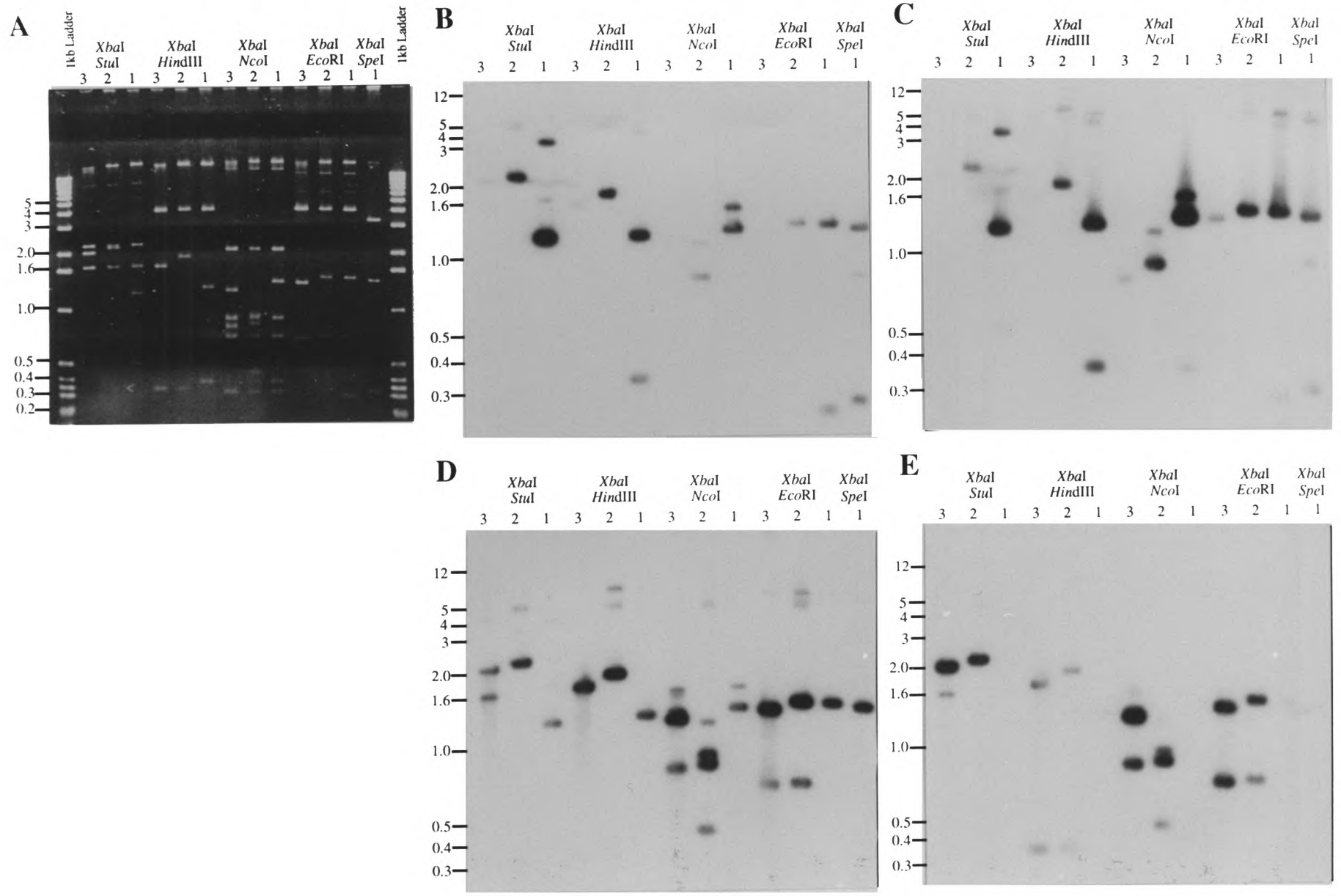
In order to determine the relationships between the three cDNA clones, restriction digests of the inserts were fractionated on agarose gels, transferred onto nitrocellulose membranes, by Southern blotting, and probed with each of the cDNA inserts and the original 6.2kb *Bam*HI DNA fragment, used to isolate the cDNA clones G7-1 and G7-2. The results (Figure 3.3) confirmed that G7-2 and G7-3 are related to one another and that G7-3 does not hybridise with G7-1 or the original probe. In addition, the results demonstrated that clones G7-1 and G7-2 only cross hybridise with one another over a limited region and some sequences in clones G7-1 and G7-3 are not hybridised by either of the other cDNA inserts or the genomic DNA probe.

3.2.3 Nucleotide sequence analysis of the cDNA clones

The three cDNA inserts were sequenced using the following strategies:

3.2.3.1 *G7-1*

This cDNA clone was sequenced using a directed cloning strategy. Specific *Hin*dIII and *Sty*I/*Spe*I fragments covering the region which hybridised to G7-2 and the genomic probe (Figure 3.3 E) were isolated and subcloned into *Sma*I cut M13mp10 [section 2.7]. Single-stranded DNA was prepared from the white plaques obtained and the inserts were compared by T-tracking [section 2.9.1]. Clones with different T-track patterns were subsequently sequenced using the M13 universal primer [section 2.9]. This sequence did not cover the whole of the 900bp *Sty*I/*Spe*I fragment so the isolated insert was digested with *Dde*I, end-filled and shotgun cloned into *Sma*I cut


A
1kb Ladder
XbaI StuI
XbaI HindIII
XbaI NcoI
XbaI EcoRI
XbaI SpeI
1kb Ladder
3 2 1 3 2 1 3 2 1 3 2 1 1
5
4
3
2.0
1.6
1.0
0.5
0.4
0.3
0.2
B
XbaI StuI
XbaI HindIII
XbaI NcoI
XbaI EcoRI
XbaI SpeI
3 2 1 3 2 1 3 2 1 3 2 1 1
12
5
4
3
2.0
1.6
1.0
0.5
0.4
0.3
C
XbaI StuI
XbaI HindIII
XbaI NcoI
XbaI EcoRI
XbaI SpeI
3 2 1 3 2 1 3 2 1 3 2 1 1
12
5
4
3
2.0
1.6
1.0
0.5
0.4
0.3
D
XbaI StuI
XbaI HindIII
XbaI NcoI
XbaI EcoRI
XbaI SpeI
3 2 1 3 2 1 3 2 1 3 2 1 1
12
5
4
3
2.0
1.6
1.0
0.5
0.4
0.3
E
XbaI StuI
XbaI HindIII
XbaI NcoI
XbaI EcoRI
XbaI SpeI
3 2 1 3 2 1 3 2 1 3 2 1 1
12
5
4
3
2.0
1.6
1.0
0.5
0.4
0.3

F

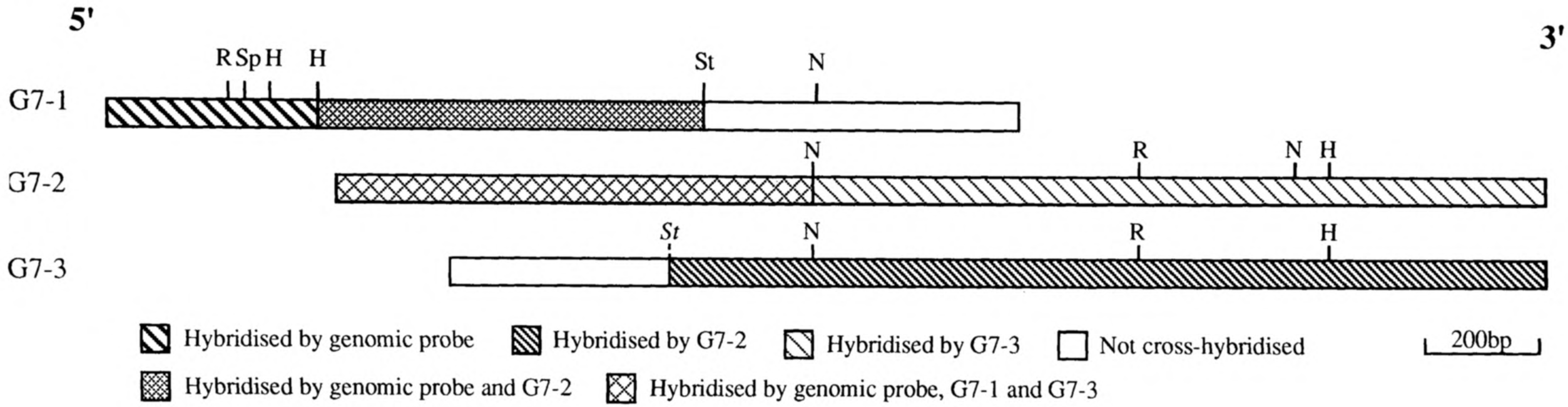


Figure 3.3 Cross-hybridisation of G7 cDNA clones.
(**A**) DNA from the G7 cDNA clones G7-1 (1), G7-2 (2) and G7-3 (3) was fractionated on a 0.8% (w/v) agarose gel. Southern blots of G7 cDNA inserts were probed with the 6.2 kb BamHI DNA fragment (**B**), G7-1 cDNA insert (**C**), G7-2 cDNA insert (**D**) and G7-3 cDNA insert (**E**). Sizes are in kb. (**F**) Pattern of cross hybridisation between the G7 cDNA clones. Restriction sites mapped within the cDNA inserts *Eco*RI (R), *Spe*I (Sp), *Hin*dIII (H), *Stu*I (St) and *Nco*I(N) are shown. The key for the hybridisation patterns is shown beneath the map.

M13mp10 [section 2.7]. The sequence obtained from these fragments completed the coverage of this region (Figure 3.4).

3.2.3.2 *G7-2*

To sequence this 2.2kb cDNA clone the whole insert was excised from the CDM8 vector using *Xho*I, isolated [section 2.4.6] and digested with each of the frequent cutting enzymes *Dde*I, *Hin*fI, *Hae*III and *Sty*I, end-filled if necessary, and shotgun cloned into *Sma*I cut M13mp10 [section 2.7]. The sequence data obtained was entered into the computer and assembled using the GCG program SAP [section 2.9.7]. A single 2,155bp contig (Figure 3.4) was generated which contained restriction sites that corresponded to the restriction map of G7-2 (Figure 3.3). The sequence was completed in both orientations by the isolation, subcloning into *Sma*I cut M13mp10 and sequencing of specific, *Dde*I/*Eco*RI and *Xba*I/*Eco*RI, fragments.

At one end of the sequence, a canonical polyadenylation signal AATAAA (Proudfoot and Brownlee 1976) was identified 10bp upstream from two deoxyadenosine residues at the end of the insert, thus defining the 3′ end of the cDNA. To confirm the sequence data, subclones spanning the insert were sequenced in both orientations using dITP mixes from the Sequenase sequencing kit. This resulted in an overall degeneracy of 6.9 for the 2,155bp insert.

3.2.3.3 *G7-3*

The similarity between the restriction maps and the cross hybridisation patterns of the inserts of clones G7-2 and G7-3 allowed a more directed approach to be used in the sequencing of G7-3. The cDNA insert was excised from CDM8 using *Xho*I, isolated [section 2.4.6] and subcloned, in both orientations, into *Hin*cII cut pGEM-3Zf [section 2.7]. Single-stranded DNA was recovered and used as template to sequence both ends of the insert with the M13 universal primer [section 2.9]. At the 3′ end the G7-3 cDNA sequence was identical to that of the G7-2 clone and included the polyadenylation signal. At the other end, the sequence did not correspond to any part of the sequences obtained from clones G7-1 or G7-2.

In order to sequence across the cDNA, a series of deletions (Figure 3.4) were prepared by religation of the clones containing the G7-3 insert in both orientations between the *Eco*RI, *Hind*III, *Sma*I or *Nco*I sites in the insert and *Eco*RI, *Hin*dIII and *Sma*I sites in the polylinker [section 2.7.4]. This allowed most of the cDNA insert to be sequenced using the M13 universal primer on single-stranded DNA templates

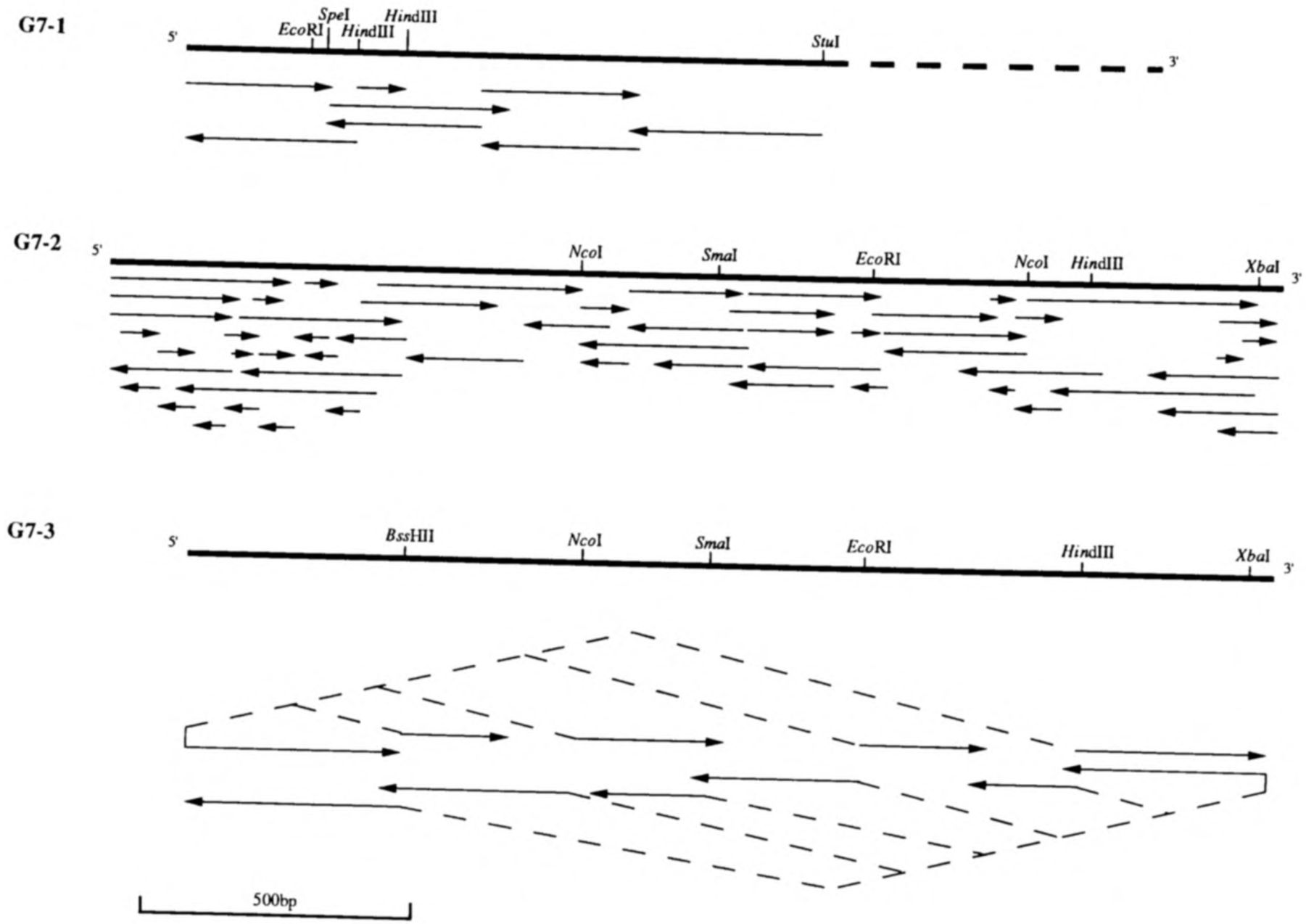


Figure 3.4 Sequencing strategies for the G7 cDNA clones.
The cDNA inserts are shown as horizontal bars with the positions of restriction sites indicated. The region of the G7-1 cDNA clone not hybridised by the other clones is shown as a dashed line. The arrows beneath the restriction maps indicate the orientation and extent of the sequence obtained from each subclone. Deletions generated by religation of the G7-3 clone are shown by dashed lines.

and the Sp6 promoter reverse primer on double-stranded DNA templates derived from the religated clones. By religating a *Bss*HII site, determined by sequencing, with the *Bam*HI site in the vector, in clones oriented in both directions, the sequence of the cDNA was completed using the M13 universal primer and single-stranded DNA templates (Figure 3.4) [section 2.9].

3.2.3.4 *Alignment of cDNA sequences*

Alignment of the sequence data obtained from the three cDNA inserts (Figure 3.5) demonstrated that clone G7-1 overlaps with the 5′ end of G7-2 over regions of 155bp, 35bp, 187bp and 63bp which are separated by a 4bp insertion containing a donor splice site in G7-1, an 83bp insertion in G7-2 and a 125bp insertion in G7-1 containing a donor splice site (Figure 3.5). The sequences then diverge at a donor splice site in G7–1 suggesting that the sequence at the 3′ end which is not hybridised by G7-2 is derived from an unspliced intron.

The G7-3 sequence overlaps that of G7-2 for 152bp and 1.4kb at the 3′ end, with the regions of overlap separated by a 73bp deletion in G7-3 (Figure 3.5). At the 5′ end, 450bp of sequence, unrelated to G7-1, G7-2 and the 6.2kb *Bam*HI genomic probe (Figure 3.3), containing part of an *Alu* repeat sequence (Figure 3.5; Schmid and Jelinek 1982), appears to be the product of a coligation event prior to cloning of the insert into CDM8. Within the overlapping region the sequences diverge for 48bp (Figures 3.5 & 3.6), which include the *Nco*I site in G7-2 [section 3.2.2]. There is also a single base substitution of a T (G7-2) to a G (G7-3) at position 1971 (Figure 3.5).

3.2.4 Confirmation of U937 sequence at position 1971

The sequence variation detected between cDNA clones G7-2 and G7-3 at position 1971 (Figure 3.5) could be either due to allelic variation at the G7 locus, a transcriptional error during mRNA synthesis or an error during reverse transcription in cDNA synthesis. The importance of this position in generating a stop codon in the G7-2 sequence [section 3.2.6] necessitated confirmation of the sequence at the cDNA level in U937 cells and at the genomic level in ICE5 DNA [section 3.3].

The region corresponding to nucleotides 1518-2421/2404 (Figure 3.5) was amplified by PCR using the 5′ sense primer G7OL3, $^{5'}$CTCTCAGAGGAGAAGCTGCAC$^{3'}$ (T_d=66°C), and the antisense 3′ primer G7OL4, $^{5'}$GCCACGAGCCACAAGCTTGTC$^{3'}$

```
G7-1 GGCGGTCGGTCAGCGGGGCG TTCTCCCACCTGTAGCGACT CAGAGCCTCCAAGCTCATGG CCTCCTTAGGAGCGAACCCA AGGAGGACACCGCAGGGACC GAGACCTGGGGCGGCCTCCT   120

G7-1 CCGGCTTCCCCAGCCCGGCC CCAGTGCCGGGCCCCAGGGA GGCCGAGGAGGAGGAAGTCG AGGAGGAGGAGGAGCTGGCC GAGATCCATCTGTGTGTGCT GTGGAATTCAGGATACTTGG   240

G7-1 GCATTGCCTACTATGATACT AGTGACTCCACTATCCACTT CATGCCAGATGCCCCAGACC ACGAGAGCCTCAAGCTTCTC CAGAGAGTTCTGGATGAGAT CAATCCCCAGTCTGTTGTTA   360

G7-1 CGAGTGCCAAACAGGATGAG AATATGACTCGATTTCTGGG AAAGCTTGCCTCCCAGGAGC ACAGAGAGCCTAAAAGACCT GAAATCATATTTTTGCCAAG TGTGGATTTTGGTCTGGAGA   480

G7-1 TAAGCAAACAACGCCTCCTT TCTGGAAACTACTCCTTCAT CCCAGACGCCATGACTGCCA CTGAGAAAATCCTCTTCCTC TCTTCCATTATTCCCTTTGA CTGCCTCCTCACACCCCCAG   600
G7-1                                                                                                                          G   600
G7-2

G7-1 GAGATTTAAGATTTACCCCG ATTCCACTGCTGATCCCCTC CCAGGTTCGAGCACTTGGAG GGCTGCTGAAGTTCCTGGGT CGAAGAAGAATCGGGGTTGA ACTGGAAGACTATAATGTCA   720
G7-2 GAGATTTAAGATTTACCCCG ATTCCACTGCTGATCCCCTC CCAGGTTCGAGCACTTGGAG GGCTGCTGAAGTTCCTGGGT CGAAGAAGAATCGGGGTTGA ACTGGAAGACTATAATGTCA   720

G7-1 GCGTCCCCATCCTGGGCTTT AAGAAATTTATGTTgtagGACTCA TCTGGTGAACATAGATCAAG ACACTTACAG---------- ---------- ---------- ----------   836
G7-2 GCGTCCCCATCCTGGGCTTT AAGAAATTTATGTT    GACTCA TCTGGTGAACATAGATCAAG ACACTTACAGTGTTCTACAG ATTTTTAAGAGTGAGTCTCA CCCCTCAGTGTACAAA   836
G7-3                                                                                                                  cttgacagtccgctggttaaagtg

G7-1 ---- ---------- ----------GAATCCT CAACAGATGCCACTGTAAGT GGGGAGAGAAGCTGCTCAGG CTATGGTTCACACGTCCGAC TCATGACCTGGGGGAG   956
G7-2 GTGG CCAGTGGACTGAAGGAGGGG CTCAGCCTCTTTGGAATCCT CAACAGATGCCACTGTAAGT GGGGAGAGAAGCTGCTCAGG CTATGGTTCACACGTCCGAC TCATGACCTGGGGGAG   956
G7-3 attctcctgcctcagcctcccgagtagctgaaattacaggcgccggccaccacgcctggctagcttttgtgtttttagtagagacggggtttcaccgtgttggccaggctggtcttgaactcctga
```

```
G7-1 CTCA GTTCTCGTCTGGACGTCATT CAGTTTTTTCTGCTGC CCCAGAATCTGGACATGGCTCAGA TGCTGCATCGGCTCCTGGGT CACATCAAGAACGTGCCTgtgcccagggtggagggca 1058
G7-2 CTCA GTTCTCGTCTGGACGTCATT CAGTTTTTTCTGCTGC CCCAGAATCTGGACATGGCTCAGA TGCTGCATCGGCTCCTGGGT CACATCAAGAACGTGCCT                   1058
G7-3 cctcaaatgacccacctctgcctcccaaagtgctgggattccaggcgtgagccaccgtgccaggacccaaggcccttaagtttaacgtctcattcttcagtcaggttttccttgttcctgcgtgtt

G7-1 gggaggtggggaaggaggttgagggctgatactgggcagtgggcttcttgaggggcattagagtgagggaagagaaaacagcggctgtaaccttgtctgactgtagCT GATTCTGAAACGCATGA 1077
G7-1                                                                                                          CT GATTCTGAAACGCATGA 1077
G7-2
G7-3 cagccatttgtttttaagtttgtgttgaaggagaaactaacaacgaaaatggacttgttgacggaagaaaagtaggaatgcagcctctggtgctgtttgagtgatccctctgccccaggcctgctg

G7-1 AGT TGTCCCACACCAAGGTCAGC GACTGGCAGGTTCTCTACAA Ggtaaggcct
G7-2 AGT TGTCCCACACCAAGGTCAGC GACTGGCAGGTTCTCTACAA GACTGTGTACAGTGCCCTGG GCCTGAGGGATGCCTGCCGC TCCCTGCCGCAGTCCATCCA GCTCTTTCGGGACATTG 1197
G7-3 cgcgctgctgtgttctggaaaggcgcattgtgccctcgctgtggcagACTGTGTACAGTGCCCTGG GCCTGAGGGATGCCTGCCGC TCCCTGCCGCAGTCCATCCA GCTCTTTCGGGACATTG 1197

G7-2 CCC AAGAGTTCTCTGATGACCTG CACCATATCGCCAGCCTCAT TGGGAAAGTAGTGGACTTTG AGGGCAGCCTTGCTGAAAAT CGCTTCACAGTCCTCCCCAA CATAGATCCTGAAATTG 1317
G7-3 CCC AAGAGTTCTCTGATGACCTG CACCATATCGCCAGCCTCAT TGGGAAAGTA---------- ---------- ---------- ---------- 1317

G7-2 ATG AGAAAAAGCGAAGACTGATG GGACTTCCCAGTTTCCTTAC TGAGGTTGCCCGCAAGGAGC TGGAGAATCTGGACTCCCGT ATTCCTTCATGCAGTGTCAT CTACATCCCTCTGATTG 1437
G7-3 --- ---AAAAGCGAAGACTGATG GGACTTCCCAGTTTCCTTAC TGAGGTTGCCCGCAAGGAGC TGGAGAATCTGGACTCCCGT ATTCCTTCATGCAGTGTCAT CTACATCCCTCTGATTG 1437

G7-2 GCT TCCTTCTTTCTATTCCCCGC CTGCCTTCCATGGTAGAGGC CAGTGACTTTGAGATTAATG GACTGGACTTCATGTTTCTC TCAGAGGAGAAGCTGCACTA TCGTAGTGCCCGAACCA 1557
G7-3 GCT TCCTTCTTTCTATTCCCCGC CTGCCTTCCATGGTAGAGGC CAGTGACTTTGAGATTAATG GACTGGACTTCATGTTTCTC TCAGAGGAGAAGCTGCACTA TCGTAGTGCCCGAACCA 1557
```

```
G7-2 AGG AGCTGGATGCATTGCTGGGG GACCTGCACTGCGAGATCCG GGACCAGGAGACGCTGCTGA TGTACCAGCTACAGTGCCAG GTGCTGGCACGAGCAGCTGT CTTAACCCGAGTATTGG 1677
G7-3 AGG AGCTGGATGCATTGCTGGGG GACCTGCACTGCGAGATCCG GGACCAGGAGACGCTGCTGA TGTACCAGCTACAGTGCCAG GTGCTGGCACGAGCAGCTGT CTTAACCCGAGTATTGG 1677

G7-2 ACC TTGCCTCCCGCCTGGACGTC CTGCTGGCTCTTGCCAGTGC TGCCCGGGACTATGGCTACT CAAGGCCGCGTTACTCCCCA CAAGTCCTTGGGGTACGAAT CCAGAATGGCAGACATC 1797
G7-3 ACC TTGCCTCCCGCCTGGACGTC CTGCTGGCTCTTGCCAGTGC TGCCCGGGACTATGGCTACT CAAGGCCGCGTTACTCCCCA CAAGTCCTTGGGGTACGAAT CCAGAATGGCAGACATC 1797

G7-2 CTC TGATGGAACTCTGTGCCCGA ACCTTTGTGCCCAACTCCAC AGAATGTGGTGGGGACAAAG GGAGGGTCAAAGTCATCACT GGACCCAACTCATCAGGGAA GAGCATATACCTCAAAC 1917
G7-3 CTC TGATGGAACTCTGTGCCCGA ACCTTTGTGCCCAACTCCAC AGAATGTGGTGGGGACAAAG GGAGGGTCAAAGTCATCACT GGACCCAACTCATCAGGGAA GAGCATATACCTCAAAC 1917

G7-2 AGG TAGGCTTGATCACATTCATG GCCCTGGTAGGCAGCTTTGT GCCAGCAGAG**T**AGGCCGAAA TTGGGGCAGTAGACGCCATC TTCACACGAATTCATAGCTG CGAATCCATCTCCCTTG 2037
G7-3 AGG TAGGCTTGATCACATTCATG GCCCTGGTAGGCAGCTTTGT GCCAGCAGAG**G**AGGCCGAAA TTGGGGCAGTAGACGCCATC TTCACACGAATTCATAGCTG CGAATCCATCTCCCTTG 2037

G7-2 GCC TCTCCACCTTCATGATCGAC CTCAACCAGGTGGCGAAAGC AGTGAACAATGCCACTGCAC AGTCGCTGGTCCTTATTGAT GAATTTGGAAAGGGAACCAA CACGGTGGATGGGCTCG 2157
G7-3 GCC TCTCCACCTTCATGATCGAC CTCAACCAGGTGGCGAAAGC AGTGAACAATGCCACTGCAC AGTCGCTGGTCCTTATTGAT GAATTTGGAAAGGGAACCAA CACGGTGGATGGGCTCG 2157

G7-2 CGC TTCTGGCCGCTGTGCTCCGA CACTGGCTGGCACGTGGACC CACATGCCCCCACAGTCTTT GTGGCCACCAACTTTCTGAG CCTTGTTCAGCTACAACTGC TGCCACAAGGGCCCCTG 2277
G7-3 CGC TTCTGGCCGCTGTGCTCCGA CACTGGCTGGCACGTGGACC CACATGCCCCCACAGTCTTT GTGGCCACCAACTTTCTGAG CCTTGTTCAGCTACAACTGC TGCCACAAGGGCCCCTG 2277

G7-2 GTG CAGTATTTG---------- ---------- ACCATGGAGACCTGTGAGGA TGGCAACGATCTTGTCTTCT TCTATCAGGTT TGCGAAGGTGTTGCGAAGGC CAGCC 2365
G7-3 GTG CAGTATTTGGATTGGCCAAG GGTTTCAGGACAGGAAGGAG ---------- ---------- --------GTT TGCGAAGGTGTTGCGAAGGC CAGCC 2348

G7-2 ATGCCTCCCACACAG CTGCCCAGGCTGGGCTTCCT GACAAGCTTGTGGCTCGTGG CAAGGAGGTCTCAGACTTGA TCCGCAGTGGAAAACCCATC AAGCCTGTCAAGGATTTGCT AAAGA 2485
G7-3 ATGCCTCCCACACAG CTGCCCAGGCTGGGCTTCCT GACAAGCTTGTGGCTCGTGG CAAGGAGGTCTCAGACTTGA TCCGCAGTGGAAAACCCATC AAGCCTGTCAAGGATTTGCT AAAGA 2468

G7-2 AGAACCAAATGGAAA ATTGCCAGACATTAGTGGAT AAGTTTATGAAACTGGATTT GGAAGATCCTAACCTGGACT TGAACGTTTTCATGAGCCAG GAAGTGCTGCCTGCTGCCAC CAGCA 2605
G7-3 AGAACCAAATGGAAA ATTGCCAGACATTAGTGGAT AAGTTTATGAAACTGGATTT GGAAGATCCTAACCTGGACT TGAACGTTTTCATGAGCCAG GAAGTGCTGCCTGCTGCCAC CAGCA 2588

G7-2 TCCTCTGAGAGTCCT TCCAGTGTCCTCCCCAGCCT CCTGAGACTCCGGTGGGCTG CCATGCCCTCTTTGTTTCCT TATCTCCCTCAGACGCAGAG TTTTTAGTTTCTCTAGAAAT TTTGT 2725
G7-3 TCCTCTGAGAGTCCT TCCAGTGTCCTCCCCAGCCT CCTGAGACTCCGGTGGGCTG CCATGCCCTCTTTGTTTCCT TATCTCCCTCAGACGCAGAG TTTTTAGTTTCTCTAGAAAT TTTGT 2708

G7-2 TTCATATTAGGAATA AAGTTTATTTTGAA 2754
G7-3 TTCATATTAGGAATA AAGTTTATTTTGAA 2737
```

Figure 3.5 Alignment of G7 cDNA insert nucleotide sequences.

The G7 cDNA insert sequences are aligned over regions where they overlap. Exonic sequences [section 3.3.3] are shown in upper case and non-contiguous or intronic insertions are shown in lower case. Deleted exonic sequences are denoted by dashes. The sequence variations at position 1971 and invariant **g t** residues at 5' donor splice sites are shown in bold. The partial *Alu* repeat element in the non-contiguous sequence at the 5' end of the G7-3 insert is underlined. The polyadenylation signal AATAAA at the 3' end is double underlined. Bases are numbered from the 5' end of the G7-1 cDNA insert and do not include intronic insertions or non-contiguous sequences.

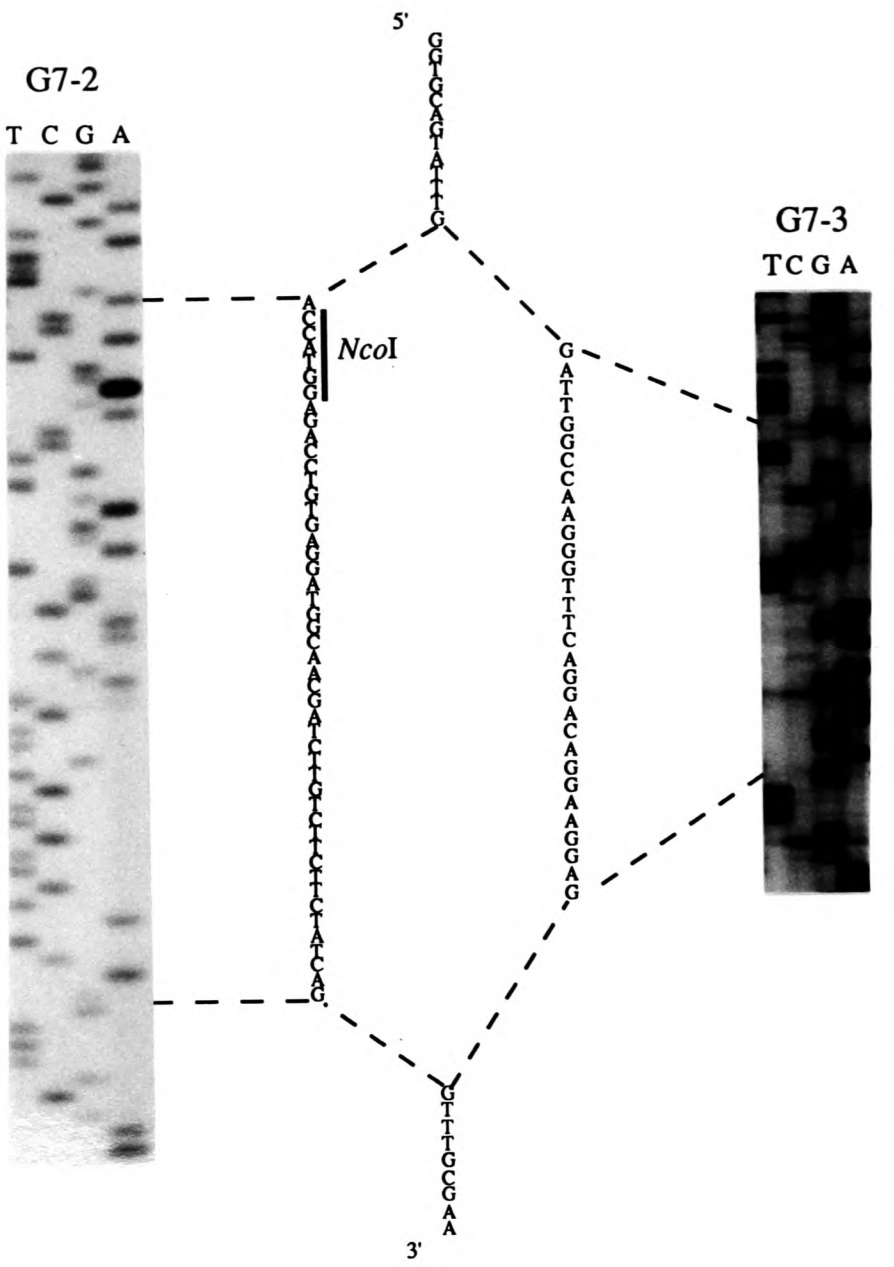


Figure 3.6 Alternative sequences in the G7 cDNA inserts.
The nucleotide sequences of cDNA clones G7-2 and G7-3 diverge for 48bp and one sequence contains the *Nco*I site as shown.

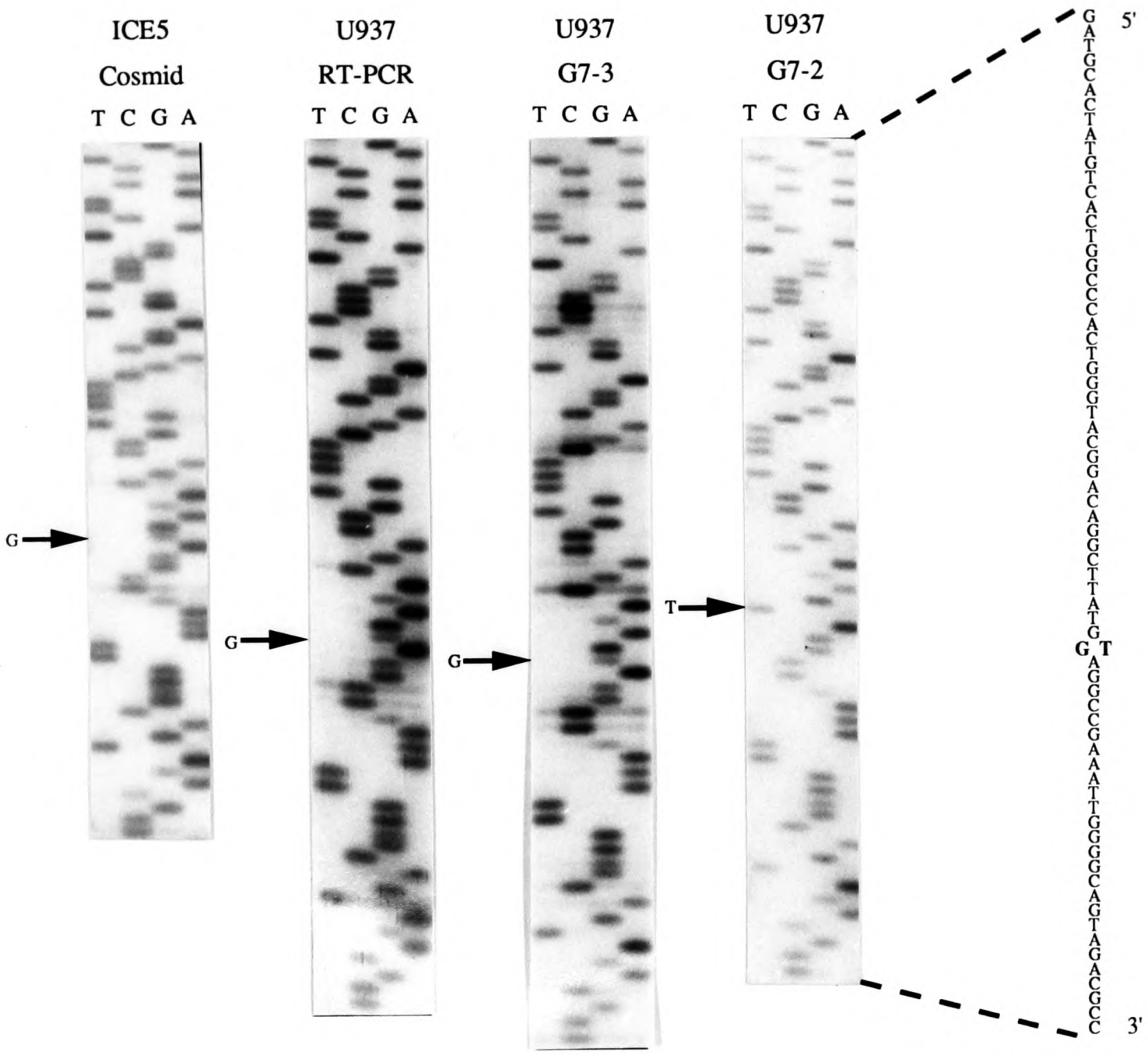


Figure 3.7 G7 sequences at cDNA nucleotide position 1971.
Genomic DNA sequence data from cosmid 9N and cDNA sequences from U937 amplified by RT-PCR, insert G7-3 and insert G7-2 are shown. The arrows indicate the base at positions corresponding to nucleotide 1971 in the cDNA sequence.

(T_d=68°C). 200ng of U937 total RNA was reverse transcribed using oligo dT primer [section 2.3.7] and PCR amplification [section 2.3.8] was carried out for 30 cycles to obtain an ~900bp product (Figure 3.21). A 285bp *SmaI/Eco*RI fragment from the PCR product was subcloned into *Sma*I cut M13mp10 [section 2.7] and sequenced [section 2.9] in both orientations. The sequence data confirmed that in ICE5 genomic DNA, U937 RT-PCR sequence and in clone G7-3, the nucleotide at position 1971 is G. Only clone G7-2 has T at position 1971 (Figure 3.7).

3.2.5 Mapping the 5′ end of G7 mRNA by primer extension

An oligonucleotide, G7OL5, of sequence $^{5'}$CTCTGAGTCGCTACAGGTGG$^{3'}$ (T_d=64°C), complementary to nucleotides 25–45 from the G7-1 insert (Figure 3.5), was end-labelled with [γ-^{32}P]-ATP [section 2.7.2], hybridised at 40°C with total RNA isolated from U937 cells and extended using reverse transcriptase, as described in section 2.10.1. Analysis of the extension products by denaturing polyacrylamide gel electrophoresis consistently revealed two products of 220bp and 343bp, both of which were not detected in tRNA controls (Figure 3.8). This extends the length of the longer cDNA by up to 299bp at the 5′ end to 3053bp.

3.2.6 Determination of the G7 cDNA structure

The G7 cDNA sequences obtained in section 3.2.3 (Figure 3.5) span 2737 and 2754bp. and primer extension analysis [section 3.2.5] indicates that these sequences are extended by 176 or 299bp at the 5′ end, thereby totalling 2913, 2930, 3036 or 3053bp. This data corresponds well with the mRNA of ~3kb detected by Northern blot analysis. These cDNA sequences contain numerous variations *i.e.* deletions and insertions, and in order to deduce the putative ORF encoding the G7 gene product, all three cDNA sequences were translated in three phases [section 2.9.7] using the polyadenylation signal sequence in clones G7-2 and G7-3 to define the orientation of the clones (Figure 3.9).

In G7-1 the longest ORF starts with the first methionine codon at nucleotide 57, which is preceded by an in-frame stop codon at nucleotide 33, and is terminated by a stop codon at nucleotide 756. This stop codon lies in the 4bp insertion in G7-1, which is absent from G7-2 (Figure 3.5). In G7-2, the ORF of the 155bp overlap between G7-1 and G7-2 continues in G7-2 to the stop codon at nucleotide 1971 from the start of

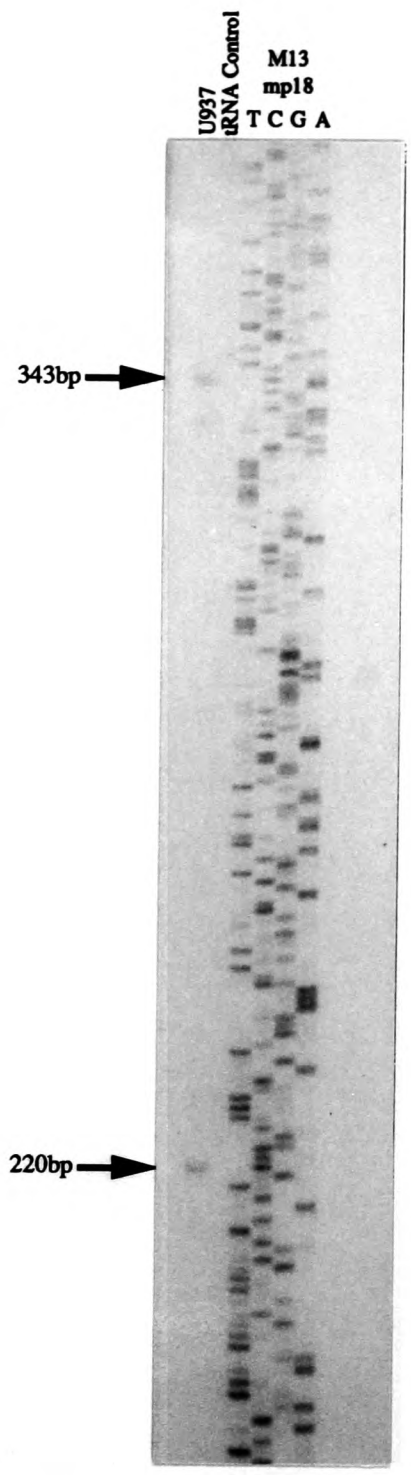


Figure 3.8 Primer extension analysis of the 5' end of G7 mRNA.
Reverse transcription was primed from oligonucleotide G7OL5 hybridised at 40°C with 40µg of total RNA isolated from U937 cells and 40µg of tRNA. Following reverse transcription samples were electrophoresed on a 6% (w/v) acrylamide/ 7M urea denaturing gel with an M13mp18 sequence ladder as size markers. Specific products of 220bp and 343bp in the track containing products derived from RNA isolated from U937 cells are indicated. These correspond to cDNA extensions of 176bp and 299bp.

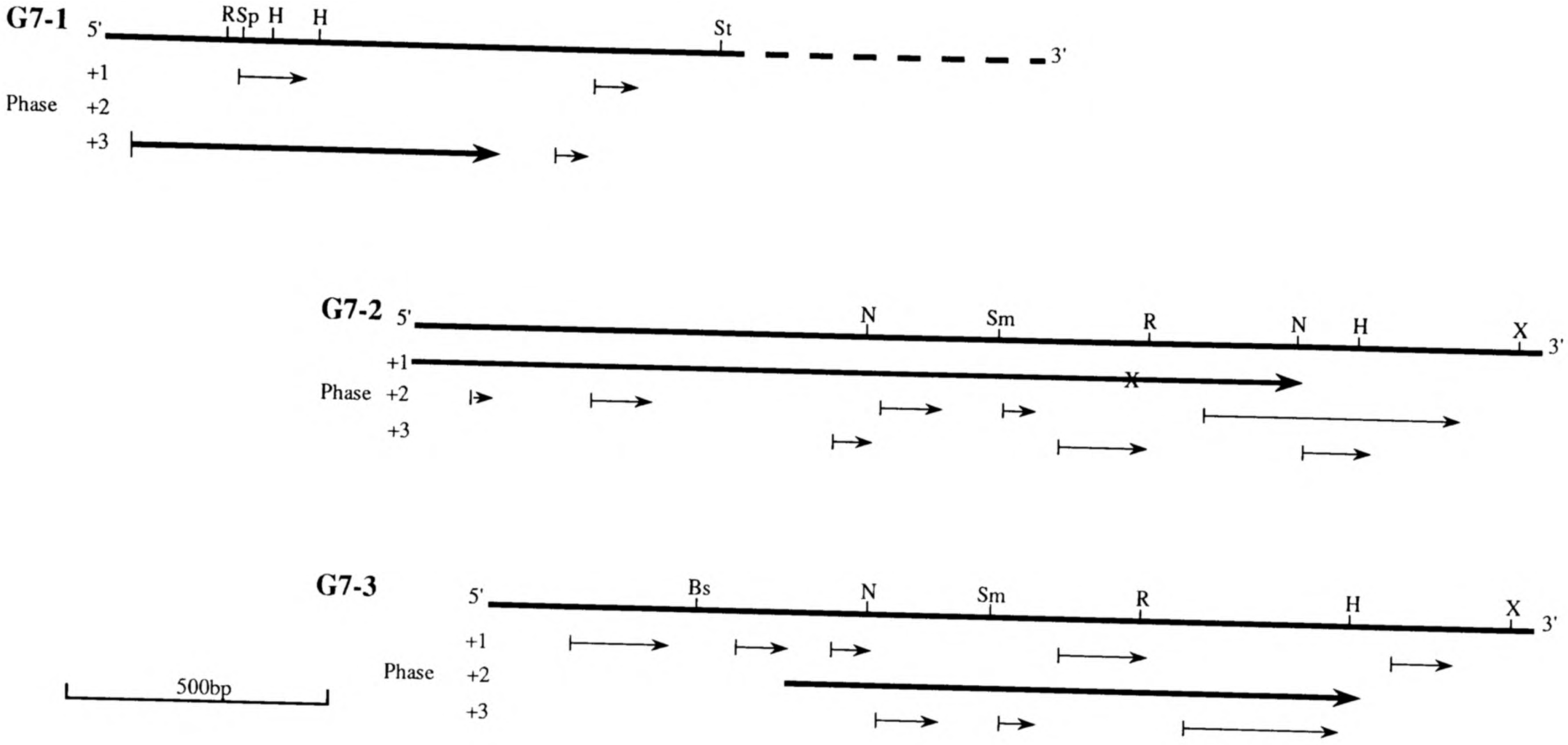


Figure 3.9 Open reading frames in the G7 cDNA inserts.
The G7 cDNA inserts are shown as horizontal lines and are aligned according to their overlapping sequence [section 3.2.3] with restriction sites *Eco*RI (R), *Spe*I (Sp), *Hin*dIII (H), *Stu*I (St), *Bss*HII (Bs), *Nco*I (N), *Sma*I (Sm) and *Xba*I (X) marked. ORFs >200bp are represented by arrows in the phase in which they occur. Methionine codons at the 5' ends are shown by vertical lines on the arrows. The longest continuous ORF that spans all three clones is marked by a bold arrow. The position of nucleotide 1971 in clone G7-2 that generates a stop codon is marked by an X on the bold arrow.

G7–1. The insertions and deletion in G7-1 continually disrupt the phase of translation so that multiple stop codons were encountered.

Where the G7-2 sequence overlaps G7-3 the ORF is again disrupted by a 22bp frameshift deletion in G7-3 that results in premature termination of the ORF. However, after this deletion, the overlapping sequence between G7-2 and G7-3 maintains the ORF until position 1971 where a stop codon is encoded in G7-2 [section 3.2.3]. However, incorporating the single base change in G7-3 allows the ORF of G7-2 to be extended to a stop codon at nucleotide 2304 resulting in a putative ORF of 2247bp (Figure 3.10). The further sequence divergence between G7-2 and G7-3 allows the ORF in G7-3 to be extended to position 2421 resulting in a longer putative ORF of 2364bp (Figure 3.10).

The longest potential ORF of G7 (Figure 3.10) starts with the first methionine codon which is located at position 57 of G7-1. The nucleotide sequence at this position is consistent with the translation start consensus $^{A}/_{G}$NCAUG of Cavener and Ray (1991) and Kozak (1991) with a C at position -1, but not the $^{A}/_{G}$ at -3 and is preceded by an in-frame stop codon at position 33. This ORF extends for 2364bp and encodes a putative polypeptide of 788 amino acids. In cDNA clone G7-2 another stop codon at position 2304 reduces the ORF to 2247bp encoding a 749 amino acid putative polypeptide. Therefore the G7 cDNAs consist of a 5′ untranslated region of 232 or 355bp, probable translated ORFs of 2247 or 2307bp and a 3′ untranslated region of 317 or 450bp containing a polyadenylation signal AATAAA.

3.2.7 Correction of a G7 cDNA clone

In order to generate a single cDNA clone from the inserts containing the sequences obtained in section 3.2.3 (Figure 3.5), clones G7-1 and G7-2 were combined. The insertions and deletions in G7-1 disrupt the ORF [section 3.2.6] but the 155bp overlap between clones G7-1 and G7-2 allows the ORF to be maintained from G7-1 to G7-2. This sequence contains no unique restriction site at which the two clones could be ligated together, however, the position at which G7-1 first diverges from G7-2 corresponds to a cleavage site for *Hin*cII in clone G7-2, allowing the following strategy to be used:

i) The *Xho*I insert of clone G7-2 was isolated [section 3.2.3.2] and subcloned into the *Hin*cII site of pBluescript [section 2.7] for manipulation of the corrected insert.

```
                                *                                   M  A  S  L  G  A  N  P  R  R  T  P  Q  G  P  R  P  G  A  A  S     21
GGCGGTCGGTCAGCGGGGCGTTCTCCCACCTGTAGCGACTCAGAGCCTCCAAGCTCATGGCCTCCTTAGGAGCGAACCCAAGGAGGACACCGCAGGGACCGAGACCTGGGGCGGCCTCCT  120

S  G  F  P  S  P  A  P  V  P  G  P  R  E  A  E  E  E  E  V  E  E  E  E  E  L  A  E  I  H  L  C  V  L  W  N  S  G  Y  L     61
CCGGCTTCCCCAGCCCGGCCCCAGTGCCGGGCCCCAGGGAGGCCGAGGAGGAGGAAGTCGAGGAGGAGGAGGAGCTGGCCGAGATCCATCTGTGTGTGCTGTGGAATTCAGGATACTTGG  240

G  I  A  Y  Y  D  T  S  D  S  T  I  H  F  M  P  D  A  P  D  H  E  S  L  K  L  L  Q  R  V  L  D  E  I  N  P  Q  S  V  V    101
GCATTGCCTACTATGATACTAGTGACTCCACTATCCACTTCATGCCAGATGCCCCAGACCACGAGAGCCTCAAGCTTCTCCAGAGAGTTCTGGATGAGATCAATCCCCAGTCTGTTGTTA  360

T  S  A  K  Q  D  E  N  M  T  R  F  L  G  K  L  A  S  Q  E  H  R  E  P  K  R  P  E  I  I  F  L  P  S  V  D  F  G  L  E    141
CGAGTGCCAAACAGGATGAGAATATGACTCGATTTCTGGGAAAGCTTGCCTCCCAGGAGCACAGAGAGCCTAAAAGACCTGAAATCATATTTTTGCCAAGTGTGGATTTTGGTCTGGAGA  480

I  S  K  Q  R  L  L  S  G  N  Y  S  F  I  P  D  A  M  T  A  T  E  K  I  L  F  L  S  S  I  I  P  F  D  C  L  L  T  P  P    181
TAAGCAAACAACGCCTCCTTTCTGGAAACTACTCCTTCATCCCAGACGCCATGACTGCCACTGAGAAAATCCTCTTCCTCTCTTCCATTATTCCCTTTGACTGCCTCCTCACACCCCCAG  600

G  D  L  R  F  T  P  I  P  L  L  I  P  S  Q  V  R  A  L  G  G  L  L  K  F  L  G  R  R  R  I  G  V  E  L  E  D  Y  N  V    221
GAGATTTAAGATTTACCCCGATTCCACTGCTGATCCCCTCCCAGGTTCGAGCACTTGGAGGGCTGCTGAAGTTCCTGGGTCGAAGAAGAATCGGGGTTGAACTGGAAGACTATAATGTCA  720

S  V  P  I  L  G  F  K  K  F  M  L  T  H  L  V  N  I  D  Q  D  T  Y  S  V  L  Q  I  F  K  S  E  S  H  P  S  V  Y  K  V    261
GCGTCCCCATCCTGGGCTTTAAGAAATTTATGTTGACTCATCTGGTGAACATAGATCAAGACACTTACAGTGTTCTACAGATTTTTAAGAGTGAGTCTCACCCCTCAGTGTACAAAGTGG  840

A  S  G  L  K  E  G  L  S  L  F  G  I  L  N  R  C  H  C  K  W  G  E  K  L  L  R  L  W  F  T  R  P  T  H  D  L  G  E  L    301
CCAGTGGACTGAAGGAGGGGCTCAGCCTCTTTGGAATCCTCAACAGATGCCACTGTAAGTGGGGAGAGAAGCTGCTCAGGCTATGGTTCACACGTCCGACTCATGACCTGGGGGAGCTCA  960

S  S  R  L  D  V  I  Q  F  F  L  L  P  Q  N  L  D  M  A  Q  M  L  H  R  L  L  G  H  I  K  N  V  P  L  I  L  K  R  M  K    341
GTTCTCGTCTGGACGTCATTCAGTTTTTTCTGCTGCCCCAGAATCTGGACATGGCTCAGATGCTGCATCGGCTCCTGGGTCACATCAAGAACGTGCCTCTGATTCTGAAACGCATGAAGT 1080

L  S  H  T  K  V  S  D  W  Q  V  L  Y  K  T  V  Y  S  A  L  G  L  R  D  A  C  R  S  L  P  Q  S  I  Q  L  F  R  D  I  A    381
TGTCCCACACCAAGGTCAGCGACTGGCAGGTTCTCTACAAGACTGTGTACAGTGCCCTGGGCCTGAGGGATGCCTGCCGCTCCCTGCCGCAGTCCATCCAGCTCTTTCGGGACATTGCCC 1200

Q  E  F  S  D  D  L  H  H  I  A  S  L  I  G  K  V  V  D  F  E  G  S  L  A  E  N  R  F  T  V  L  P  N  I  D  P  E  I  D    421
AAGAGTTCTCTGATGACCTGCACCATATCGCCAGCCTCATTGGGAAAGTAGTGGACTTTGAGGGCAGCCTTGCTGAAAATCGCTTCACAGTCCTCCCCAACATAGATCCTGAAATTGATG 1320

E  K  K  R  R  L  M  G  L  P  S  F  L  T  E  V  A  R  K  E  L  E  N  L  D  S  R  I  P  S  C  S  V  I  Y  I  P  L  I  G    461
AGAAAAAGCGAAGACTGATGGGACTTCCCAGTTTCCTTACTGAGGTTGCCCGCAAGGAGCTGGAGAATCTGGACTCCCGTATTCCTTCATGCAGTGTCATCTACATCCCTCTGATTGGCT 1440

F  L  L  S  I  P  R  L  P  S  M  V  E  A  S  D  F  E  I  N  G  L  D  F  M  F  L  S  E  E  K  L  H  Y  R  S  A  R  T  K    501
TCCTTCTTTCTATTCCCCGCCTGCCTTCCATGGTAGAGGCCAGTGACTTTGAGATTAATGGACTGGACTTCATGTTTCTCTCAGAGGAGAAGCTGCACTATCGTAGTGCCCGAACCAAGG 1560
```

```
E  L  D  A  L  L  G  D  L  H  C  E  I  R  D  Q  E  T  L  L  M  Y  Q  L  Q  C  Q  V  L  A  R  A  A  V  L  T  R  V  L  D    541
AGCTGGATGCATTGCTGGGGGACCTGCACTGCGAGATCCGGGACCAGGAGACGCTGCTGATGTACCAGCTACAGTGCCAGGTGCTGGCACGAGCAGCTGTCTTAACCCGAGTATTGGACC 1680

L  A  S  R  L  D  V  L  L  A  L  A  S  A  A  R  D  Y  G  Y  S  R  P  R  Y  S  P  Q  V  L  G  V  R  I  Q  N  G  R  H  P    581
TTGCCTCCCGCCTGGACGTCCTGCTGGCTCTTGCCAGTGCTGCCCGGGACTATGGCTACTCAAGGCCGCGTTACTCCCCACAAGTCCTTGGGGTACGAATCCAGAATGGCAGACATCCTC 1800

L  M  E  L  C  A  R  T  F  V  P  N  S  T  E  C  G  G  D  K  G  R  V  K  V  I  T  G  P  N  S  S  G  K  S  I  Y  L  K  Q    621
TGATGGAACTCTGTGCCCGAACCTTTGTGCCCAACTCCACAGAATGTGGTGGGGACAAAGGGAGGGTCAAAGTCATCACTGGACCCAACTCATCAGGGAAGAGCATATACCTCAAACAGG 1920

V  G  L  I  T  F  M  A  L  V  G  S  F  V  P  A  E  */E  A  E  I  G  A  V  D  A  I  F  T  R  I  H  S  C  E  S  I  S  L  G    661
TAGGCTTGATCACATTCATGGCCCTGGTAGGCAGCTTTGTGCCAGCAGAGTAGGCCGAAATTGGGGCAGTAGACGCCATCTTCACACGAATTCATAGCTGCGAATCCATCTCCCTTGGCC 2040
                                                                  GAG

L  S  T  F  M  I  D  L  N  Q  V  A  K  A  V  N  N  A  T  A  Q  S  L  V  L  I  D  E  F  G  K  G  T  N  T  V  D  G  L  A    701
TCTCCACCTTCATGATCGACCTCAACCAGGTGGCGAAAGCAGTGAACAATGCCACTGCACAGTCGCTGGTCCTTATTGATGAATTTGGAAAGGGAACCAACACGGTGGATGGGCTCGCGC 2160

L  L  A  A  V  L  R  H  W  L  A  R  G  P  T  C  P  H  S  L  C  G  H  Q  L  S  E  P  C  S  A  T  T  A  A  T  R  A  P  G    741
TTCTGGCCGCTGTGCTCCGACACTGGCTGGCACGTGGACCCACATGCCCCCACAGTCTTTGTGGCCACCAACTTTCTGAGCCTTGTTCAGCTACAACTGCTGCCACAAGGGCCCCTGGTG 2280

      A  V  F  D  H  G  D  L  *  749
G7-2  CAGTATTTGACCATGGAGACCTGTGAGGATGGCAACGATCTTGTCTTCTTCTATCAGGTTTGCGAAGGTGTTGCGAAGGCCAGCCATGCCTCCCACACAGCTGCCCAGGCTGGGCTTCCT 2400
      A  V  F  G  L  A  K  G  F  R  T  G  R  R  F  A  K  V  L  R  R  P  A  M  P  P  T  Q  L  P  R  L  G  F  L  T  S  L  W  L    800
G7-3  CAGTATTTGGATTGGCCAAGGGTTTCAGGACAGGAAGGAGGTTTGCGAAGGTGTTGCGAAGGCCAGCCATGCCTCCCACACAGCTGCCCAGGCTGGGCTTCCTGACAAGCTTGTGGCTCG

G7-2  GACAAGCTTGTGGCTCGTGGCAAGGAGGTCTCAGACTTGATCCGCAGTGGAAAACCCATCAAGCCTGTCAAGGATTTGCTAAAGAAGAACCAAATGGAAAATTGCCAGACATTAGTGGAT 2520
      V  A  R  R  S  Q  T  *  788
G7-3  TGGCAAGGAGGTCTCAGACTTGATCCGCAGTGGAAAACCCATCAAGCCTGTCAAGGATTTGCTAAAGAAGAACCAAATGGAAAATTGCCAGACATTAGTGGATAAGTTTATGAAACTGGA 2520

G7-2  AAGTTTATGAAACTGGATTTGGAAGATCCTAACCTGGACTTGAACGTTTTCATGAGCCAGGAAGTGCTGCCTGCTGCCACCAGCATCCTCTGAGAGTCCTTCCAGTGTCCTCCCCAGCCT 2640
G7-3  TTTGGAAGATCCTAACCTGGACTTGAACGTTTTCATGAGCCAGGAAGTGCTGCCTGCTGCCACCAGCATCCTCTGAGAGTCCTTCCAGTGTCCTCCCCAGCCTCCTGAGACTCCGGTGGG 2640

G7-2  CCTGAGACTCCGGTGGGCTGCCATGCCCTCTTTGTTTCCTTATCTCCCTCAGACGCAGAGTTTTTAGTTTCTCTAGAAATTTTGTTTCATATTAGGAATAAAGTTTATTTTGAA  2754
G7-3  CTGCCATGCCCTCTTTGTTTCCTTATCTCCCTCAGACGCAGAGTTTTTAGTTTCTCTAGAAATTTTGTTTCATATTAGGAATAAAGTTTATTTTGAA  2738
```

Figure 3.10 The complete G7 cDNA sequences.
The derived amino acid sequence from the longest ORF is shown above the cDNA sequence including the variations resulting from the base change at position 1971 and the alternative sequences in cDNA inserts G7-2 and G7-3.

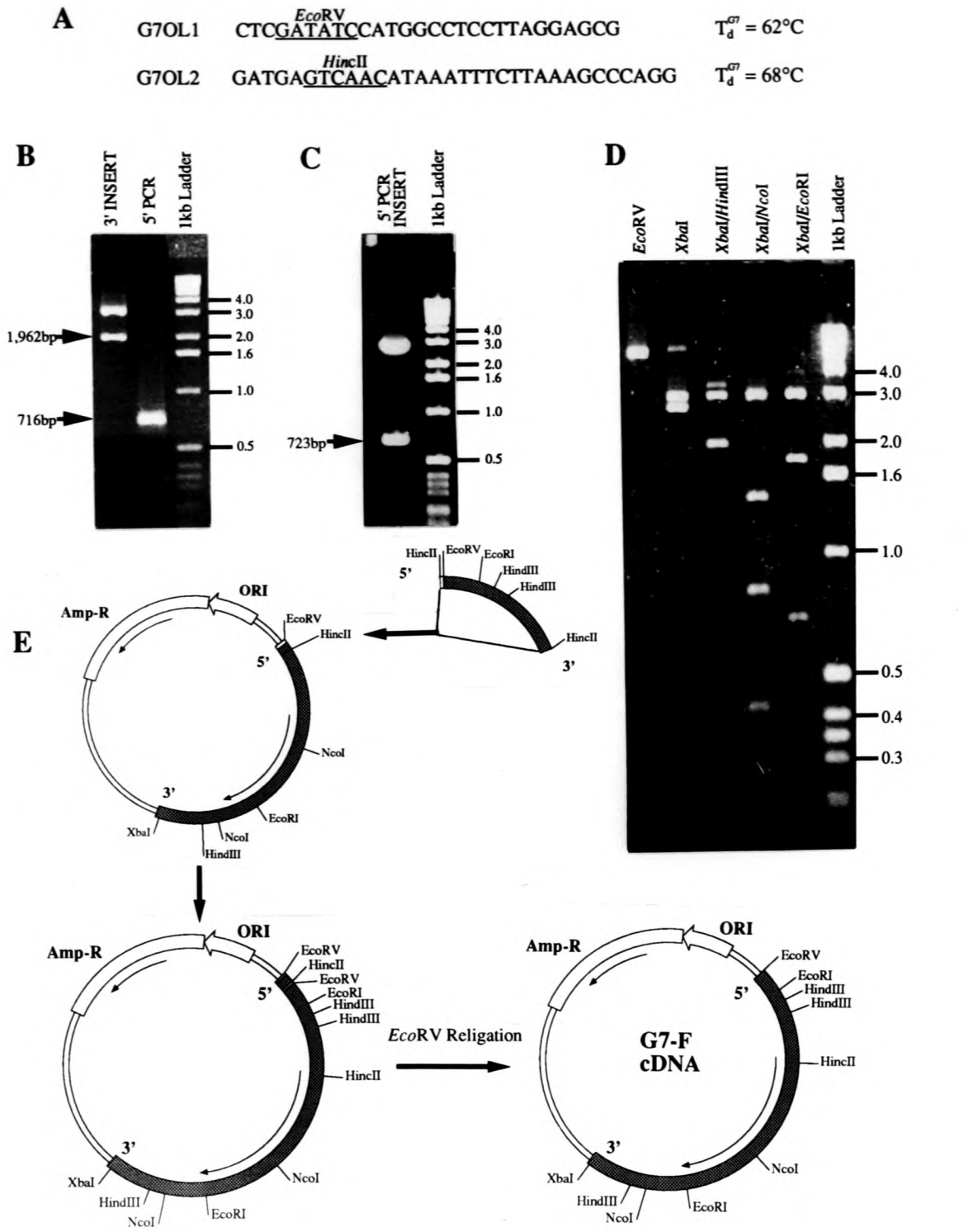


Figure 3.11 G7 cDNA correction strategy.

(**A**) Oligonucleotides G7OL1 and G7OL2 containing adaptor sequences and restriction sites are shown. T_d^{G7} is the dissociation temperature for the nucleotides that hybridise to the G7 cDNA sequence. (**B**) The 3' 1962bp insert was excised from CDM8 and subcloned into pBluescript. The 716bp 5' fragment was amplified by PCR to generate the required specific product. (**C**) The 5' PCR product was subcloned into pBluescript and excised with *Hinc*II as a 723bp fragment. (**D**) Restriction enzyme digests of the corrected, religated G7–F cDNA clone to confirm the orientation of the 5' fragment in the clone.
(**E**) Subcloning and religation manipulations of the G7 cDNA inserts in pBluescript. The G7 cDNA sequences are shaded. The ampicillin resistance gene (Amp-R) and the origin of replication (ORI) are shown. The direction of transcription of the ampicillin resistance gene and the orientation of the G7 cDNA sequences are shown by arrows. Sizes are in kb.

ii) The required sequence from clone G7-1 was amplified by PCR [section 2.3.8] using the G7OL1 and G7OL2 primers with adaptor sequences containing restriction sites incorporated into them (Figure 3.11). At the 5′ end, primer G7OL1 contains an *Eco*RV site upstream of the methionine start codon [section 3.2.6] which would allow in-frame ligation of the cDNA into the *Sma*I site of pGEX-2T [section 3.7.1]. At the 3′ end, primer G7OL2 contains the antisense sequence of G7-2 which allows cleavage with *Hin*cII for ligation with G7-2. The PCR product of 717bp was kinase treated [section 2.7.2], to phosphorylate the ends, and directly ligated into *Eco*RV cut pBluescript. This insert was subcloned, in overlapping fragments, into *Sma*I cut M13mp10 [section 2.7] and sequenced [section 2.9] to check for PCR errors; none were found.

iii) The 5′ insert was excised using *Hin*cII, which cleaves at the 5′ end in the vector and at the 3′ end in the adaptor sequence, and the ampicillin resistance gene was inactivated by digestion with *Pvu*I. The 5′ insert was subcloned into the cut and phosphatase treated *Hin*cII site of the clone that contained the 3′ insert. The resultant colonies were screened by hybridisation with separate probes that identify only the 5′ or 3′ inserts. DNA isolated from duplicating positives was screened using restriction enzyme digestion to identify clones with the 5′ insert in the correct orientation.

iv) Finally, the duplicated, overlapping, sequence at the 5′ end of the corrected clone was deleted by religation of the *Eco*RV sites in the vector and at the 5′ end of the insert [section 2.7.4].

This generated a single, corrected cDNA clone (G7-F) from the G7-1 and G7-2 cDNA inserts.

3.2.8 Southern and Northern blot analysis

To verify the authenticity of the corrected G7 cDNA clone, the G7-F insert was excised from the pBluescript vector by digestion with *Eco*RV and *Xba*I, isolated from LGT agarose and used as a probe in Southern and Northern blot analysis.

This 2.7kb probe was hybridised to Southern blots [section 2.6] of cloned DNA from cosmid 9N and uncloned DNA from the HLA homozygous cell line ICE5, from which the cosmid library was constructed (Figure 3.12; Sargent *et al.* 1989a). The DNA was digested with the restriction enzymes *Hin*dIII, *Bgl*II and *Bam*HI and the pattern of hybridisation obtained from the cosmid blot was consistent with that

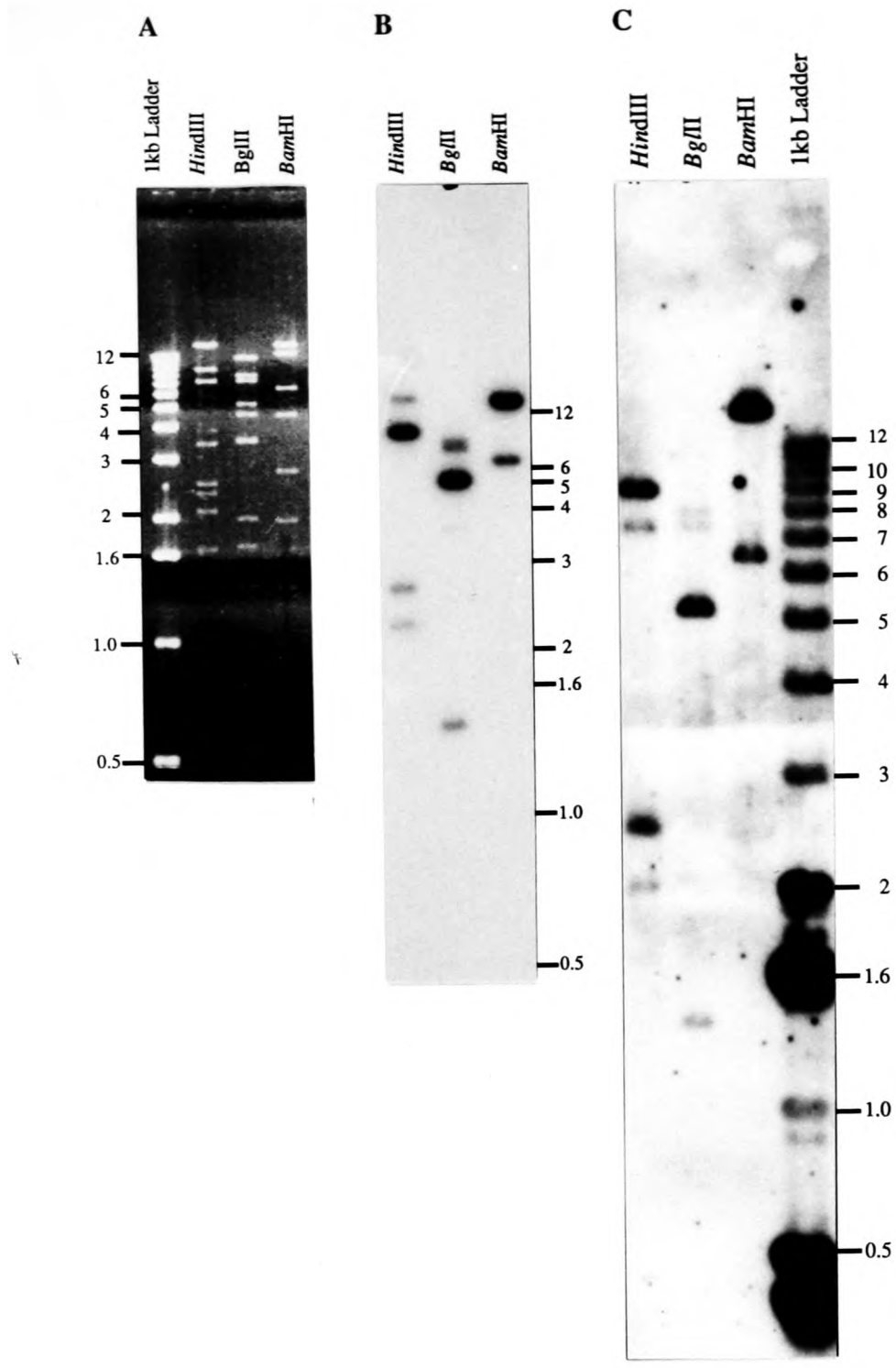


Figure 3.12 Southern blot analysis of G7.

A Digested cosmid 9N DNA fractionated on a 0.8% (w/v) agarose gel. Southern blots of cosmid 9N DNA (**B**) and ICE5 total genomic DNA (**C**) hybridised with the G7-F *Eco*RV/*Xba*I cDNA insert. Sizes are in kb.

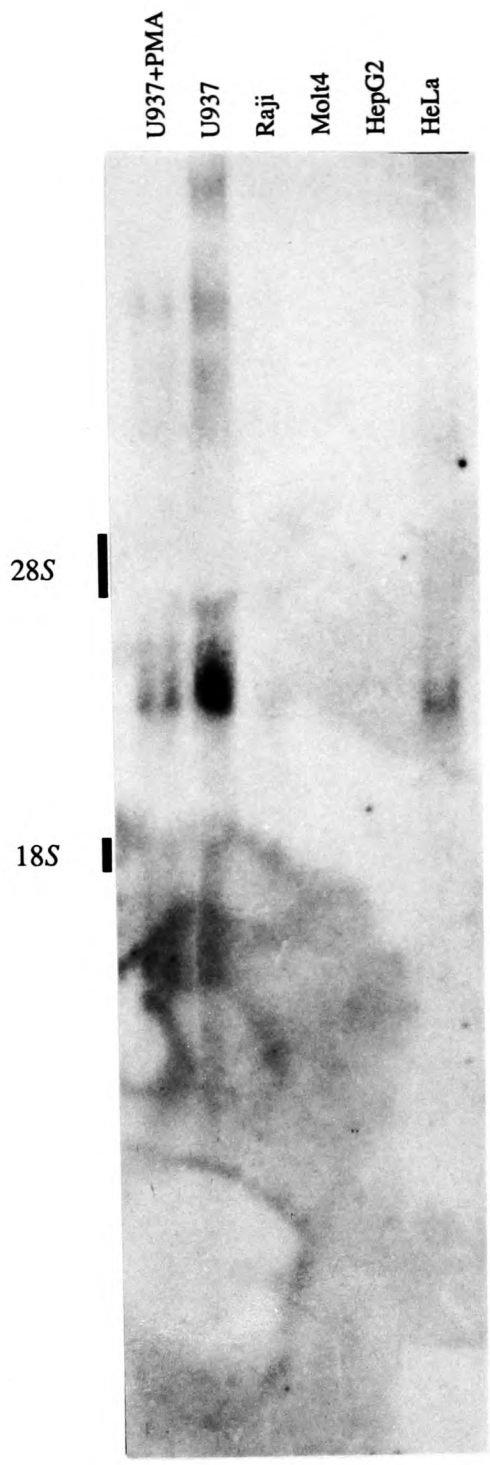


Figure 3.13 Northern blot analysis of G7.
Total RNA (20μg) isolated from the cell lines shown was fractionated on a 1% (w/v) agarose-formaldehyde denaturing gel and transferred onto nitrocellulose. The Northern blot was hybridised with the G7-F *Eco*RV/*Xba*I cDNA insert. The positions of migration of the 18*S* and 28*S* rRNA are shown.

predicted from figure 3.2. The hybridisation pattern on the blot of uncloned DNA was identical to that expected from the restriction sites previously mapped in genomic DNA (Figure 3.1; Sargent *et al.* 1989a). The 7.5kb *Hin*dIII fragment detected in uncloned DNA is not present on the cosmid blot as it extends past the end of the cosmid insert, but is incorporated into the 19kb band which includes the pDVcos vector and the other end of the cosmid insert. No additional bands were seen on the genomic DNA blot, confirming the integrity of the sequences in the G7-F cDNA and demonstrating that G7 is a single copy gene in the human genome.

The G7-F probe was also hybridised to a Northern blot [section 2.6] of total RNA prepared from cultured cell lines representing monocytes (U937), macrophages (U937 cells stimulated with Phorbol 12-myristate 13-acetate [PMA]) B lymphocytes (Raji), T lymphocytes (Molt4), liver (HepG2) and epithelium (HeLa). A mRNA of ~3kb was visible in the cell lines representing monocyte, macrophage and epithelial cell lineages (Figure 3.13), but was not detected in the T and B lymphocyte or liver cell lineages, consistent with the results of Sargent *et al.* (1989a).

3.3 DETERMINATION OF G7 GENE STRUCTURE

The observation of numerous variations between the different cDNA clones necessitated determination of the exonic sequence of G7 in order to confirm the cDNA sequence and to determine the source of the variation. In addition, determination of the organisation of the G7 gene could assist in the identification of the boundaries of functional domains of the G7 protein product encoded within discrete exons

3.3.1 Mapping and subcloning the G7 gene

The fragments generated by *Bgl*II digestion of cosmid 9N, that hybridise to probe G7–F, are 1.3kb, 3.5kb, 5.0kb, 7.3kb and 7.9kb in size (Figure 3.12). These fragments were ligated directly into *Hin*cII cut pBC [section 2.7] and screened with the G7-F cDNA probe [section 2.6]. All the fragments except the largest, 7.9kb, fragment were successfully subcloned. In addition, the 1.6kb *Bgl*II fragment which lies within the gene, but is not hybridised by the probe was also isolated. The orientation and

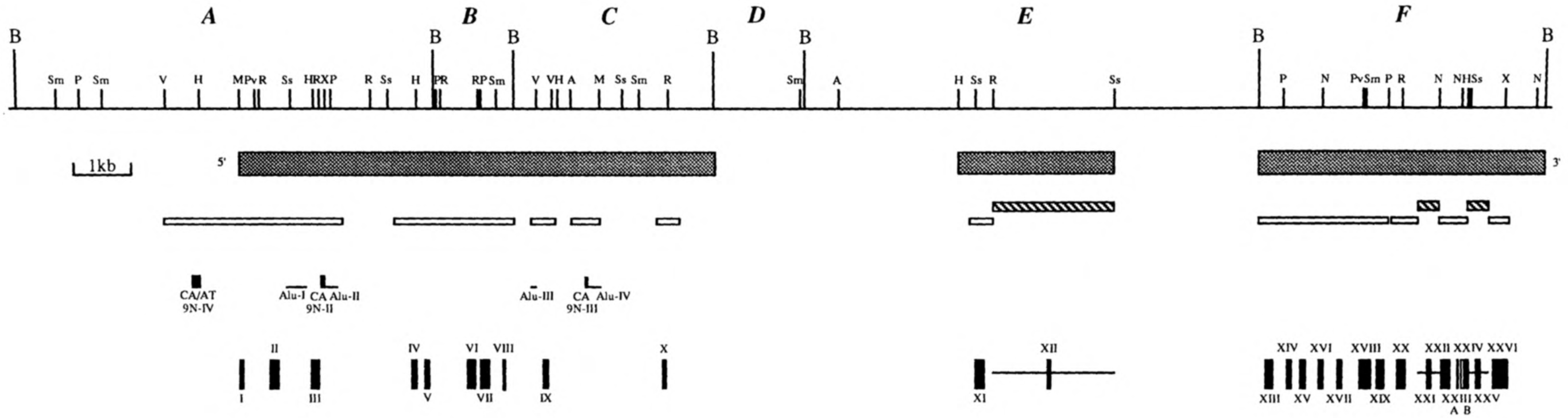


Figure 3.14 Molecular map of the organisation of the G7 gene.

The G7 gene was mapped by hybridisation to the regions shown by the shaded boxes using restriction enzymes *Bam*HI (M), *Bgl*II (B), *Hin*dIII (H) and *Sst*I (Ss).The five *Bgl*II fragments *A-E* were characterised and additional restriction sites for *Eco*RI (R), *Pst*I (P), *Sma*I (Sm), *Xba*I (X) and *Pvu*II (Pv) were mapped and identified by sequencing. Open boxes indicate regions of DNA subcloned and sequenced to locate the G7 exons. The hatched boxes indicate regions from which partial sequence was obtained that could not be positioned. The locations of $(CA)_n$ (small filled boxes) and *Alu* (short lines) repeat sequences identified are shown. The G7 exons I-XXII, XXIIIA, XXIIIB and XXIV-XXVI are shown as filled boxes. Horizontal lines through the boxes indicate that the precise position of the exon could not be determined and the limits of the lines indicate the possible locations of the exons. The region containing exons XXII, XXIIIA, XXIIIB and XXIV is shown in greater detail in figure 3.17.

authenticity of all the inserts was confirmed by restriction enzyme digestion to check the sizes and to compare restriction maps of the inserts with those expected from cosmid 9N digested with *Hin*dIII, *Bgl*II, *Bam*HI, *Eco*RV and *Asp*718. Additional digests were performed in order to map the positions of *Eco*RI, *Pst*I, *Sst*I, *Sma*I and *Xba*I sites in the inserts (Figure 3.14). These *Bgl*II fragments and the subclones containing them were designated A-F from the 5′ end (Figure 3.14) including the 1.6kb fragment.

3.3.2 Sequencing G7 exons

The G7 genomic subclones A, B, C and F [section 3.3.1] (subclone D was not hybridised by the cDNA clones and fragment E was not subcloned) were digested individually with *Alu*I and *Sau*3A restriction enzymes, which cut frequently within the cDNA sequence, and *Rsa*I, which, although it is a frequent-cutting enzyme, cuts only four times in the cDNA sequence. The fragments generated were end-filled [section 2.4.4], if necessary, and ligated directly into *Sma*I cut M13mp10 [section 2.7]. Plaques were screened using the G7-F cDNA [section 2.6], single-stranded DNA was prepared from the positive clones [section 2.9.1] and the inserts were sequenced using the M13 universal primer [section 2.9.3]. The sequence data obtained by this process was entered into the computer and the set of sequences from each clone assembled separately using the GCG program, SAP [section 2.9.7]. Sequence corresponding to nucleotides 1-919 of the G7 cDNA was obtained from subclones A, B and C, whilst only partial sequence between nucleotides 1122-2754 was obtained from clone F.

To obtain more sequence data, including the alternative sequence included in G7–3, but not in G7-F, the procedure was repeated using the restriction enzymes *Hae*III, *Alu*I and *Sau*3A to digest genomic DNA from clone F and the resulting plaques were screened with the G7-3 insert. The new sequence data obtained was used to supplement the sequence already in the database. The remaining sequence from subclone F was obtained directly by using double-stranded plasmid DNA [section 2.9.4], religated at specific sites [section 2.7.4], as templates and the M13 universal primer in all except one case when the specific primer G7OL3 was used. A 400bp *Nco*I fragment containing the two alternative sequences from G7-2 and G7-3 was isolated, subcloned into *Sma*I cut M13mp10 and sequenced in order to confirm

the organisation of this region of the gene.

To sequence the regions of the gene that encode nucleotides 921-1121, a 297bp *Rsa*I fragment corresponding to nucleotides 832-1128 was isolated as a probe from cDNA clone G7-2 [section 2.4.3] and mapped by hybridisation to a region which included fragment E. This probe was then used to screen M13 plaques containing *Hin*fI and *Dde*I fragments of cosmid 9N. The sequence data obtained still did not complete the exonic sequence of G7, but subcloning and sequencing of fragments derived from *Hin*fI/*Dde*I, *Hin*fI/*Stu*I, *Dde*I/*Stu*I and *Alu*I/*Stu*I double digests of cosmid 9N did.

In order to characterise the 5′ end of the G7 gene and the upstream control sequences, the 1.21kb *Bam*HI/*Eco*RV DNA fragment, located upstream of the start of the cDNA sequence, was isolated and subcloned into *Eco*RV cut pBluescript [section 2.7] in both orientations. The insert was mapped with a variety of restriction enzymes and religated [section 2.7.4] at *Hin*dIII, *Spe*I and *Ssp*I sites mapped in the insert. These were then used as templates for double-stranded sequence analysis of this region.

3.3.3 G7 gene structure

3.3.3.1 *Exonic organisation of G7*

By comparison of all the cDNA sequences with the 11kb of genomic sequence obtained in section 3.3.2, the exon sequences, the exon-intron boundaries and the extent of the G7 gene were defined (Figures 3.14 & 3.15). The G7 gene spans ~22kb and is arranged into a total of 27 exons, I–XXII, XXIIIA, XXIIIB, XXIV–XXVI. These vary in size from 31bp in XXIIIA and 36bp in VIII, to 253bp in XXVI. There is no variation between the sequences encoded in the cDNA and the exonic sequences in genomic DNA, except at position 1971 in cDNA clone G7-2 [section 3.2.4].

All the exon-intron boundaries were sequenced and all the splice sites conform to the GT/AG rule (Krainer and Maniatis 1988). The complete splice sites are shown in figure 3.16 except that at the 5′ end of exon XII only the invariant AG sequence flanking the exon was obtained.

The two alternative sequences in G7-2 and G7-3 are encoded by two discrete exons, XXIIIB and XXIIIA, respectively, which lie 38bp apart (Figure 3.15). Usage of these two exons appears to be mutually exclusive and occurs by the splicing patterns shown in figure 3.17.



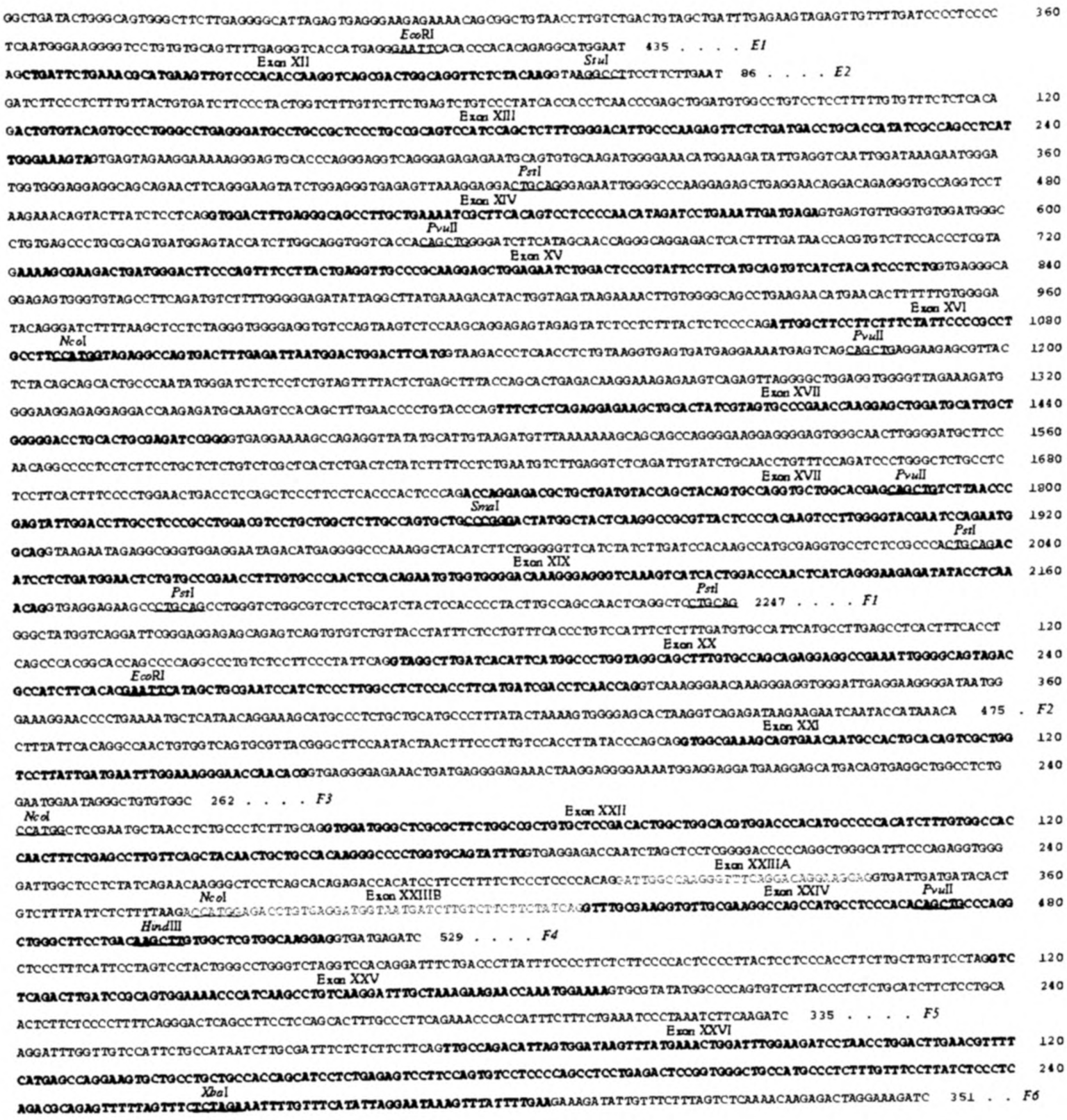


Figure 3.15 Nucleotide sequence of the G7 exons.

The G7 exons are shown in blue, except for the alternative exons XXIIIA and XXIIIB which are green, and numbered as in figure 3.14. $(CA)_n$ and $(AT)_n$ repeats are shown in pink. *Alu* repeats are shown in red. Restriction enzyme recognition sequences are underlined. Contigs are named A1-F6 in accordance with the *Bgl*II fragments from which they were subcloned and ordered from 5' to 3'.

Exon	I	AGCGACTCAG **gt**tactgaaa	intron	1	391bp	tcacttttgcatccgc**ag** AGCCTCCAAG	Exon	II	160bp
Exon	II	GCTGGCCGAG **gt**ctctgagg	intron	2	549bp	tttcttccttgctggac**ag** ATCCATCTGT	Exon	III	124bp
Exon	III	CTCCAGAGAG **gt**ggggatgg	intron	3	1.6kb	tctctgttctccttcca**ag** TTCTGGATGA	Exon	IV	81bp
Exon	IV	GGAAAGCTTG **gt**aaggactt	intron	4	151bp	tttctttcctcccccac**ag** CCTCCCAGGA	Exon	V	63bp
Exon	V	GTGGATTTTG **gt**atctcctt	intron	5	667bp	gcttgcctccctcaaata**g** GTCTGGAGAT	Exon	VI	122bp
Exon	VI	CCTCCTCACA **gt**gagattgg	intron	6	112bp	ttccctggtgcttctgc**ag** CCCCCAGGAG	Exon	VII	161bp
Exon	VII	AATTTATGTT **gt**aggtgatt	intron	7	246bp	ttttgttttctgtcctc**ag** GACTCATCTG	Exon	VIII	36bp
Exon	VIII	ACACTTACAG **gt**aaagaggt	intron	8	0.5kb	ctccatttctcctcgac**ag** TGTTCTACAG	Exon	IX	83bp
Exon	IX	AGCCTCTTTG **gt**aggtgtgc	intron	9	2.0kb	cctgccttatccctcac**ag** GAATCCTCAA	Exon	X	46bp
Exon	X	AGCTGCTCAG **gt**gagtgggt	intron	10	5.4kb	cccaaaccctcacttcc**ag** GCTATGGTTC	Exon	XI	139bp
Exon	XI	GAACGTGCCT **gt**gagcccag	intron	11	N.K.	**ag** CTGATTCTGA	Exon	XII	63bp
Exon	XII	TCTCTACAAG **gt**aaggcctt	intron	12	N.K.	ttttgtgtttctctcac**ag** ACTGTGTACA	Exon	XIII	129bp
Exon	XIII	TGGGAAAGTA **gt**gagtagaa	intron	13	255bp	acagtacttatctcctc**ag** GTGGACTTTG	Exon	XIV	73bp
Exon	XIV	ATTGATGAGA **gt**gagtgttg	intron	14	144bp	gtgtcttccaccctcgt**ag** AAAAGCGAAG	Exon	XV	110bp
Exon	XV	CATCCCTCTG **gt**gagggcag	intron	15	221bp	tcctctttactctcccc**ag** ATTGGCTTCC	Exon	XVI	81bp
Exon	XVI	GGACTTCATG **gt**aagaccct	intron	16	246bp	tttgaacccctgtaccc**ag** TTTCTCTCAG	Exon	XVII	88bp
Exon	XVII	GAGATCCGGG **gt**gaggaaaa	intron	17	268bp	cttcctcacccactccc**ag** ACCAGGAGAC	Exon	XVIII	190bp
Exon	XVIII	AGAATGGCAG **gt**aagaatag	intron	18	118bp	gcctctccgcccactgc**ag** ACATCCTCTG	Exon	XIX	127bp
Exon	XIX	CCTCAAACAG **gt**gaggagaa	intron	19	0.3kb	tgtctccttccctattc**ag** GTAGGCTTGA	Exon	XX	147bp
Exon	XX	CCTCAACCAG **gt**caaaggga	intron	20	N.K.	ttgtccaccttatacccc**ag** CAGGTGGCGA	Exon	XXI	78bp
Exon	XXI	AACCAACACG **gt**gaggggag	intron	21	N.K.	acctctgccctctttgc**ag** GTGGATGGGC	Exon	XXII	145bp
Exon	XXII	GCAGTATTTG **gt**gaggagac	intron	22	134bp	cttttctccctccccac**ag** GATTGGCCAA	Exon	XXIIIA	31bp
Exon	XXII	GCAGTATTTG **gt**gaggagac	intron	22	202bp	tcttttattctcttttaa**g** ACCATGGAGA	Exon	XXIIIB	48bp
Exon	XXIIIA	CAGGAAGGAG **gt**gattgatg	intron	23	85bp	TCTTGTCTTCTTCTATC**AG** GTTTGCGAAG	Exon	XXIV	90bp
Exon	XXIV	TGGCAAGGAG **gt**gatgagat	intron	24	N.K.	ccttcttgcttgttcct**ag** GTCTCAGACT	Exon	XXV	74bp
Exon	XXV	AAATGGAAAA **gt**gcgtatat	intron	25	N.K.	cgatttctctcttcttc**ag** TTGCCAGACA	Exon	XXVI	253bp

Figure 3.16 Exon-intron boundaries in the G7 gene.

The G7 exon sequences at the exon-intron boundaries are shown in upper case and the intron sequences, including the polypyrimidine tracts at the 3' ends, flanking the G7 exons are shown in lower case. The invariant **gt** and **ag** sequences at the splice sites are shown in bold and the sizes of the exons and the introns are included. The splice site at the 5' end of exon XXIV is the 3' and of exon XXIIIB and is in upper case. N.K. not known.

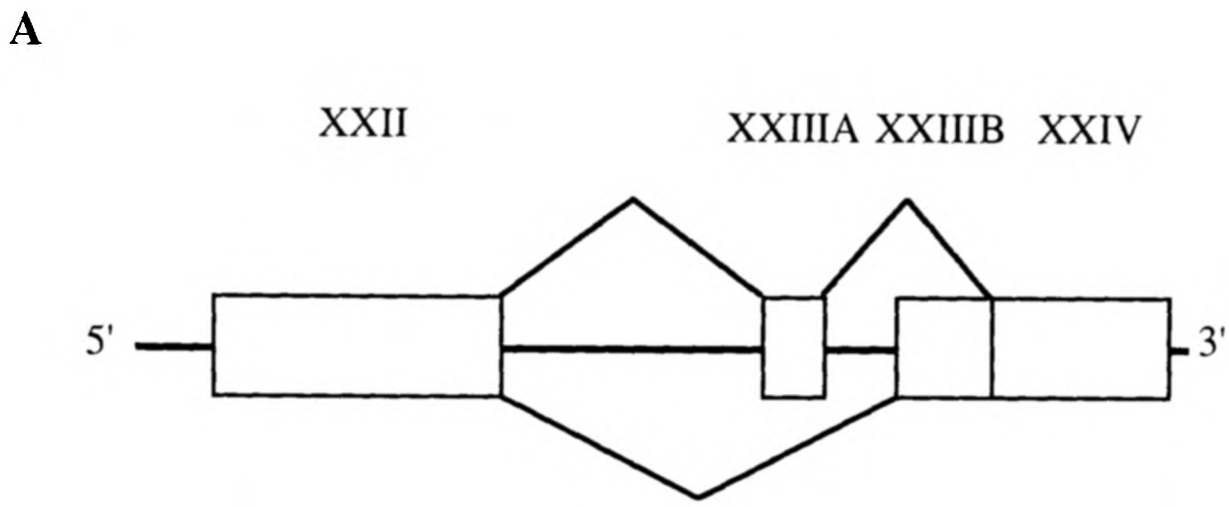


Figure 3.17 Alternative exon usage in G7.
A Organisation of exons XXII, XXIIIA, XXIIIB and XXIV in G7 (shown as open boxes) and alternative splicing patterns. The different splicing patterns result in two different products that include exon XXIIIA (**B**) or exon XXIIIB (**C**) which contain two different ORFs that differ in length as indicated by arrows above the spliced products.

The 5′ untranslated region of the mRNA is encoded in exons I and II. The ORFs [section 3.2.6] start in exon II and extend to exons XXIIIB or XXV so that the 3′ untranslated region is encoded by exons XXIIIB–XXVI or exons XXV–XXVI.

3.3.3.2 *Alu repeats*

Within the partial intronic sequences obtained from the G7 gene, one complete *Alu* repeat and three end fragments of others were identified. The *Alu* family of short interspersed repeat elements (SINES) are the most abundant family of dispersed repetitive DNA sequences in the human genome. It has been estimated that there are 300,000 to 500,000 copies of *Alu* repeats per haploid genome and as such they account for 3-6% of the total DNA (Schmid and Jelinek 1982).

Alu repeats (Figure 3.18) are approximately 300bp in length and are imprecise dimers of two, ~130bp, monomer sequences separated by a central, A-rich, region with a 31bp insertion in the second monomer. At the 3′ end there is an A-rich tract that can be polymorphic (Economou *et al.* 1990) and the sequences are flanked by short direct repeats of 7-20bp that are unique to each sequence. *Alu* repeats have very similar sequences to 7SL RNA molecules and may have been generated by reverse transcription followed by insertion into the genome. The distances between *Alu* repeats is variable but the average separation is ~8kb. *Alu* sequences are often located within the introns of genes, where they are transcribed and later spliced out, but they can also occur in exons encoding the 5′ and 3′ untranslated regions of mRNAs.

One complete *Alu* repeat and three end fragments of others were identified within the sequences flanking G7 exons (Figures 3.14 & 3.15). The complete *Alu* repeat, I, (Figure 3.18) contains a mixed poly(A) tract (Economou *et al.* 1990) at the 3′ end, consisting of three $(A_3G)_n$ repeats and ten adenines. This repeat lies in the second intron and is oriented 5′→3′ with respect to G7.

The next repeat, II, (Figure 3.18) contains 225bp of the *Alu* repeat consensus sequence and has a mixed poly(A) tract of ten adenines followed by a polymorphic (dG-dT).$(dC\text{-}dA)_n$ repeat of 23 dinucleotides in the haplotype sequenced. This sequence lies in the third intron and is oriented 3′→5′ with respect to G7, towards exon III and the first *Alu* repeat.

Within DNA fragment C two more *Alu* repeat sequence fragments were sequenced. The first, III, (Figure 3.18) in the eighth intron is oriented 5′→3′ within G7 and consists of 50bp at the 3′ end of the repeat. This repeat contains a poly(A) tract of

23 bases and is described as type I by Economou *et al.* (1990). The last *Alu* repeat, IV, (Figure 3.18) lies in the ninth intron, oriented 3′→5′ towards exon IX. 225bp of this *Alu* repeat was sequenced and it contains a type II, or patterned, poly(A) tract at the 3′ end which is a polymorphic $(CA)_n$ repeat of 17 dinucleotides in the haplotype sequenced.

The *Alu* repeats were aligned using the CLUSTAL V program (Higgins and Sharp 1988) (Figure 3.18) with the consensus *Alu* repeat sequence of Jurka and Smith (1988). From the sequences of these repeats all four appear to belong to class II (Britten *et al.* 1988) or the consensus, branch *a* of the S subfamily (Jurka and Smith 1988).

3.3.4 Explanation of cDNA sequence variations

With knowledge of the exon boundaries and the flanking intron sequences it is possible to explain the origin of the multiple insertions and deletions observed in the G7-1 and G7-3 cDNA clones [section 3.2.3.4].

Clone G7-1 is encoded by exons I–VII at which point it overlaps clone G7-2. At the 3′ end of exon VII the G7-1 mRNA has utilised a cryptic donor splice site within the following intron resulting in the insertion of four bases (GUAG) including an in-frame stop codon (UAG). The remainder of the intron has been successfully spliced out to join exon VIII, but exon VIII has then been spliced directly to exon X, resulting in a frameshift deletion of exon IX.

Exon X has been successfully spliced to exon XI, except that at the 3′ end, 125bp of intron have been incorporated, by utilisation of a further cryptic donor splice site. Downstream of exon XII the intron was not spliced out and beyond the *Stu*I site the cDNA insert does not hybridise with the other G7 cDNA clones (Figure 3.3).

cDNA clone G7-2 is composed of exons VII–XXVI, including exon XXIIIB and the G→T mutation in exon XX [section 3.2.6].

The 5′ end of clone G7-3 represents a coligated insert and the cDNA insert starts with exon XIII. Exon XIV, of 73bp, has been deleted, resulting in a frameshift. Exons XV-XXVI, including exon XXIIIA and the correct sequence in exon XX, provide the 3′ end of the ORF [section 3.2.6].

```
ALU  GGCCGGGCG-GGTGGCTCAC GCCTGTAATCCCAGCACTTT GGGAGGCCGAGGCGGGCGGA TCACCTGAGGTCAGGAGTTC
I    AGCCGGGCGCGGTGGCTCAC ACCTGTAATCTTAGCACTTT GGGAGACTGAGGCGGGCGGA TCACCTGAGCTCAGGAGTTG
IV   -------------------- -------------------- ------------------GGA TCCCCTGAGGTCAGGAGTTC
II   -------------------- -------------------- ------------------GA TCAC--AAGGTCAGGAGTTT
III  -------------------- -------------------- -------------------- --------------------

ALU  GAGACCAGCCTGGCCAACAT GGTGAAACCCCGTCTCTACT AAAAATACA-----AAAATT AGCCGGGCGTGGTGGCGCGC
I    GAGACCAGCCTGGGCAACAT GGCGAAACCCCGTCTCTACT AAAAATATA-----AAAATT AGCCGAGCGTGGTGGCACGT
IV   AAGACCAGCCTGGGCAACAT GGTGAAACCCCGTCTCTACT AAAAATACA-----AAAATT AGCTGGGCGTGGTGGCAGGT
II   GAGACCACCCTGGCCAACAT GGTGAAACACCGTCTGTGCT AAAAACACACACAAAAAACT AGCTGGGTGTGGTGGCACAC
III  -------------------- -------------------- -------------------- --------------------

ALU  GCCTGTAATCCCAGCTACTC GGGAGGCTGAGGCAGGAGAA TCGCTTGAACCCGGGAGGCG GAGGTTGCAGTGAGCCGAGA
I    GCCTGTTATCCCAGCTACTG GGAAGGCTGAGGCAGGAGAA TCACTGTAACTCGGGAGGCG GAGGTTGCAATGAGCTGAGA
IV   GCCTGTAATCCCAGTTACTC AGGAGGCTGAGGCAGGAGAA TCGCTTGAACCCAGGAGGCA GAGGTTGCATTGAGCCGAGA
II   GCCTGTAGTCCCAGCTACTC GGGAGGTTGAGTCAGGAGAA TTGCTTGAACCCAGGAGGTG GAGGCTGCAGTGAGCCAAGA
III  -------------------- -------------------- -------------------- ----------CTGAAA

ALU  TCGCGCCACTGCACTCCAGC CTGGGCGACA-GAGCGAGAC TCCGTCTC A-rich tract
I    TTGCTCCACTGCACTTCAGC CTGGGCGACA-GAGCAAGAC TCCGTCTC (AAAG)3 AAAAAAAAA
IV   TCGCACCATTGCACTCCAGC CTGGGGGACAAGAGCGAGAC TTCGTCT (CA)17 AAA
II   TTGCGCCACTGCACTTGAGC CTGGGTAACA-GAGCAAGAC TCTGTCT AGAAAAAAAAAATA (TG)23
III  TTGCACCACTGCACTCCAGC CTGGGCGACA-GAGCGAGAC TCCATCTC AAAAAAAAAAAAAAAAAAAAAA
```

Figure 3.18 Alignment of the G7 *Alu* repeat sequences with the consensus sequence.
The *Alu* repeat sequences obtained from the G7 genomic sequence (I-IV) are aligned with the consensus sequence (ALU) of Jurka and Smith (1988). The alignment was produced by the CLUSTAL program (Higgins and Sharp 1988) and adjusted manually.

3.3.5 Upstream sequence analysis of the G7 gene

The genomic sequence obtained at the 5′ end of G7 [section 3.3.2] included the *Bss*HII restriction sites used to identify the HTF-island associated with the 5′ end of the G7 gene (Sargent *et al.* 1989a). The CpG content of 3kb of sequence including exons I-III and 1.5kb of upstream sequence was analysed using the GCG program CPGPLOT (Figure 3.19). The result confirmed the presence of a CpG-rich sequence between positions 1100 and 1600, which includes the start of the 5′ end of the G7 cDNA.

A 1.5kb region of DNA sequence upstream of the 5′ end of the G7 cDNA (Figure 3.20), which extends to within 1.3kb of the 5′ end of G6, was examined for sequences involved in the control of gene expression. In this sequence a potential 3′ splice site was observed that conforms to the consensus sequence described by Mount (1982), 7bp upstream of the 5′ end of the cDNA (Figure 3.20). This suggests the possibility that there is another exon upstream of exon I. However, this would be unusual in that the start methionine would lie in the third exon and the G7 gene would include two 5′ exons consisting of untranslated sequence. In the absence of any RNase mapping data or extended sequence from anchored PCR experiments it is not possible to confirm whether or not there is another exon upstream of the start of the cDNA sequence.

The 1.5kb of upstream sequence was analysed for the presence of transcription factor binding sites by comparison with the consensus elements reported by Locker and Buzard (1990) and Jones *et al.* (1988) and the results are shown in figure 3.20.

Multiple CAAT boxes and GC boxes but only a single TATA box were identified within the sequence. These promoter elements are normally located at specific positions relative to the transcription start site of a gene. In addition multiple transcription factor binding sites are present in the sequence and these appear to be clustered in two regions between positions 180–510 and 1060–1470 (Figure 3.20).

If there is no additional upstream exon in G7 then the two potential start sites, mapped by primer extension analysis at positions 1225 and 1348, lie within the second cluster of putative transcription factor binding sites (Figure 3.20). This positions them more than 140bp downstream of the TATA box, which is usually located ~30bp upstream of the transcription start site, and there is also no apparent correlation between the locations of the GC and CAAT boxes and the potential start sites. However, the presence of multiple GC boxes has been shown to promote

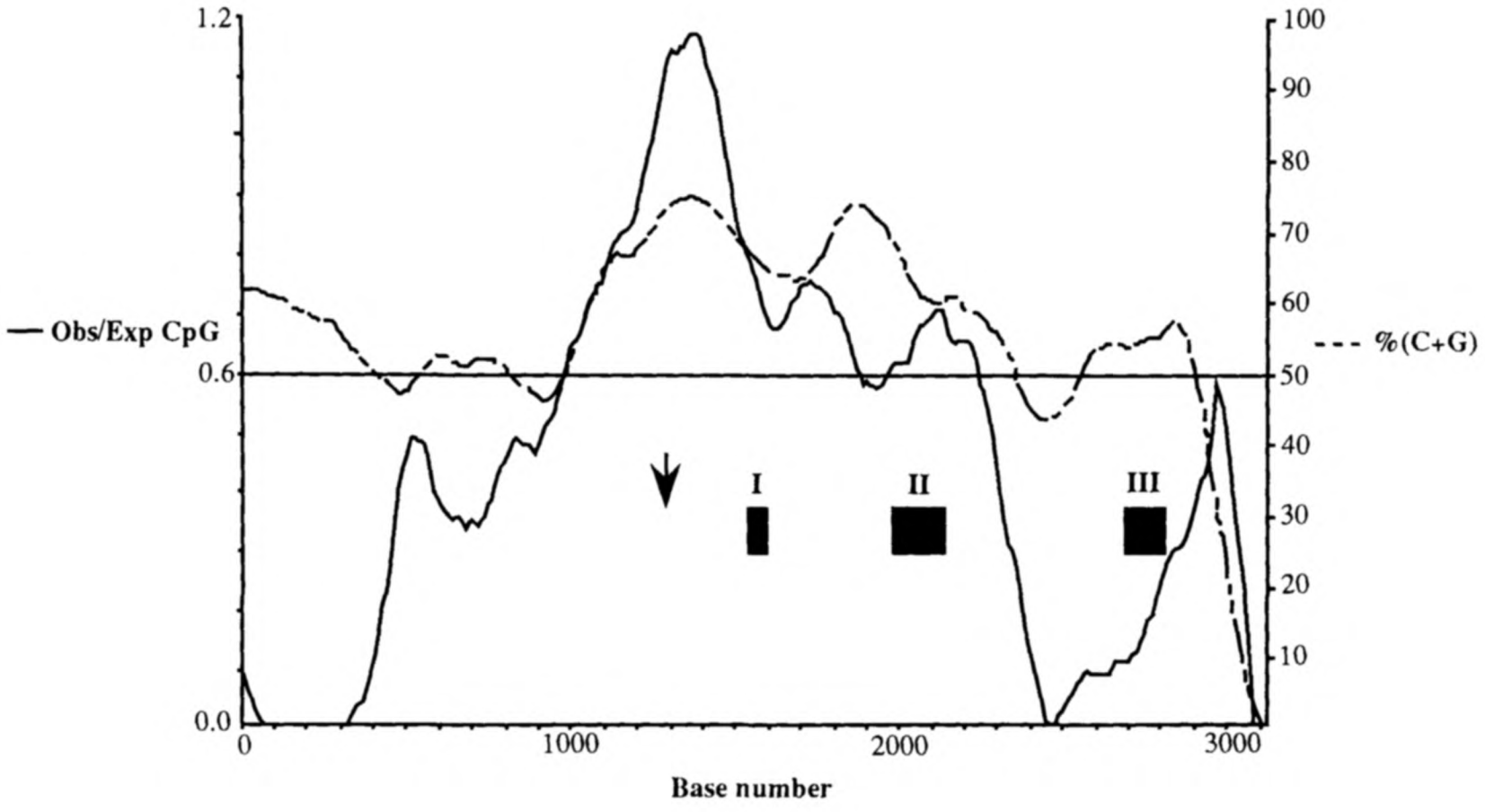


Figure 3.19 Gardiner-Garden Plot of the 5' end of G7.

The CpG content of the nucleotide sequence from contig *A1* (Figure 3.15) analysed by the CPGPLOT program, using a 200 base window and shift increment of 10, is shown. Exons I, II and III of the G7 geneare shown as filled boxes and the position of the furthest 5' extension is shown as a vertical arrow.

Obs/Exp is (No. of CpG dinucleotides/No. of C x No. of G) x N

where N is the window size. (Gardiner-Garden and Frommer 1987)

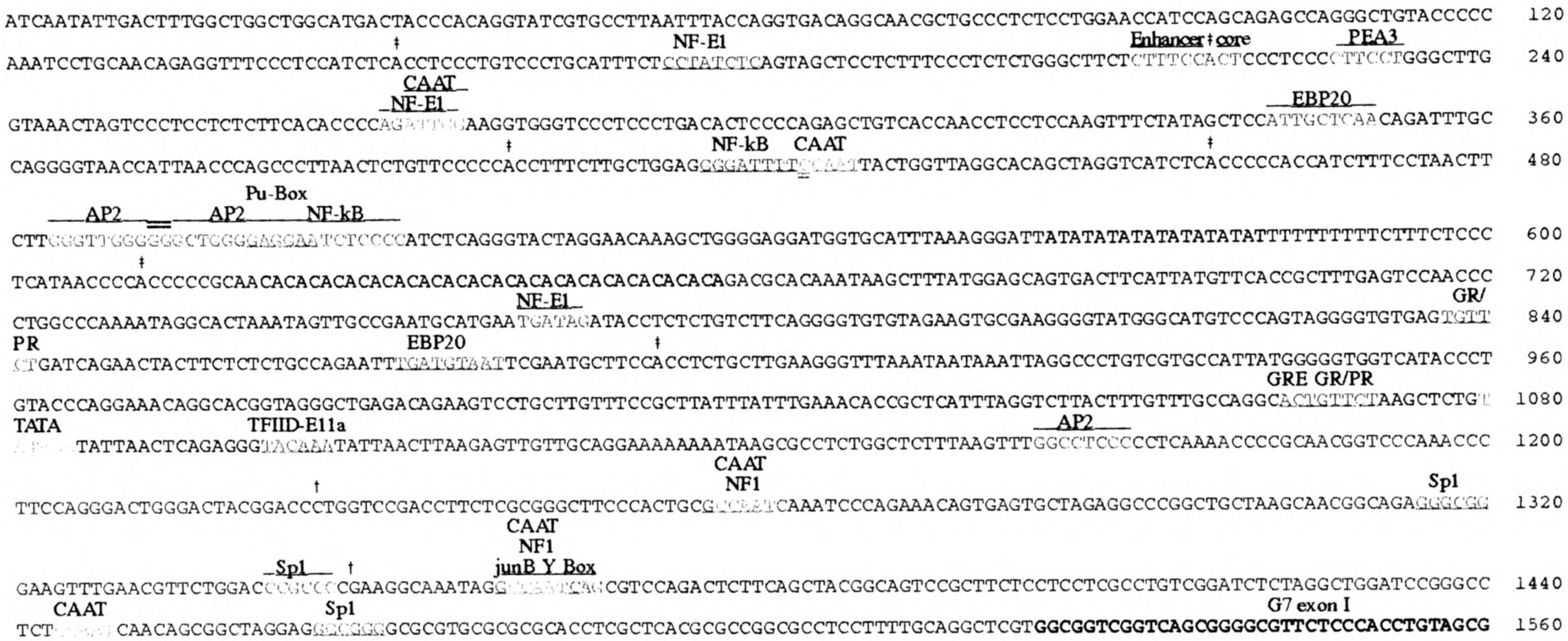


Figure 3.20 Analysis of the 5' end of the G7 gene.

The 5' end of exon I is shown in blue and the HTF island in figure 3.19 is shown in pink. CAAT and TATA boxes sequences are shown in light blue. Potential binding sites for transcription factors are shown in green, those oriented 5' to 3' are underlined and those oriented 3' to 5' are overlined. $(CA)_n$ and $(AT)_n$ repetitive sequences are shown in red. † indicates the 5' limits of primer extension products and ‡ indicates the position of transcription start site consensus sequences YYCAYYYYY (Corden et al. 1980).

transcription from multiple start sites, and potential binding sites for many other gene-regulating elements have also been identified within this region (Figure 3.20).

Potential binding sites for transcription factors AP2, GR/PR, GRE and TFIID are located in the region between nucleotides 1060–1370 (Figure 3.20) which includes the potential start sites. The GC and CAAT boxes lie between positions 1260 and 1470 which is downstream of the mapped, potential start sites and much closer to the 5′ end of the cDNA sequence which starts at position 1550.

From this information it is unclear where transcription of G7 is initiated. The start site may lie closer to the upstream cluster of transcription factor binding sites between positions 180–510 since the only sequence resembling the transcription start consensus sequence YYCAYYYY (Corden *et al.* 1980) lies between positions 150–890 (Figure 3.20). However, most 5′ CpG islands begin before the transcription start sites and extend much further downstream into the gene rather than upstream into the 5′ untranscribed flanking DNA (Gardiner-Garden and Frommer 1987), suggesting that G7 transcription is initiated within ~200bp of the 5′ end of the cDNA sequence.

3.4 EXPRESSION OF G7 mRNA

3.4.1 Tissue distribution of G7 expression

Northern blot analysis of G7 (Figure 3.13) indicates that G7 mRNA is expressed in monocytes, macrophages and in epithelial cell lines, but not in T or B lymphocyte or liver cell lines. In order to confirm this result, total RNA from monocytic (U937 and HL60), B lymphocyte (Daudi), T lymphocyte (Molt4), liver (HepG2) and epithelial (HeLa) cell lines was used in an RT-PCR assay. Equal amounts of total RNA were reverse transcribed and amplified by PCR using oligonucleotides G7OL3 and G7OL4 [section 3.2.4]. Thirty cycles of PCR (94°C for 60s; 64°C for 30s: 72°C for 120s) were carried out using the reverse transcription product from ~200ng total RNA as template. To confirm that the 900bp amplified products were derived from equal amounts of template, control reactions were carried out by amplifying single-stranded G9 cDNA, generated by reverse transcription of equal amounts of total RNA, using sequence specific primers. G9 is a novel MHC-linked gene which is constitutively expressed in all cell lines (Sargent *et al.* 1989a).

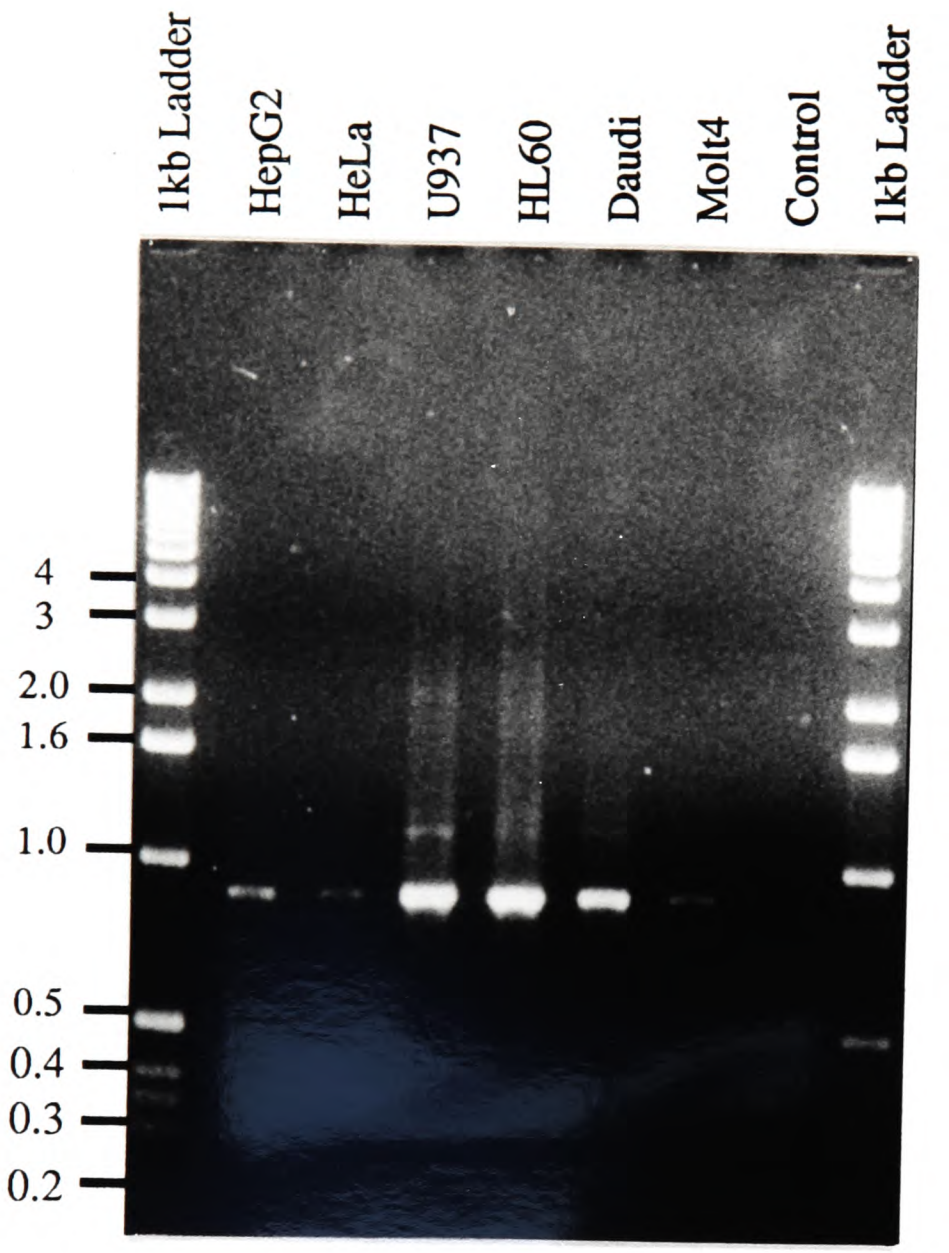


Figure 3.21 Analysis of G7 expression by PCR.
RT-PCR of mRNA (~200ng) from the cell lines shown was carried out for thirty cycles of 94°C for 60s, 64°C for 30s and 72°C for 120s using oligo-nucleotide primers G7OL3 and G7OL4. The reaction products are shown following fractionation on a 0.8% (w/v) agarose gel. Products of ~900bp are visible in the samples from all of the cell lines examined. Sizes are in kb.

The results indicate that G7 is expressed in all the cell lines examined (Figure 3.21). However, the amount of G7 RT-PCR product generated from RNA isolated from U937, HL60 and Daudi cells was higher than that from the other cell lines. In the control experiment, the product generated from the mRNA isolated from Daudi cells was increased by a similar amount, indicating that this sample may have contained more template than the other cell lines. In addition, the level of G7 product generated from mRNA isolated from HeLa cells was less than from U937 or HL60 cells, indicating that the relative levels of mRNA production indicated by the Northern blot data was due to unequal loading of the total RNA samples on the gel.

From this analysis it is apparent that, the distribution of G7 expression is ubiquitous although the levels generated in the monocytic cell lines U937 and HL60 appear to be elevated. The Northern blot data [section 3.2.8] can be explained by the fact that the amount of G7 mRNA is very low and the signal was present, but undetectable in Molt4, Raji and HepG2 cell lines.

3.4.2 Exon usage in G7 mRNA

The cDNA sequences obtained in section 3.2.3 (Figure 3.5) include two alternatively spliced exons, XXIIIA and XXIIIB, [section 3.3.3.1] which result in two different putative protein products [section 3.2.6]. In order to determine which putative G7 protein, if either, is the major product produced, the relative usage of these two exons was investigated.

Using the oligonucleotide primers G7OL3 and G7OL4 [section 3.2.4], single-stranded cDNA derived from U937, HeLa, HepG2 and Daudi cell lines was amplified by PCR [section 3.4.1]. A 900bp product was generated which includes the alternative exon sequences. These vary by 17bp in length and by the presence or absence of an *Nco*I restriction site. Exon XXIIIA consists of 31bp whilst exon XXIIIB consists of 48bp, resulting in the 17bp difference in size between the two species. Restriction enzyme digestion using *Eco*RI cleaves the ~900bp PCR product to produce two DNA fragments of 490bp and either 393bp or 410bp that can be resolved by agarose gel electrophoresis (Figure 3.22). In addition, cleavage by *Nco*I is also diagnostic for the G7-2 specific sequence derived from exon XXIIIB, resulting in DNA fragments of 128bp and 283bp. Exon XXIIIA does not contain a cleavage site for *Nco*I. From figure 3.22 it is clear that two fragments of ~400bp are generated following digestion of the

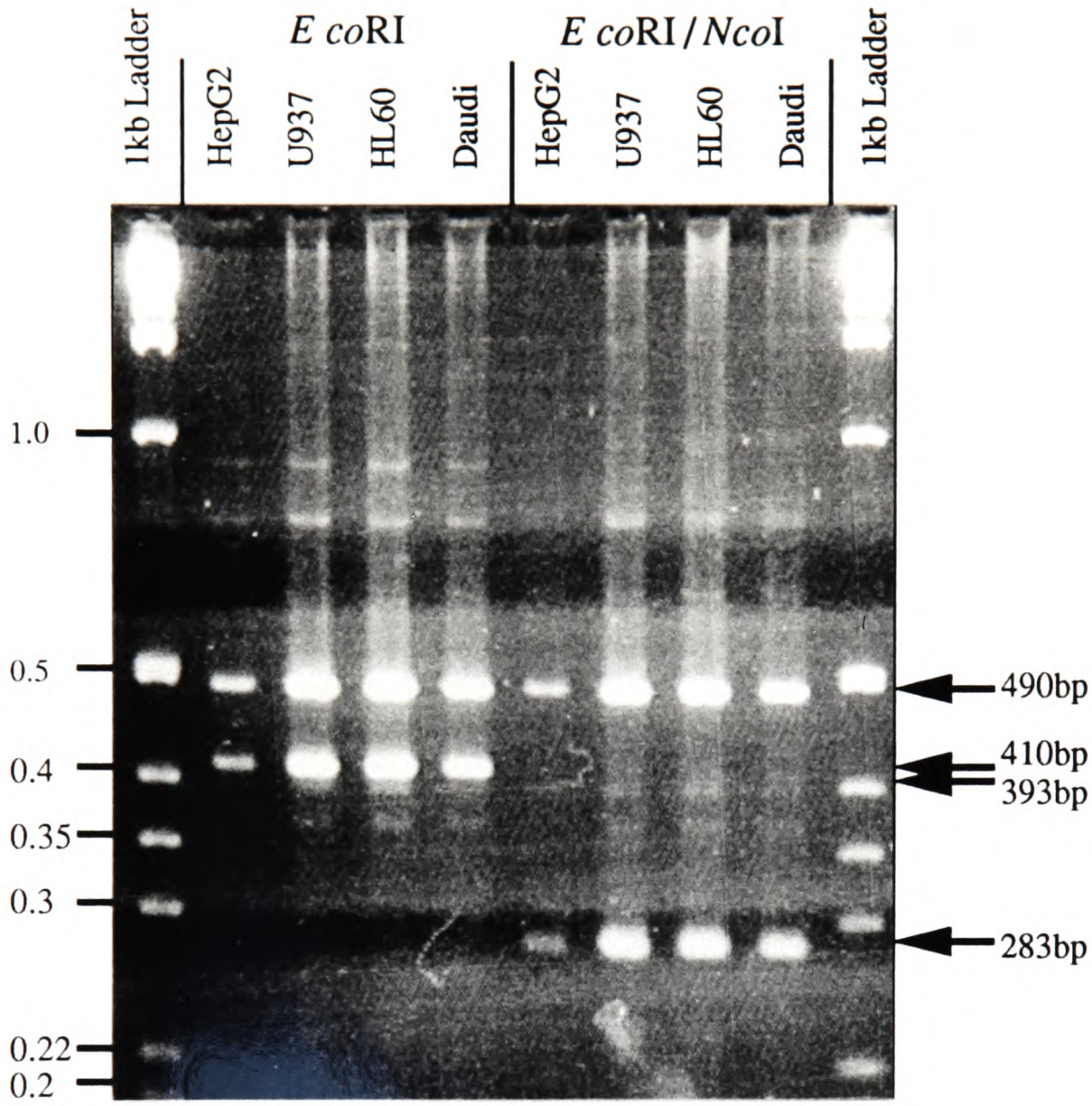


Figure 3.22 Analysis of alternative exon usage in G7 mRNA.
The RT-PCR products from cell lines HepG2, U937, HL60 and Daudi, amplified using oligonucleotide primers G7OL3 and G7OL4 and digested with *Eco*RI and *Eco*RI/*Nco*I are shown. The samples were fractionated on a 2% (w/v) agarose gel at 4°C to reduce diffusion of the products. Two products are visible following *Eco*RI digestion of 393bp and 410bp confirming the use of two different exons. The presence of exon XXIIIB in the more abundant 410bp product is confirmed by the subsequent digestion using *Nco*I which generates the 283bp product shown.Sizes are in kb.

PCR products with *Eco*RI with the major product being the upper, 410bp one. Upon cleavage with *Nco*I this major product is cleaved to produce 283bp and 127bp fragments whilst the minor, 393bp product is not cleaved. These two results indicate that the major product of G7 utilises exon XXIIIB generating the smaller 2247bp ORF which terminates at the stop codon at nucleotide 2304.

There are, however, several PCR products present (Figures 3.21 & 3.22) which are detected by hybridisation with G7 cDNA that cannot be explained by the cDNA sequence data. Following PCR there are products of ~1.2kb and ~800bp present and digestion with *Eco*RI generates additional fragments of ~900bp, 800bp and 360bp. The 900bp product may represent uncleaved PCR product due to mutation of the *Eco*RI site during PCR, but the 1.2kb product, which is cleaved by *Eco*RI, and the 800bp and 360bp fragments may represent other minor G7 products. These fragments either do not contain the *Eco*RI site, resulting in the 800bp fragment, or contain different sequences between exons XX and XXIV to leave the 360bp product. Further sequence analysis of these minor products will be necessary to determine the differences with the cDNA clones resulting in these products.

3.5 CHARACTERISATION OF THE G7 GENE PRODUCTS

3.5.1 Characteristics of the G7 protein sequence

The longest ORF identified in the various G7 cDNA sequences, which was confirmed by the sequencing of the G7 exons in genomic DNA, encodes two different proteins of 749 amino acids and 788 amino acids (Figure 3.23), of which the smaller, 749 residue protein is the major G7 product [section 3.4.2].

The N-terminal sequence of the predicted G7 protein products does not contain a hydrophobic leader peptide (von Heijne 1988) and a hydrophobicity plot of the putative protein sequence (Figure 3.24) indicates that there are no hydrophobic regions characteristic of transmembrane segments. As a result, the predicted G7 protein products are probably intracellular proteins. Both of the predicted protein products contain 14 cysteine residues, which in an intracellular environment will remain as reduced thiol groups (Barbosa *et al.* 1989). The sequences also contain 6 potential N-glycosylation sites ($NX^{S}/_{T}$), but these will remain unglycosylated in an

```
G7-1      MASLGANPRRTPQGPRPGAA SSGFPSPAPVPGPREAEEEE VEEEEELAEIHLCVLWNSGY LGIAYYDTSDSTIHFMPDAP DHESLKLLQRVLDEINPQSV 100
G7-1/2    VTSAKQDENMTRFLGKLASQ EHREPKRPEIIFLPSVDFGL EISKQRLLSGNYSFIPDAMT ATEKILFLSSIIPFDCLLTP PGDLRFTPIPLLIPSQVRAL 200
G7-1/2    GGLLKFLGRRRIGVELEDYN VSVPILGFKKFMLTHLVNID QDTYSVLQIFKSESHPSVYK VASGLKEGLSLFGILNRCHC KWGEKLLRLWFTRPTHDLGE 300
G7-2      LSSRLDVIQFFLLPQNLDMA QMLHRLLGHIKNVPLILKRM KLSHTKVSDWQVLYKTVYSA LGLRDACRSLPQSIQLFRDI AQEFSDDLHHIASLIGKVVD 400
G7-2/3    FEGSLAENRFTVLPNIDPEI DEKKRRLMGLPSFLTEVARK ELENLDSRIPSCSVIYIPLI GFLLSIPRLPSMVEASDFEI NGLDFMFLSEEKLHYRSART 500
G7-2/3    KELDALLGDLHCEIRDQETL LMYQLQCQVLARAAVLTRVL DLASRLDVLLALASAARDYG YSRPRYSPQVLGVRIQNGRH PLMELCARTFVPNSTECGGD 600
G7-3      KGRVKVITGPNSSGKSIYLK QVGLITFMALVGSFVPAEEA EIGAVDAIFTRIHSCESISL GLSTFMIDLNQVAKAVNNAT AQSLVLIDEFGKGTNTVDGL 700
G7-2      ..................... ..................*. ..................... ..................... ..................... 700

G7-3      ALLAAVLRHWLARGPTCPHS LCGHQLSEPCSATTAATRAP GAVFGLAKGFRTGRRFAKVL RRPAMPPTQLPRLGFLTSLW LVARRSQT* 788
G7-2      ..................... ..................... ....DHGDL* 749
```

Figure 3.23 Derived amino acid sequences from the G7 cDNA sequences.

The amino acid sequences translated from the G7 cDNA clones are shown. Lines are labelled to show the cDNA clones from which the derived amino acid sequences were obtained. Identical amino acids are indicated by a dot (.) and stop codons are indicated by an asterisk (*).

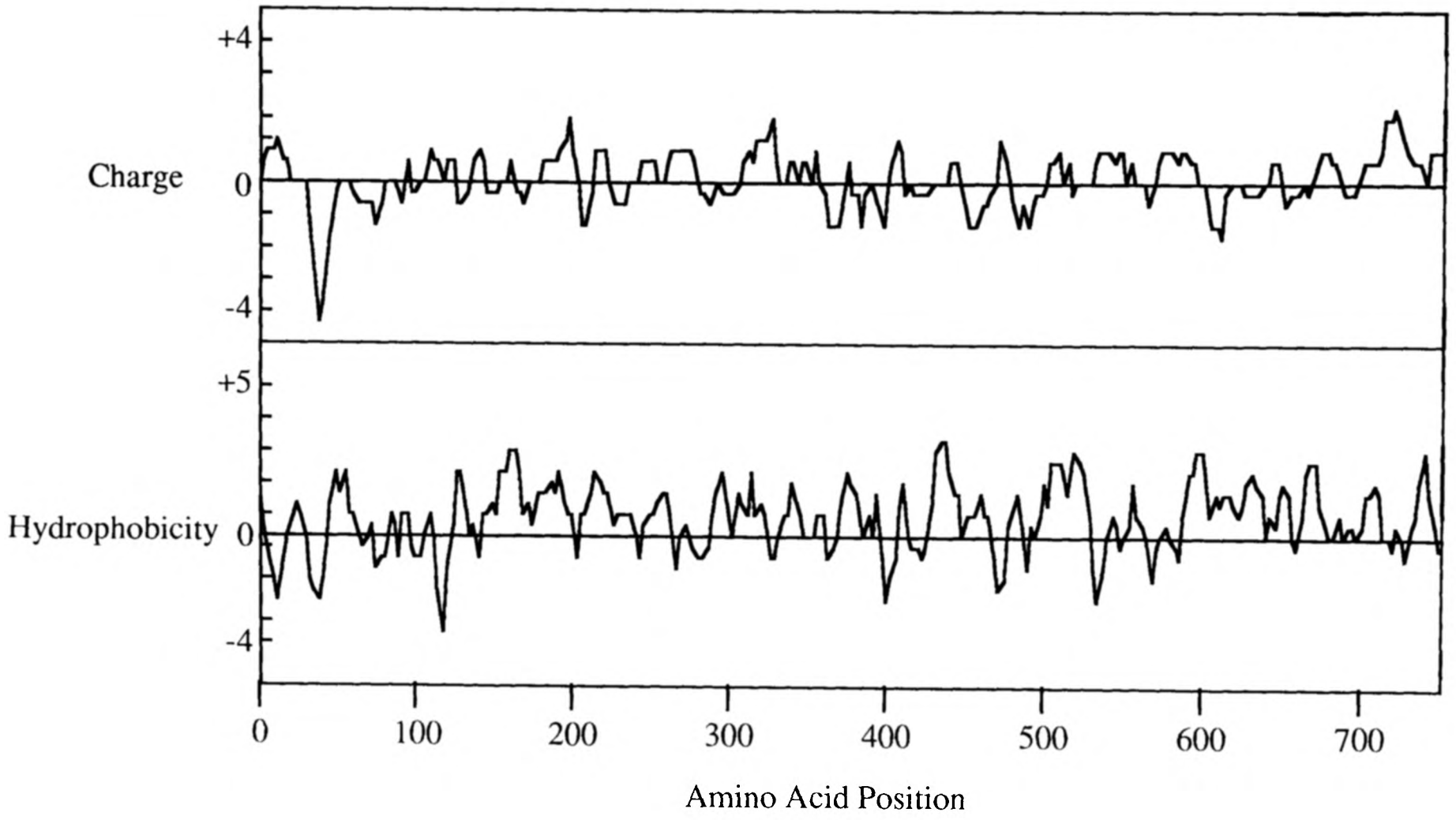


Figure 3.24 Hydrophobicity and Charge profiles of the putative G7 protein sequence. The plots were generated using the Protein Interpretation Program (PIP) to analyse the 749 residue putative G7 protein sequence using the default parameters.

PROTEIN	POSITION	% SIMILARITY	LENGTH	POSITION IN G7
S.cerevisiae MIBP	468-937	22.9%	475 aa	270-727
Human Divergent Upstream Protein (DUP)	547-998	27.8%	464 aa	258-711
S.cerevisiae MutS protein homologue 2 (MSH2)	267-788	24.7%	530 aa	202-709
S.pombe Swi4	426-858	24.3%	444 aa	270-708
S,typhimurium DNA mismatch repair protein MutS	422-707	31.9%	295 aa	409-702
E.coli DNA mismatch repair protein MutS	333-707	28.4%	388 aa	318-702
S.cerevisiae MutS protein homologue 1 (MSH1)	260-724	21.0%	648 aa	87-724
S.pneumoniae DNA mismatch repair protein HexA	238-700	27.3%	483 aa	233-707
Mouse Repair-3 protein	510-925	23.0%	430 aa	510-684

Table 3.1 Putative G7 protein similarities.

The proteins sharing most significant amino acid sequence similarity detected with the 749 residue putative G7 protein sequence using the FASTA program (Pearson and Lipman 1988) are listed. The regions sharing similarity to the putative G7 sequence, the length of the region and the % identity are shown.

aa amino acids

intracellular environment. In the N-terminal region there is a negatively charged region (Figure 3.24) containing 10 glutamic acid residues out of 12 amino acids (EAEEEEVEEEEE). Similar regions of negatively charged, acidic residues have previously been identified in a number of nuclear proteins, such as, nucleolin (Srivastava *et al.* 1990), human major centromere autoantigen (CENP-B) (Earnshaw *et al.* 1987) and the non-histone chromosomal proteins (HMG-1 and HMG-2) (Wen *et al.* 1989, Shirakawa 1990). It has been suggested that proteins containing acidic regions may be associated with chromatin (Earnshaw 1987). However, although there are clusters of basic residues at positions 210 and 230 of the putative G7 protein products, as well as in the C-terminal sequence unique to the less abundant form of G7 protein, these sequences do not constitute nuclear localisation signals.

A DIAGON plot (Devereux *et al.* 1984) of the putative G7 protein products indicated that the sequences do not contain any repetitive sequence elements.

3.5.2 G7 protein sequence similarities

The predicted G7 protein sequence was screened against the SwissProt and NBRF protein databases [section 2.12.1] to search for protein similarity with other known protein sequences which may indicate a possible function of the G7 protein. The longer predicted G7 protein sequence was used and divided into four overlapping 220 amino acid sequences to increase the sensitivity of the search.

The results of the search indicated that the only significant similarities were between 21.0% and 31.9% over between 295 and 648 amino acid residues (Table 3.1). These similarities all occurred within the region from amino acids 90 to 720. The three bacterial proteins, MutS from *Salmonella typhimurium* (Haber *et al.* 1988), MutS from *Escherichia coli* (Schlensog and Block 1991) and HexA from *Streptococcus pneumoniae* (Priebe *et al.* 1988) are DNA mismatch repair proteins from the MutHLS and the HexAB systems. Swi4, from *Schizosaccharomyces pombe,* is one of the *swi* gene products that regulate switching of mating type (Fleck *et al.* 1992). The human Divergent Upstream Protein (DUP) (Fujii and Shimada 1989) and the mouse *Rep*-3 protein, originally named *Rep* 1 (Linton *et al.* 1989), are both encoded by genes which are located upstream of Dihydrofolate Reductase (DHFR) genes, transcribed from the opposite strand. In addition the putative mismatch binding protein (MIBP or YMBP) (Valle *et al.* 1991) and the MutS protein

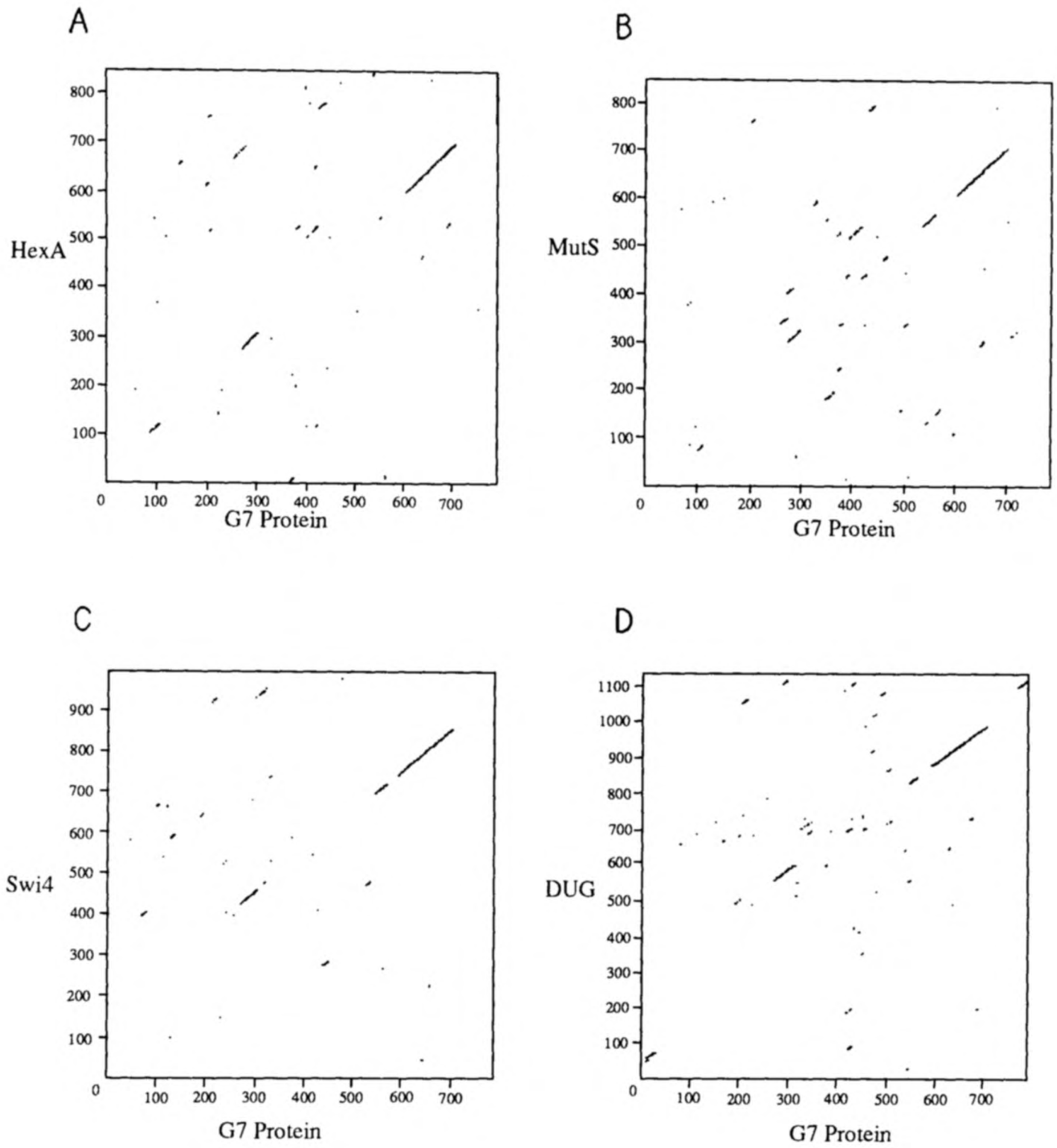


Figure 3.25 DIAGON plots of putative G7 protein sequence similarities.
This shows comparisons of the 749 residue G7 protein sequence with the *S. pneumoniae* HexA (**A**), *S.typhimurium* MutS (**B**), *S.pombe* Swi4 (**C**) and human DUP (**D**) proteins. The most similar sequences are visible in the C-terminal regions of these proteins.

homologues, MSH1 [EMBL accession no. M84169] and MSH2 [EMBL accession no. M84170] (Kolodner and Reenan unpublished), have been identified in *Saccharomyces cerevisiae*. All of these proteins with similarity to the putative G7 proteins are mismatch repair proteins or have significant homology to the DNA mismatch repair proteins.

DIAGON plots, using the SIP program from the STADEN package, using a 31 amino acid window and a proportional score of 340, of the longer putative G7 protein product versus HexA, MutS from *Salmonella typhimurium*, Swi4 and DUP (Figure 3.25) indicated that the highest regions of similarity lie towards the C–terminus of the putative G7 protein products, but the similarity extends to over half of the protein sequence. At the N-terminus, the plots with Swi4 and DUP do not pass through the origin, indicating that these proteins contain additional sequence at the N-terminus.

To determine the extent of the similarity with G7 the sequences were aligned using the CLUSTAL V program (Higgins and Sharp 1988). From this alignment, the greatest sequence similarity was localised between amino acid positions 270–709 of the putative G7 protein sequence. This sequence and the corresponding sequences from the proteins showing similarity to the putative G7 protein sequence were aligned using the multiple sequence alignment package AMPS (Alignment of Multiple Protein Sequences: Barton 1990, Barton and Sternberg 1990). The analysis was carried out using the 250 PAMs mutation data matrix (Dayhoff *et al.* 1978) with a bias parameter of 8 added to each term and a gap penalty of 6. Fifty random runs of the ten, ~450 amino acid long sequences were carried out to calculate the mean random score. The ALIGN scores of the pairwise analyses (in SD units) represent the standard deviation from the mean random score. The pairwise scores for the set of sequences are between 17 and 84 SD units indicating very significant sequence similarities between all of the sequences. The alignment is shown in figure 3.26.

SD scores can be used to plot a dendrogram as shown in figure 3.27. This dendrogram indicates that the set of mismatch repair proteins can be divided into three groups. The bacterial MutS and HexA proteins constitute one group and, unsurprisingly, the gram-negative enterobacterial MutS proteins are more closely related to each other than to the gram-positive streptococcal HexA protein. The second group include the human DUP protein and its mouse homologue *Rep*-3, the

MUTSE
MUTSS
HEXA
REP3
DUG
SWI4
MIBP
MSH2
G7
MSH1

Binding Site
Motif I
Motif II
Motif III
ATP-

Figure 3.26 Alignment of the putative G7 protein sequence with DNA mismatch repair proteins and their homologues. The putative G7 protein sequence (residues 270-708) is aligned with *E.coli* MutS (289-713), *S. typhimurium* MutS (292-713), *S.pneumoniae* HexA (274-701), mouse *Rep-3* (517-949), human DUG (559-995), *S.pombe* Swi4 (426-858), *S.cerevisiae* MIBP (468-919), *S.cerevisiae* MSH1 (447-870) and *S.cerevisiae* MSH2 (337-788). The alignment was generated using the AMPS package (Barton and Sternberg 1990) and formatted using the ALSCRIPT program (Barton 1993). At positions where 8 or more sequences contain identical or conserved amino acid residues the residues are boxed and at positions where 8 or more identical residues occur the residues are also shaded. Conserved residues within the ATP-binding site and the conserved motifs I-III are shown in bold.

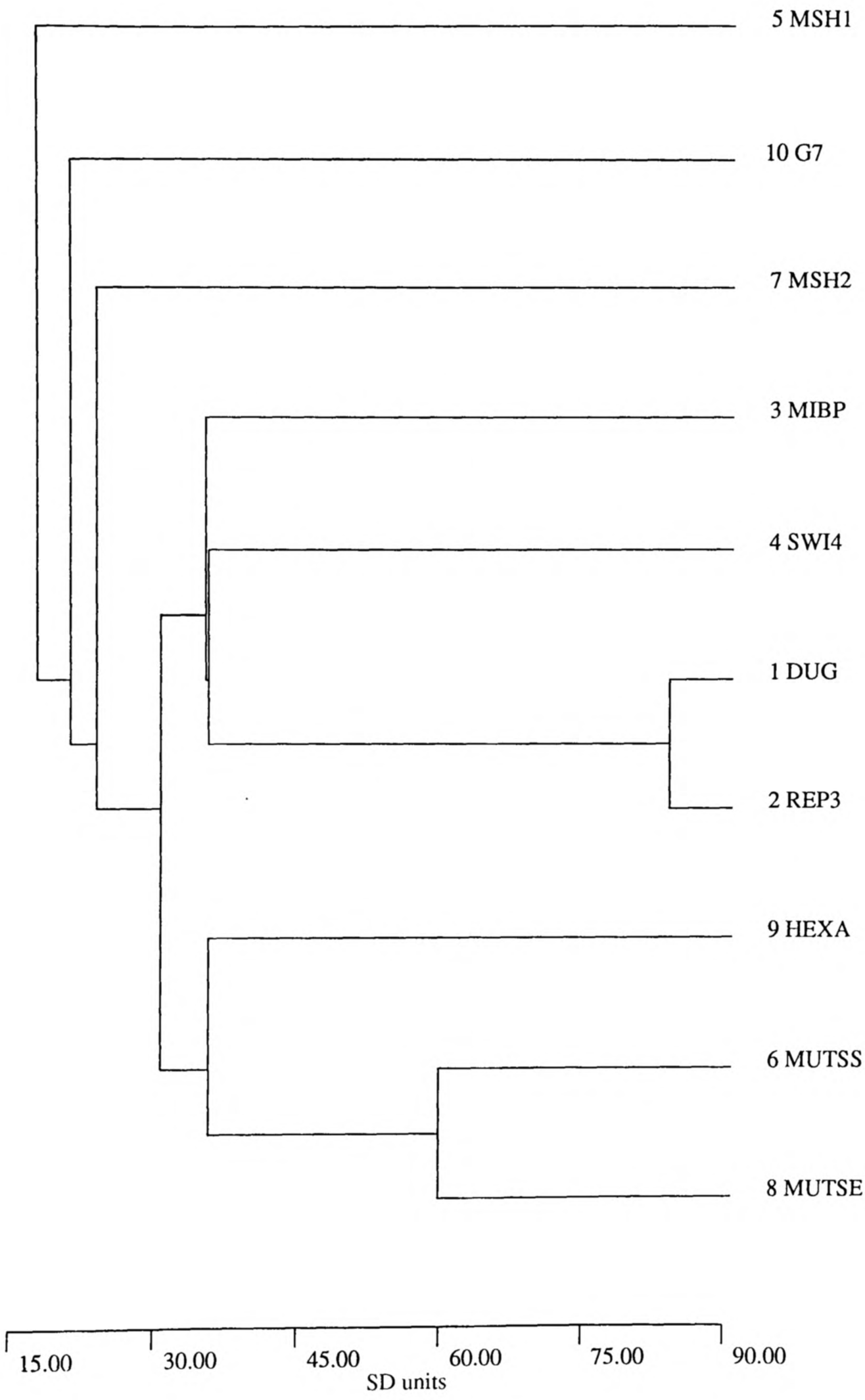


Figure 3.27 A dendrogram of the DNA mismatch repair proteins and similar proteins. The ALIGN scores generated using the AMPS package were used to plot the dendrogram to show the similarity, in SD units, between the bacterial DNA mismatch repair proteins, their homologues and similar proteins.

mating-type switching protein, Swi4, from *Schizosaccharomyces pombe* and MIBP from *Saccharomyces cerevisiae*. Both of the yeast proteins appear to have diverged from the mammalian sequences at approximately the same time. The third group, consisting of the G7 protein, MSH1 and MSH2, from *Saccharomyces cerevisiae*, are all divergent from each other and the other similar proteins, including MIBP, also from *Saccharomyces cerevisiae*, although they do still share very significant similarity. This group may be members of a structurally and functionally similar protein family, but the closer similarity between the bacterial mismatch proteins and DUP, *Rep*–3, Swi4 and MIBP suggest that these are more likely to be members of the same family, involved in the recognition of mismatched DNA. Negative complementation of HexA in *E. coli* (Prudhomme *et al.* 1991) indicates that HexA binds to mismatched DNA and, by association, MutS, DUP, *Rep*–3, Swi4 and MIBP may play a similar role.

The most conserved region in all of the mismatch repair proteins and their homologues is near the C-terminus, spanning amino acids 606 to 702 in the putative G7 protein sequence. This sequence includes a type A nucleotide-binding site for ATP or GTP (Saraste *et al.* 1990) which is present in all the mismatch repair proteins and their homologues (Figure 3.26) and has been shown to exhibit ATPase activity in MutS from *Salmonella typhimurium* (Haber and Walker 1991). In addition there are other highly conserved motifs in this region. The region of similarity observed by Haber and Walker (1991) contains two highly conserved motifs, I and II (Figure 3.26), flanking a 25 amino acid sequence which is most divergent in G7, Swi4, MIBP, MSH1 and MSH2. Fleck *et al.* (1992) observed another motif, DELGRGTSTXDG, included in motif III (Figure 3.26). So far no structure or function has been assigned to these highly conserved sequences which are all unique to this group of proteins. These highly conserved regions are also encoded within discrete exons. The ATP-binding site is encoded within exon XIX, motifs I and II are encoded within exon XX and motif III is encoded within exon XXI.

All the members of this group of proteins, except G7, contain DNA-binding site consensus sequences (Pabo and Sauer 1984) at the C-terminus and all the eukaryotic members, except G7, contain nuclear localisation signals close to casein kinase II phosphorylation sites. However, not all nuclear proteins contain recognised nuclear localisation signals (Dingwall and Laskey 1991). The putative G7 protein is the

smallest member of this group of proteins and the DNA-binding site and nuclear localisation sites occur in the additional sequence at the N- and C-termini of the other proteins. Alternatively, G7 may be located in the mitochondria or may constitute part of a protein complex in which another protein has DNA binding activity.

3.6 POLYMORPHIC ANALYSIS OF G7

Although a large number of diseases have been shown to be associated with the MHC, linkage disequilibrium has so far prevented the mapping of these diseases to specific loci within the MHC. In order to assist in determining whether or not the G7 gene is associated with any inherited disorder or susceptibility to any autoimmune disease, analysis of G7 for RFLPs associated with the gene that may be informative in disease studies was carried out. In addition, dinucleotide $(CA)_n$ repeats, often found to be highly polymorphic and very informative, have also been located within the G7 gene.

3.6.1 RFLP analysis

To search for RFLPs, a series of Southern blots were prepared using total genomic DNA [section 2.3.4] isolated from ten, HLA-typed, unrelated individuals (1 to 10 in table 3.2). The DNA was digested with the restriction enzymes *Bam*HI, *Bgl*II, *Hin*dIII, *Bst*EII, *Nco*I, *Pvu*II, *Pst*I, *Sst*I, *Eco*RI, *Rsa*I, *Taq*I and *Msp*I and fractionated on 0.8% (w/v) agarose gels, except for the *Taq*I- and *Msp*I-digested DNA which was fractionated in 1% (w/v) agarose. Probing by hybridisation with the G7-F, full length, cDNA probe detected fragments on all of the blots. Polymorphism was only observed following digestion with *Pvu*II (Figure 3.28), when an extra 1.7kb DNA fragment was observed in the DNA from individuals 4, 6, 7 and 8 (Table 3.2; Figure 3.28). Another Southern blot (a gift from S J Cross) containing *Pvu*II-digested total genomic DNA from an additional five HLA typed individuals, A1-5 (Table 3.2), was also hybridised with the G7-F probe. The same 1.7kb *Pvu*II variant was present in three of this second group of five individuals, resulting in a total of 7 (24%) in the limited sample of 29 haplotypes analysed (ICE5 is an HLA homozygous cell line and therefore only

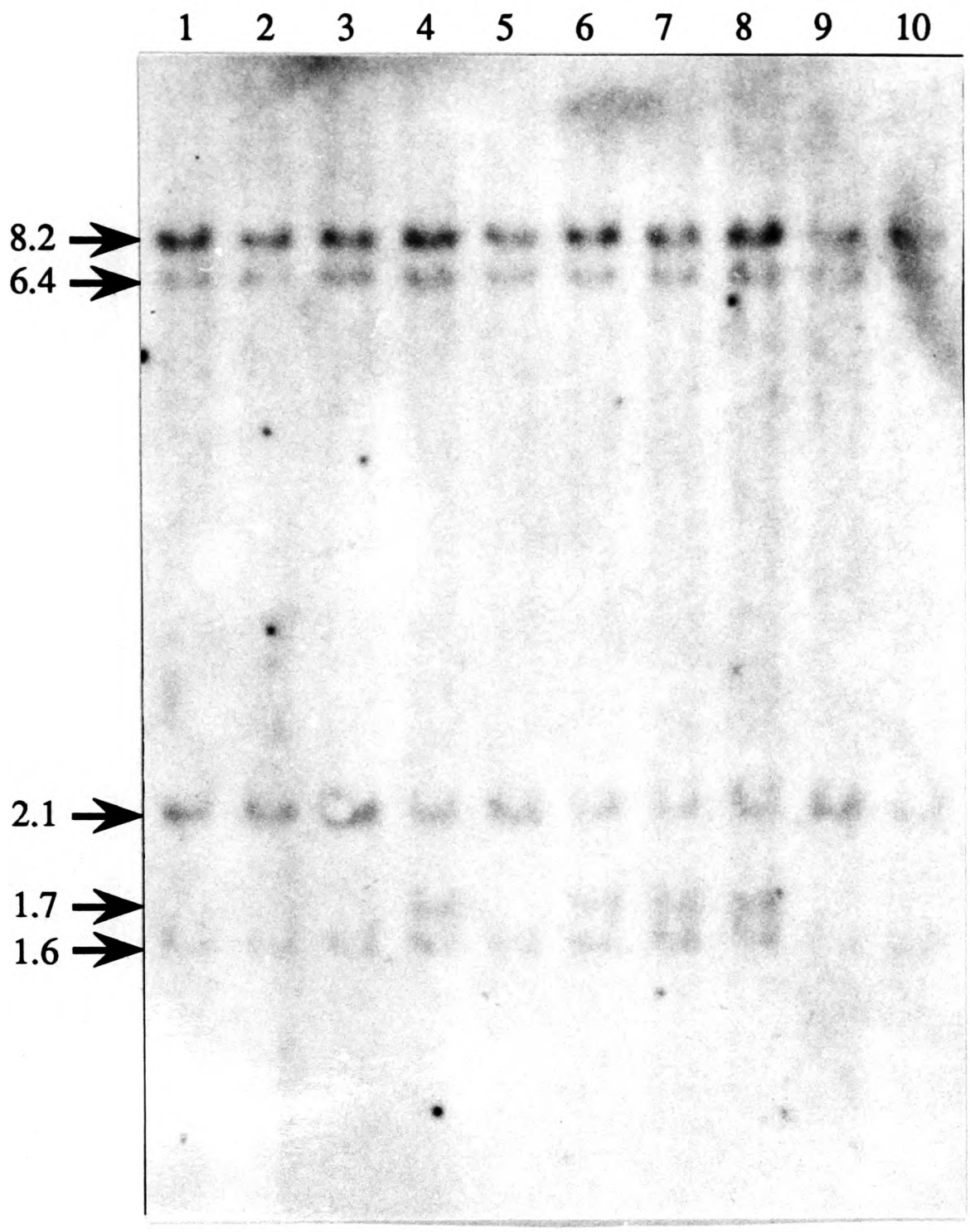


Figure 3.28 *Pvu*II RFLP in the G7 gene.

A Southern blot of genomic DNA from ten unrelated HLA-typed individuals digested with *Pvu*II, fractionated on a 0.8% (w/v) agarose gel and transferred onto nitrocellulose was hybridised with the 2.7kb *Eco*RV/*Xba*I cDNA insert. The invariant restriction fragments of 8.2, 6.4 and 1.6kb are indicated and the polymorphic fragments of 2.1 and 1.7kb, present in individuals 4,6,7 and 8, are marked.

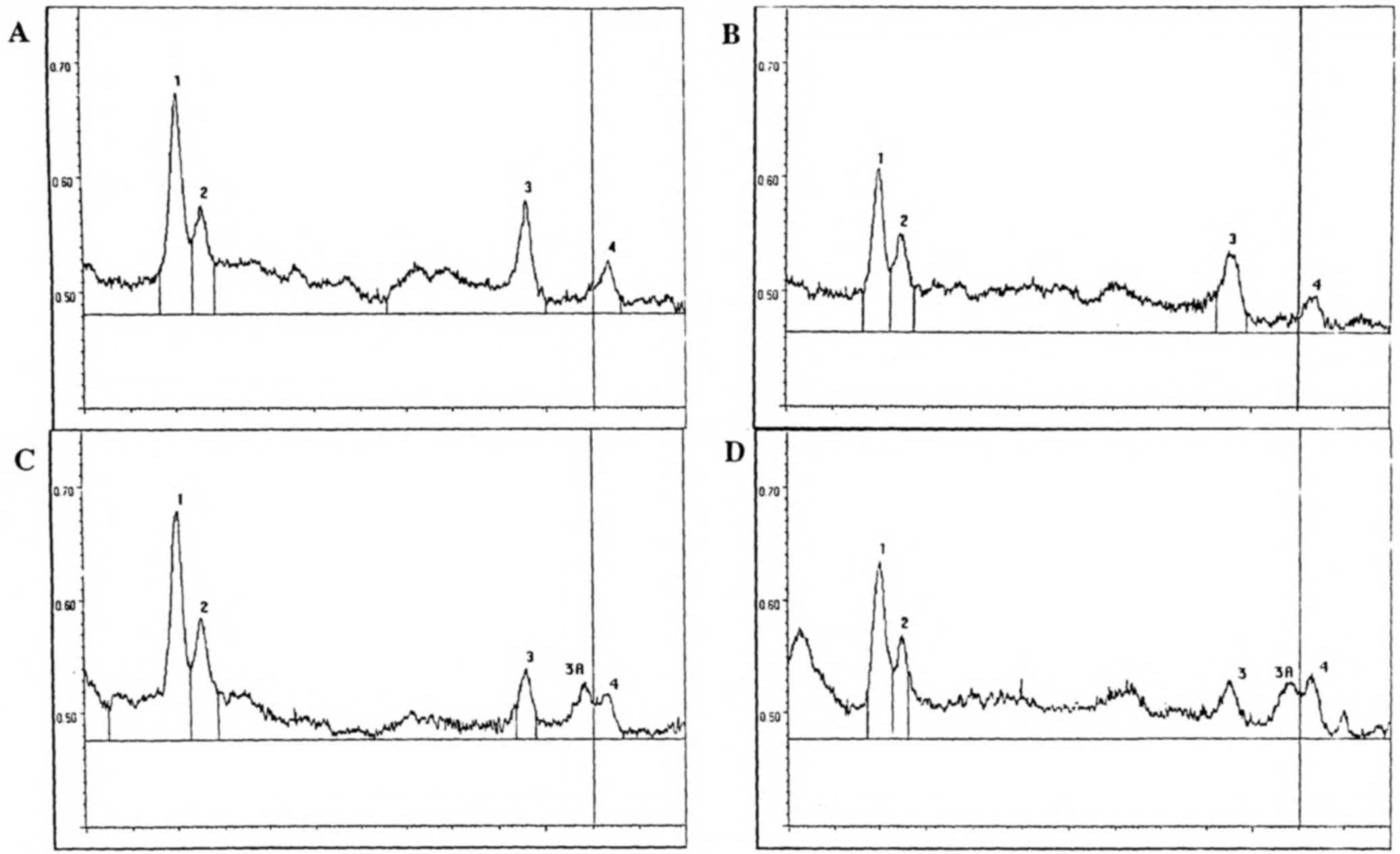


Figure 3.29 Densitometric profiles of the G7 *Pvu*II RFLP.

Lanes 1, 4, 5 and 7 from the Southern blot shown in figure 3.28 were scanned using the Molecular Dynamics Computing Densitometer and analysed using Molecular Dynamics Image Quant Software v3.0. Densitometric profiles were generated by calculating the total volume under the curve to eliminate variations within each track. The profiles from tracks 1 (**A**) and 5 (**B**) which do not contain the polymorphism and tracks 4 (C) and 7 (**D**) which contain the polymorphism are shown. Peaks 1, 2 and 4, which correspond to the invariant DNA fragments of 8.2, 6.4 and 1.6kb, are indicated. The polymorphic 2.1 and 1.7kb DNA fragments corresponds to peaks 3 and 3A in **C** and **D**. From the height of the peaks is can be seen that the presence of the polymorphic 1.7kb DNA fragment (3A) is accompanied by a reduction in the height of peak 3.

INDIVIDUALS	DR	HLA-B	HLA-C	HLA-A	1.7kb *Pvu*II
1	4,4	2,27	.	.	–
2	5,(6)	7,44	7	2,29	–
3	1,7	7,44	.	.	–
4	2,7	7,50	6,7	2,3	+
5	2	7	7	.	–
6	3,8	7,8	7	1,3	+
7	7,7	8,17	6	1,2	+
8	3,13	8,14	7,(8)	3,32	+
9	3,6	14	.	3,28	–
10	6	35	4	2	–
A1	3,13	8,14	.	1,28	+
A2	3,4	12,18	(5)	2	–
A3	4,11	18,45	6	29,31	–
A4	1,6	35,62	3,4	2,3	+
A5	2,7	45,60	6,7	23,26	+

Table 3.2 HLA-typed individuals analysed by RFLP analysis.
The HLA types of 15 unrelated individuals are shown. Serological specificities not known are shown by dots (.) and uncertain results are shown in brackets. Individuals numbered 1 to 10 are shown in figure 3.28. The results from individuals A1-5 are not shown.The presence of the 1.7kb *Pvu*II polymorphism in individuals is indicated by a cross (+).

counts as one haplotype).

To determine where the 1.7kb *Pvu*II polymorphism is derived from, the autoradiograph shown in figure 3.28 was scanned for densitometric analysis using the Molecular Dynamics Computing Densitometer and the data analysed using the Molecular Dynamics Image Quant Software v3.0 (Department of Plant Sciences, Oxford University). Densitometric profiles of the areas within tracks 1 & 5, which do not contain the polymorphism, and tracks 4 & 7, which contain the polymorphism were calculated (Figure 3.29). Comparison of the samples with and without the 1.7kb fragment shows that the intensity of the 2.1kb fragment (Peak 3) is reduced to approximately half when the 1.7kb fragment is present. This suggests that the polymorphic *Pvu*II site is located within the 2.1kb fragment. To map the location of the polymorphism within the cDNA, the 2.1kb *Pvu*II genomic DNA fragment was isolated from cosmid 9N, radiolabelled and hybridised to a Southern blot of the G7-F cDNA. The results show that the hybridisation is limited to a 700bp *Pvu*II fragment (Figure 3.30).

3.6.2 Characterisation of the G7 *Pvu*II polymorphism

The 700bp *Pvu*II fragment to which the polymorphic site was mapped [section 3.6.1] is flanked by oligonucleotides G7OL3 and G7OL4 [section 3.2.4]. This sequence was amplified by RT-PCR of total RNA isolated from HepG2, U937, HL60 and Daudi cell lines [section 3.4.1] and then digested with *Pvu*II. Analysis of the digests by agarose electrophoresis (Figure 3.31) revealed that the PCR products are cleaved at positions corresponding to nucleotides 1656 and 2380/2363 in the cDNA sequence [section 3.2.3] to generate a 724bp product. However, the products from the HepG2 and Daudi cell lines contained additional fragments of ~600bp and 130bp. Examination of the cDNA sequence ~130bp from the *Pvu*II sites revealed that there are three possible locations (2250, 2256 and 2259) (Figure 3.5) at which a single base change would create an additional *Pvu*II site. To identify which of these three positions is the polymorphic *Pvu*II site, the ~130bp product from Daudi cells was subcloned into *Sma*I cut M13mp10 [section 2.7] and sequenced using the M13 universal primer [section 2.9].

The sequence data shown in figure 3.32 shows that the polymorphic *Pvu*II site lies at nucleotide 2256 in the cDNA sequence, 124bp from the 3′ *Pvu*II site. To generate

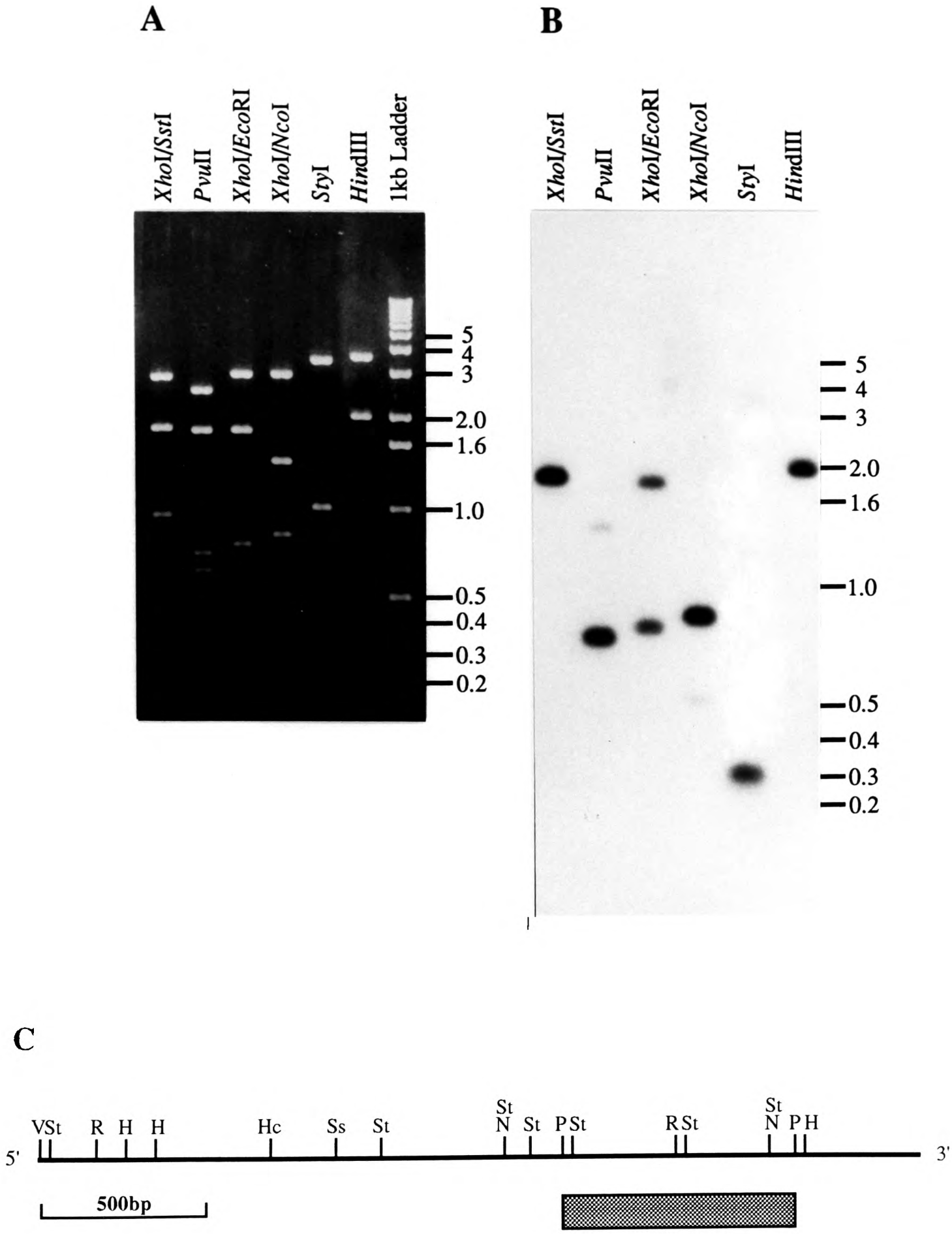


Figure 3.30 Mapping the G7 *Pvu*II RFLP within the G7 cDNA.
A cDNA clone G7-F was digested with the restriction enzymes shown and fractionated on a 0.8% (w/v) agarose gel. **B** Southern blot of G7-F DNA hybridised with the 2.1kb genomic DNA fragment containing the *Pvu*II RFLP. Sizes are in kb. **C** Restriction map of the G7-F cDNA clone containing sites for *Eco*RV (V), *Sty*I (St), *Eco*RI (R), *Hin*dIII (H), *Sst*I (Ss), *Nco*I(N) and *Pvu*II (P). The region hybridised by the 2.1kb genomic DNA fragment is shown as a shaded box.

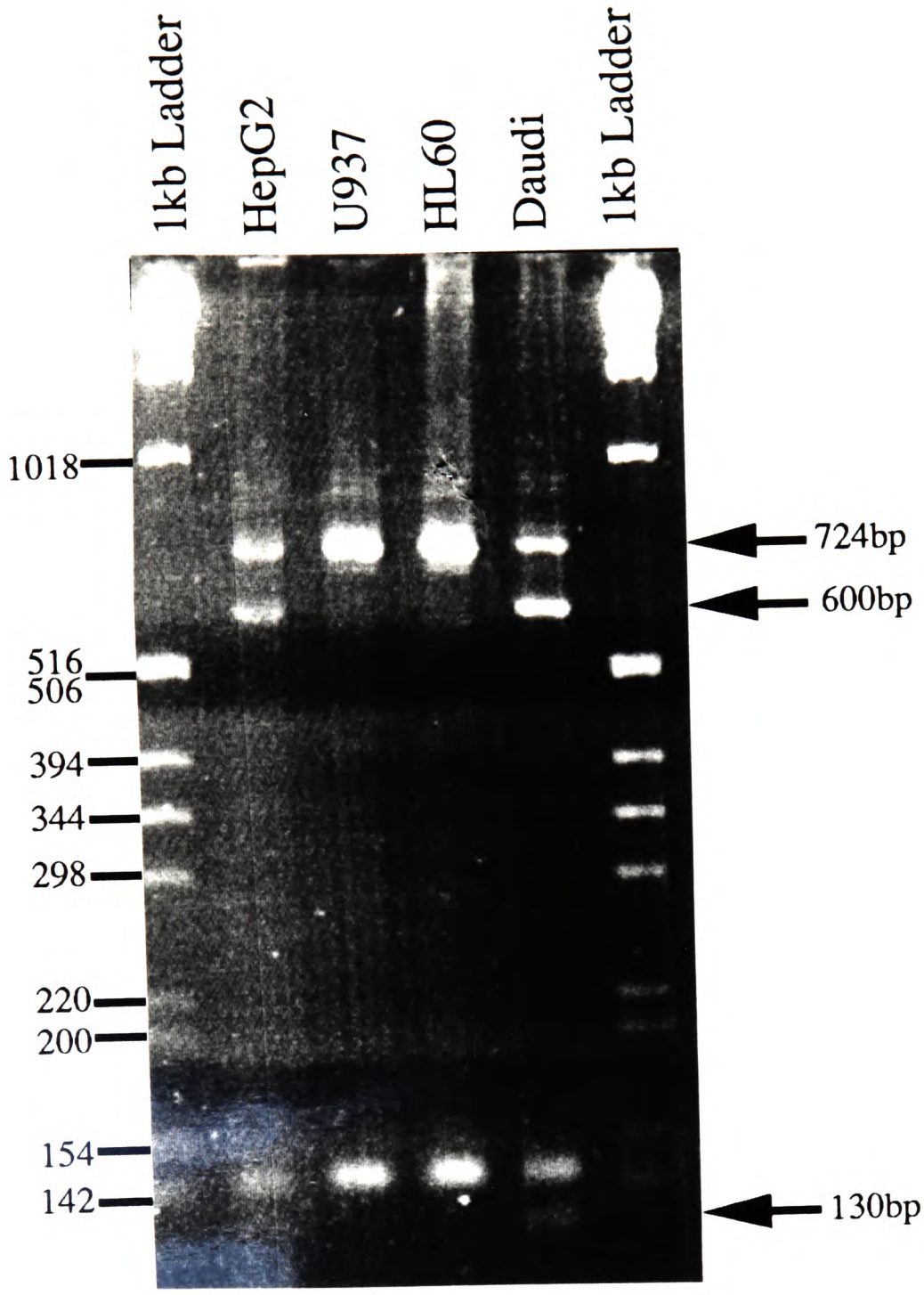


Figure 3.31 *Pvu*II polymorphism in G7 cDNA.

The G7 RT-PCR products amplified using oligonucleotide primers G7OL3 and G7OL4 from RNA isolated from the cell lines shown were digested with *Pvu*II and fractionated on a 2% (w/v) agarose gel. Samples from cell lines U937 and HL60 contain the expected 724bp fragment. The samples from HepG2 and Daudi contain additional fragments of ~130bp and ~600bp with reduced intensity of the 724bp band. This indicates that the samples from HepG2 and Daudi cells contain the *Pvu*II polymorphism within the amplified region ~130 bp from one of the *Pvu*II sites previously mapped in figure 3.30. Sizes are in bp.

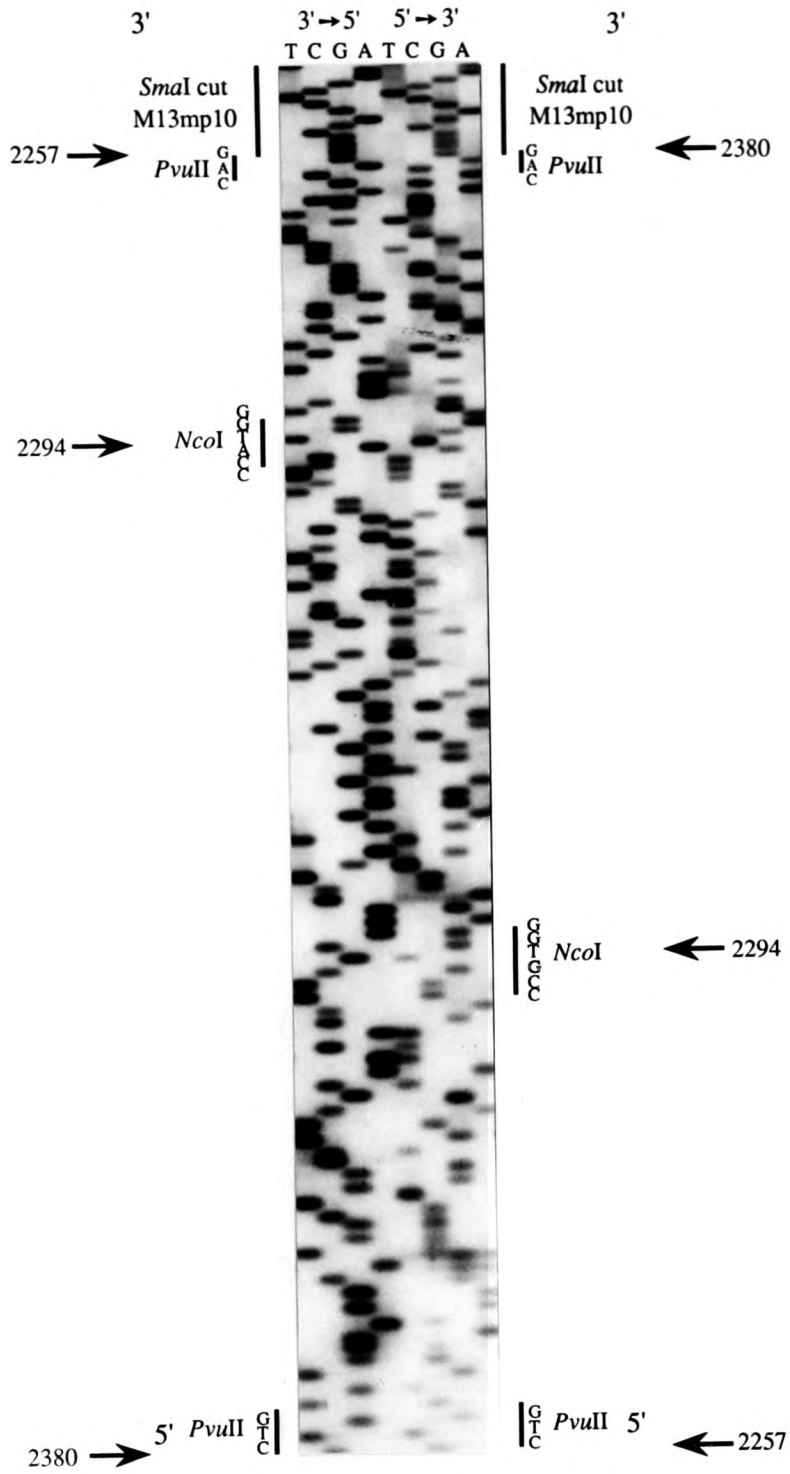


Figure 3.32 Localisation of the G7 *Pvu*II polymorphism.

The nucleotide sequence of the ~130bp *Pvu*II cDNA fragment from Daudi cells is shown in both orientations. The *Pvu*II site at position 2257 is not present in the cDNA clones derived from U937 cells. The positions of the *Pvu*II cleavage sites are shown. The recognition sequences of the *Nco*I site at position 2294 are shown, and an error in this sequences in one of the clones is shown in outline.

this site a single base change of A→G, or G→A, must have occurred. On the opposite strand this change is C→T which could have been caused by deamination of methylcytosine to thymidine. The A→G change causes a change of codon from ACT to GCT, which results in an amino acid change from threonine to alanine at residue 734, which is at the C-terminus, beyond motif III, of the predicted protein sequence.

3.6.3 Dinucleotide $(CA)_n$ repeats

$(CA)_n$ repeats, or (dC-dA).(dG-dT) repeats, are the second most abundantly occurring human microsatellite DNA sequences. These tandem, dinucleotide repeats are stable, can be highly polymorphic and tend to consist of from 6 to 30 repeating units (Weber 1990). As such, these sequences can readily be amplified by PCR and are ideal markers for constructing high resolution genetic maps in order to identify susceptibility loci involved in genetic diseases.

In the human MHC class III region twelve $(CA)_n$ microsatellite sequences have been identified (Hsieh 1992) and of these, four were isolated from cosmid 9N. One of them, 9N-I, is located in the second intron of G6, whilst the others mapped to DNA fragments corresponding to G7 (Hsieh 1992). Their locations in the G7 gene were identified by analysis of the sequences obtained in section 3.3.2.

The first repeat, 9N–IV, lies 900bp upstream of the 5′ end of the G7 cDNA and 39bp from an $(AT)_n$ repeat [section 3.3.2]; this is ~7kb away from the original position to which it was mapped. The allele sequenced possesses 19 dC-dA dinucleotides and 9 dA-dT dinucleotides. These compound repeats are normally very polymorphic, but 9N–IV could not be amplified by PCR. This is probably due to variation between the flanking sequences originally obtained (Hsieh 1992) and those shown in figure 3.15.

The next repeat, 9N–II, (Figure 3.15) lies within the third intron of G7, 61bp downstream of exon III at the 3′ end of an *Alu* repeat [section 3.3.3.2]. The allele sequenced contained 23 dC-dA dinucleotides. Seven alleles of 9N–II were detected at this locus and values of 72.3% heterozygosity and PIC (polymorphic information content)=0.705 were obtained from analysis of fifty unrelated individuals (Hsieh 1992).

The other repeat, 9N–III, (Figure 3.15) lies in intron 9 of G7, ~500bp downstream of exon IX at the 3′ end of an *Alu* repeat [section 3.3.3.2]. This is unusual as Beckman and Weber (1992) noted that only two out of fourteen $(CA)_n$ repeats were observed

adjacent to *Alu* elements. The particular allele sequenced at 9N–II contained 17 dC-dA dinucleotides, but 13 different alleles were detected with 80.6% heterozygosity and a PIC value of 0.797 from analysis of fifty unrelated individuals (Hsieh 1992).

3.7 EXPRESSION OF G7 POLYPEPTIDES

3.7.1 Expression of G7 fusion proteins

To confirm the expression of G7 at the protein level it is necessary to raise antibodies which detect G7 protein. In addition, antibodies that detect the G7 protein product could be used to determine the subcellular localisation of the protein and facilitate determination of its function. Towards this aim, antigens with the predicted protein sequence were prepared.

The *E.coli* expression vector, pGEX-2T (Smith and Johnson 1988) facilitates the expression and isolation of predicted protein sequences, such as that of G7, encoded in cDNA sequences. The pGEX-2T vector contains the coding sequence of the 26kDa, glutathione S-transferase of *Schistosoma japonicum* (Sj26) with the termination codon replaced by a polylinker sequence followed by TGA stop codons in all three reading frames such that cDNA sequences can be subcloned in all three reading frames. Induction by IPTG produces Sj26 as a fusion protein, with the predicted protein sequence encoded by the cDNA insert fused at the C-terminus. In addition, the pGEX-2T vector includes the thrombin cleavage site, Leu-Val-Pro-Arg-Gly-Ser, upstream of the cloning sites in the polylinker, which allows subsequent cleavage of the fusion protein to release the cDNA-encoded antigen.

Five different cDNA fragments from the full length G7 cDNA clone, G7-F, were subcloned, in phase with the Sj26 reading frame, into the *Sma*I site of the pGEX-2T polylinker [section 2.7]. These cDNA restriction fragments were: a 702bp *Eco*RV/*Hin*cII fragment, encoding amino acids 1–233; a 1420bp *Eco*RV/*Nco*I fragment, encoding amino acids 1–472; a 2663bp *Eco*RV/*Xba*I fragment, encoding the complete G7-F open reading frame; a 1962bp *Hin*cII/*Xba*I fragment, encoding amino acids 234–638 and an 826bp *Nco*I fragment, encoding amino acids 473–638 (Figure 3.33).

The plasmid constructs were transformed into *E.coli* NM554 competent cells

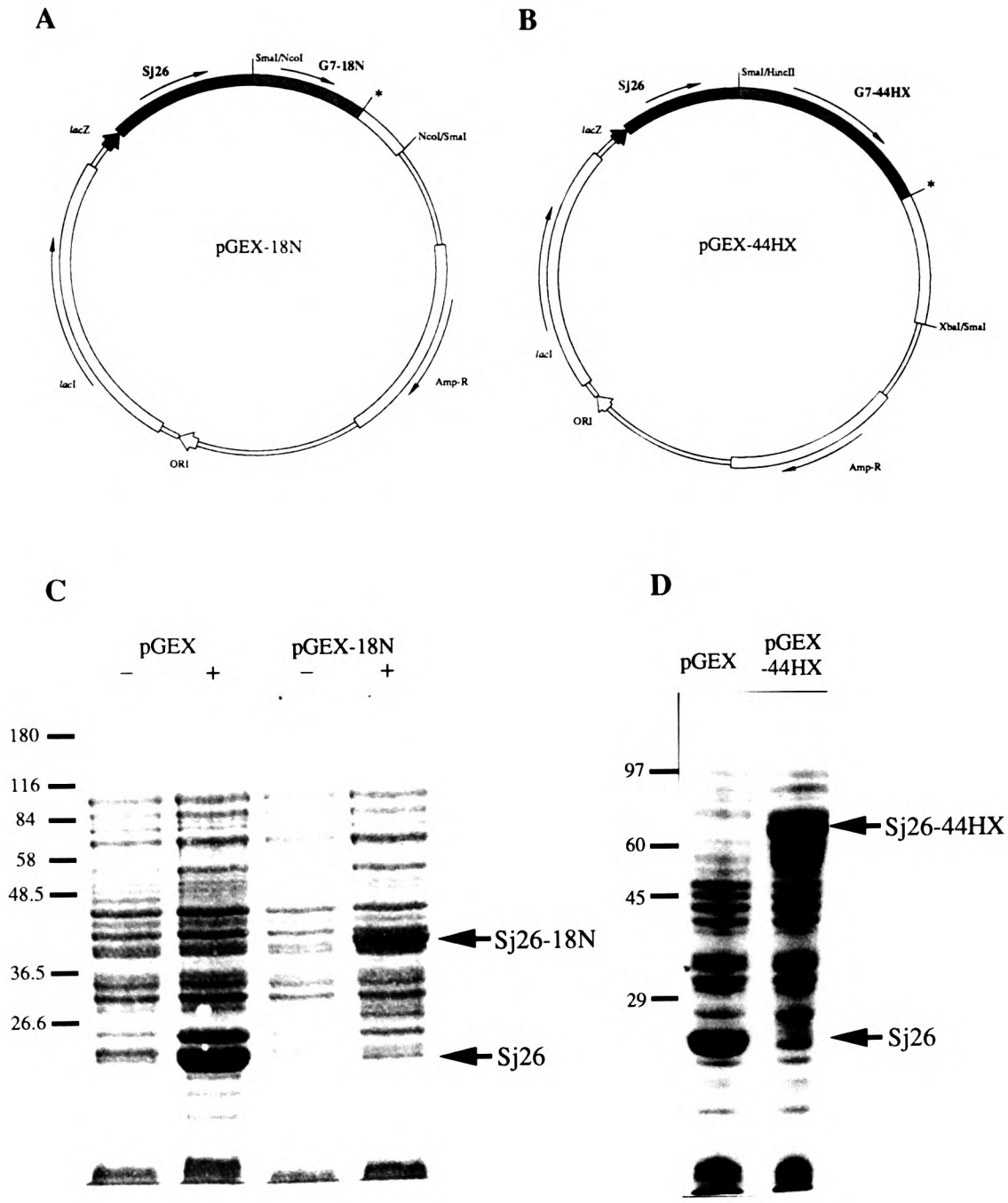


Figure 3.33 Expression of G7 fusion proteins.

Plasmid constructs pGEX-18N (**A**) and pGEX-44HX (**B**) containing G7 cDNA fragments (shaded regions) for expression as fusion proteins with Sj26 (filled regions). Arrows indicate the direction of transcription. Asterisks (*) show the positions of stop codons in the cDNA sequences. Induction of expression of the Sj26-18N (**C**) and Sj26-44HX (**D**) fusion proteins and Sj26 protein by IPTG. Total *E.coli* cell lysates fractionated on 10% (w/v) acrylamide gels by SDS-PAGE are shown following induction by IPTG (+) or in the absence of IPTG (-) (**C**). The bands corresponding to fusion proteins and Sj26 are shown. Sizes are in kDa.

[section 2.7.5] and expression of the G7 polypeptide-Sj26 fusion proteins was induced by the addition of IPTG [section 2.12.5]. Small samples of total cell lysates were analysed by SDS-PAGE on 10% (w/v) polyacrylamide gels to determine the level of expression of fusion protein by each clone compared with uninduced cells. Fusion proteins of 44kDa and 70kDa, were only induced by addition of IPTG to cultures of the clones containing the 1962bp *Hin*cII/*Xba*I and the 826bp *Nco*I fragments, respectively (Figure 3.33). The other clones did not express any fusion protein. These clones all encode the N-terminal sequence of G7, but there is no indication why this sequence should prevent expression of the fusion proteins.

3.7.2 Purification of G7 fusion proteins

The two fusion proteins expressed in *E.coli* [section 3.7.1] were isolated from cell lysates by absorption onto glutathione agarose beads [section 2.12.6] and at this stage the yields from the two clones were compared by SDS-PAGE. The larger, 70kDa, fusion protein, consisting of 26kDa Sj26 and ~44kDa generated from the 1962bp, *Hin*cII/*Xba*I G7 cDNA fragment, produced significantly poorer yields than the ~44kDa fusion protein consisting of 26kDa Sj26 and ~18kDa generated from the 826bp, *Nco*I G7 cDNA fragment. Consequently, the latter protein was used in subsequent experiments.

Although the purified ~44kDa fusion protein, was isolated on glutathione agarose beads [section 2.12.6] and was directly cleaved, on the beads, with thrombin [section 2.12.7], no G7 polypeptide was released into the supernatant. To confirm that the fusion protein had been cleaved by thrombin, samples of the supernatant and the beads were removed at 0, 30, 60 and 120 minutes after the addition of thrombin and analysed by SDS-PAGE on a 15% (w/v) polyacrylamide gel (Figure 3.34). The result indicated that the G7-Sj26 fusion protein is cleaved by thrombin, but that the 18kDa G7 polypeptide generated is apparently insoluble and remains bound to the beads.

Since the G7 polypeptide could not be isolated away from Sj26 by thrombin cleavage the intact fusion protein was eluted from the glutathione agarose beads using reduced glutathione [section 2.12.8] (Figure 3.35). However, using this procedure the yield of fusion protein was very poor (~20μg/litre of bacterial culture) making the preparation of sufficient antigen to raise antibodies an impractical task.

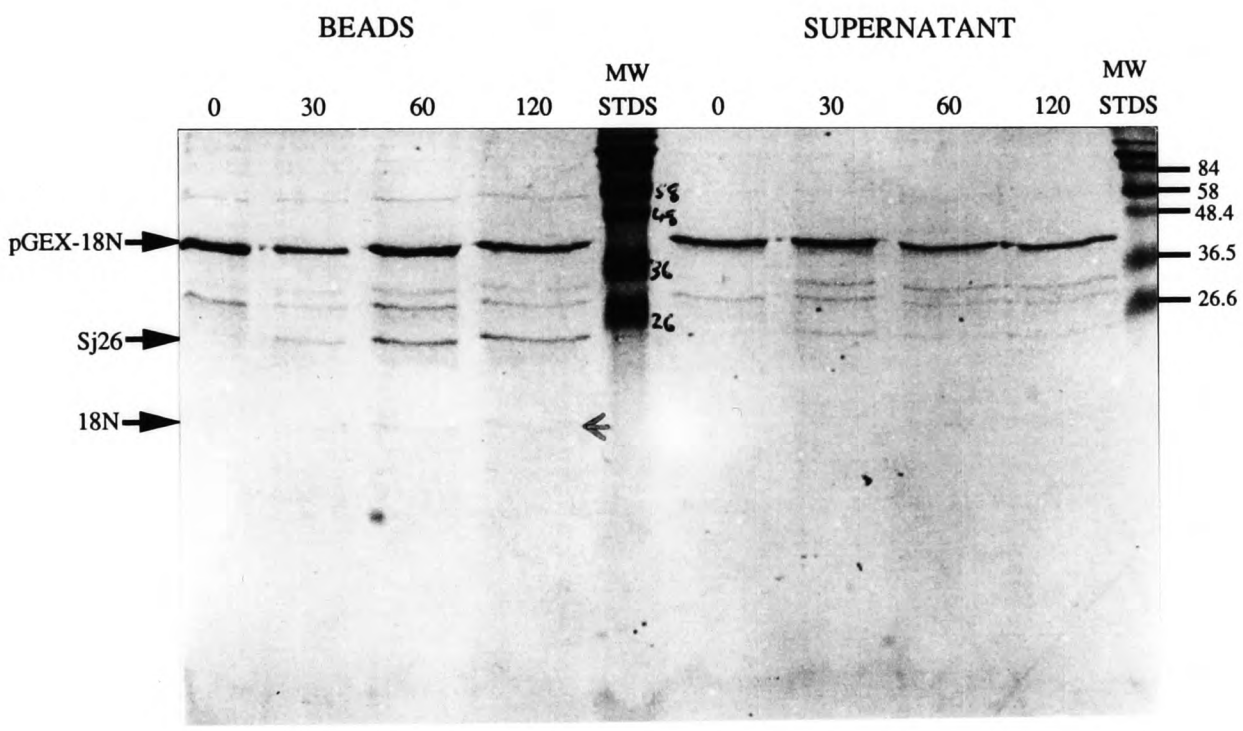


Figure 3.34 Cleavage of the pGEX-18N expression product with thrombin. The pGEX-18N expression product Sj26-18N isolated on glutathione agarose beads was cleaved with thrombin at 25°C. Aliquots were removed 0, 30, 60 and 120 min after addition of thrombin and the beads and supernatant were analysed separately by SDS-PAGE on a 15% (w/v) acrylamide gel. The pGEX-18N expression product and the cleavage products Sj26 and 18N are indicated. None of the 18N thrombin cleavage product is visible in the supernatant. Sizes are in kDa.

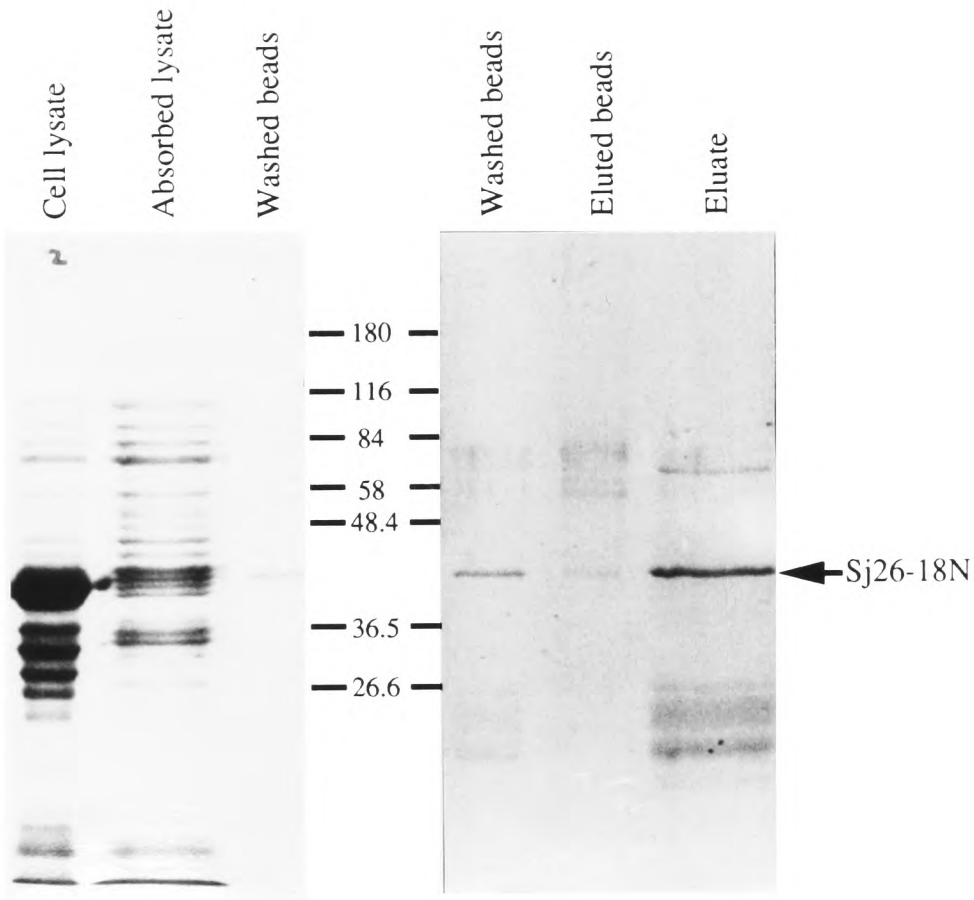


Figure 3.35 Isolation of the G7 Sj26-18N fusion protein.
The Sj26-18N fusion protein was isolated from the *E.coli* cell lysate by absorption onto glutathione agarose beads and the lysate was removed by washing. The Sj26-18N was eluted from the beads using reduced glutathione. The Sj26-18N fusion protein is shown. Sizes are in kDa.

Synthetic N- or C-terminal peptide haptens conjugated onto carrier proteins as antigens may facilitate future attempts to raise antibodies against G7.

3.8 DISCUSSION

3.8.1 The G7 gene

The G7 gene lies in the MHC class III region, 58kb telomeric of HSP70-Hom and contains 27 exons that span ~22kb of DNA. Although G7 was originally shown to express an ~3kb mRNA in monocytic and macrophage cells by Northern blot analysis (Sargent *et al.* 1989a), using the more sensitive RT-PCR technique, G7 expression has now been demonstrated in all the cell lines examined [section 3.4.1]. However, the level of G7 expression is low in most cell types and appears to be highest in monocytes and macrophages [section 3.4.1].

G7 contains three $(CA)_n$ repeats located within its introns and in the 5′ flanking region of the gene [section 3.6.3]. Two of these repeats have been characterised in detail and have been shown to be highly polymorphic and stably inherited (Hsieh 1992). As a result these repeats can be used for linkage studies to map disease susceptibility loci in the MHC class III region. The three $(CA)_n$ repeats lie within a region spanning 7kb, a much higher density of repeats than the 1 in 30kb predicted by Weber and May (1989).

Nucleotide analysis of the cDNA clones obtained in section 3.2.3 has demonstrated a large amount of variation in their splicing patterns, altering and terminating their ORFs. This observation may be due to the aberrant splicing of transcripts which can occur in cultured cell lines. For the G7 cDNAs, however, these variations have been explained on the basis of the exonic sequence data [section 3.3.4] and includes the alternative use of two exons that are both utilised in expression of G7 in cell lines [section 3.4.2]. This alternative splicing results in the production of two different putative protein products of 749 amino acids (M_r 83171), the product encoded by the higher abundance mRNA, and 788 amino acids (M_r 87624), the product encoded by the lower abundance mRNA.

3.8.2 Functions of the G7 proteins

The variation between the two predicted G7 protein sequences involves 5 amino acids at the C-terminus and an additional 37 amino acid residues in the less abundant form. No known function has been assigned to these sequences although the additional 37 amino acids do include three pairs of dibasic residues that may have some biological function.

The negatively charged, glutamic acid-rich, region of the G7 protein, which is also found in several chromatin-associated proteins, is shorter in G7 than in the other proteins, but the potential significance of this observation is unknown. However, the predominantly acidic sequences have been associated in CENP-B with the scleroderma variant CREST (Earnshaw *et al.* 1987) and in SLE with an autoantigen HMG-1 (Wen *et al.* 1989).

The most striking similarity detected within the putative G7 protein sequence is that which is shared with all the DNA mismatch repair proteins and homologues so far identified. All of these proteins have been shown to contain a type A nucleotide binding site for ATP or GTP, which does not conform to any of the previously defined subgroups (Saraste *et al.* 1990). ATPase activity has been detected that co-purified with *E.coli* MutS (Grilley *et al.* 1989) and site directed mutagenesis of the nucleotide binding site in the *S.typhimurium* MutS reduces this ATPase activity (Haber and Walker 1991).

The MutS and HexA proteins have been demonstrated to recognise mismatched DNA by direct interaction (Su and Modrich 1986) and by negative complementation (Haber *et al.* 1988). In addition, the mismatch binding activity of *S.typhimurium* MutS is affected by mutagenesis of the ATP binding domain (Haber and Walker 1991). MutS, therefore, is believed to bind to mismatched DNA and in an ATP-dependent step MutL and MutH bind so that MutH can nick the DNA at an unmethylated site initiating DNA repair (Modrich 1991).

Although G7 shows very strong similarity with the DNA mismatch repair proteins and their homologues and probably shares some structural similarity it does appear to lack two important motifs, a nuclear localisation signal and a consensus DNA binding site. However, the nuclear targeting signal consensus was only present in 56% of the nuclear proteins searched by Dingwall and Laskey (1991) and proteins <4.5nm in diameter or <60kDa, can move freely into and out of the

nucleus (von Heijne 1988). The apparent lack of DNA binding site consensus sequence (Churchill and Travers 1991) may be due to a different binding specificity in G7. For example, if G7 specifically recognises damaged DNA then the DNA structure, and therefore the binding specificity, would be different, requiring a different structure in the binding protein. The DUG protein is more similar to the bacterial proteins and is more likely to have the same function in man. An altered specificity in G7 for damaged or modified DNA is therefore more likely. This correlates with the ubiquitous expression of G7, but the elevated levels in monocytic and macrophage cell lines may be due to the increased possibility of DNA damage, and need for repair during hydroxyl free-radical production following phagocytosis.

3.8.3 Possible roles of G7 in disease

The original work of Teuscher *et al.*(1990) and Snoek *et al.* (1991) indicated that because of its location, between HSP70 and BAT5, G7 was a potential candidate gene for the *Orch-1* locus. However, subsequent work by Snoek *et al.* (1993), in addition to confirming the location of a gene homologous to G7 in the mouse *H-2S* region, has shown that the *Orch-1* locus lies centromeric of G7, so that G7 itself cannot be the *Orch-1* locus.

The putative G7 protein sequence [section 3.5.1] and the variation in G7 mRNA levels in different cell types [section 3.4.1] provide evidence that G7 may possibly play a role in other disease conditions. The negatively charged, glutamic acid rich region in the putative G7 protein sequence [section 3.5.1] may play a role in scleroderma or SLE disease aetiology [section 3.8.2], but the length of this region of negatively residues is shorter than that of the other disease associated proteins [section 3.5.1] and G7 is probably not involved in these disease states.

The significant similarity of the putative G7 protein sequence with DNA mismatch repair proteins and their homologues [section 3.5.2] indicates that G7 probably has a similar function. However, the absence of a DNA binding site similar to those found in the DNA mismatch repair proteins [section 3.5.2] may indicate that the DNA binding specificity, and subsequent protein function, is different. The putative G7 protein may play a role in other DNA repair mechanisms, but not mismatch repair.

Five main human DNA repair disorders have been identified: xeroderma

pigmentosum, Cockayne syndrome, Fanconi's anaemia, ataxia telangiectasia, Bloom's syndrome and '46 BR', identified by the cell line in which it was originally observed. These disorders are associated with twenty two complementation groups or genes, of which seven have so far been cloned (Hoeijmakers and Bootsma 1992). Patients that suffer from these different disorders, differ in their sensitivity to different agents *e.g.* ultraviolet rays and X-rays, and the proteins responsible may therefore have different specificities (Hoeijmakers and Bootsma 1992). G7 may be one of these loci responsible for one of the human DNA repair disorders.

The increased levels of G7 mRNA expression in monocytic cell lines [section 3.4.1] suggest a role for G7 in the immune system. This could be related to monocytic and, following differentiation, macrophage functions. The production of hydroxyl free-radicals, following phagocytosis, may cause some damage to DNA in the cell. However, G7 may protect monocytes and macrophages from excessive DNA damage and may also be important in other cells where this damage is not so acute. The deficiency of such an enzyme or the presence of polymorphic variants with lower activity could result in cell death, and the release of damaged DNA. The modified DNA fragments could then be recognised as antigens, leading to the production of anti-DNA antibodies such as those which have been identified in patients suffering with SLE.

CHAPTER 4

CHARACTERISATION OF G7b

4.1 INTRODUCTION

The identification of genes in the human MHC class III region, following the isolation of overlapping cosmids (Sargent *et al.* 1989a; Spies *et al.* 1989a,b, 1990; Kendall *et al.* 1990), has previously been carried out using two different approaches. Initially, Sargent *et al.* (1989a) mapped HTF-islands in this region, since these undermethylated, CpG-rich sequences have been shown to be frequently associated with the 5′ ends of housekeeping, and some tissue-specific, genes (Bird 1986, 1987; Gardiner-Garden and Frommer 1987). The identification of eleven HTF-islands in the interval from C2 to TNF-A led to the identification of fourteen novel genes; G1–G10, B144 and the HSP70 gene cluster (Sargent *et al.* 1989a,b; Milner and Campbell 1990). The second approach, adopted by Spies *et al.* (1989a,b, 1990) and Kendall *et al.* (1990), involved using whole cosmid inserts to screen cDNA libraries and this has led to the identification of seventeen novel genes between HLA-DRA and HLA-B; BAT 1–5 (Spies *et al.* 1989a), BAT 6–9 (Spies *et al.* 1989b), G11–18 (Kendall *et al.* 1990), X2–6 (Spies *et al.* 1990) and gene G9a, which lies between G9 and G10 (Milner and Campbell 1993).

Comparison of the locations of these novel genes mapped within the MHC class III region has revealed that neither method always detects all the genes present. Approximately one third of the genes within the MHC class III region are not HTF-island associated and low abundance or tissue specific mRNAs are more likely to be missed when large genomic DNA fragments are used as probes on specific cDNA libraries. As a result, BAT8 (G9a) was not originally detected by Sargent *et al.* (1989a), as it is not associated with an HTF-island, and G1, G4, G7, G8, G11, G12 and G17 were not detected by Spies *et al.* (1989a,b, 1990).

One of the potential HTF-islands identified by Sargent *et al.* (1989a), located ~15kb telomeric of HSP70-Hom, was not previously analysed for associated transcripts. This HTF-island is characterised by recognition sites for the restriction enzymes

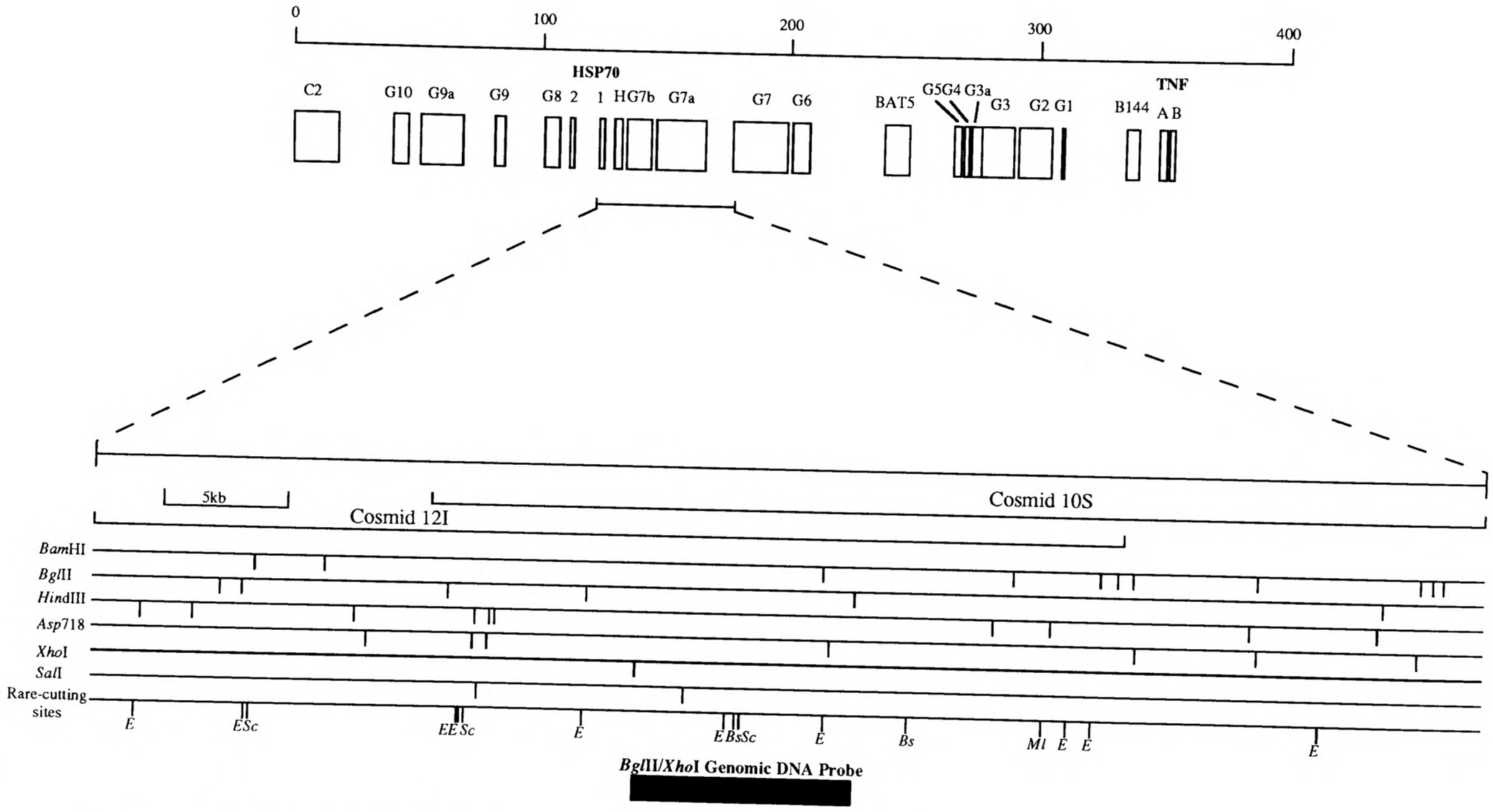


Figure 4.1 Molecular map illustrating the location of the G7b gene in the human MHC class III region.
The upper half of this figure shows the locations of the genes mapped between the C2 and TNF genes. The genes are shown as open boxes and the region corresponding to the genomic inserts of cosmids 12I and 10S is shown as a horizontal bar. The region contained in the cosmid inserts is expanded in the lower half of the figure. The positions of restriction sites in the region are indicated and the locations of restriction sites for rare-cutting enzymes *Eag*I (E), *Sac*II (Sc), *Bss*HII (Bs) and *Mlu*I (Ml) are shown together. The position of the ~9kb *Bgl*II/*Xho*I genomic DNA probe is depicted by a filled box.

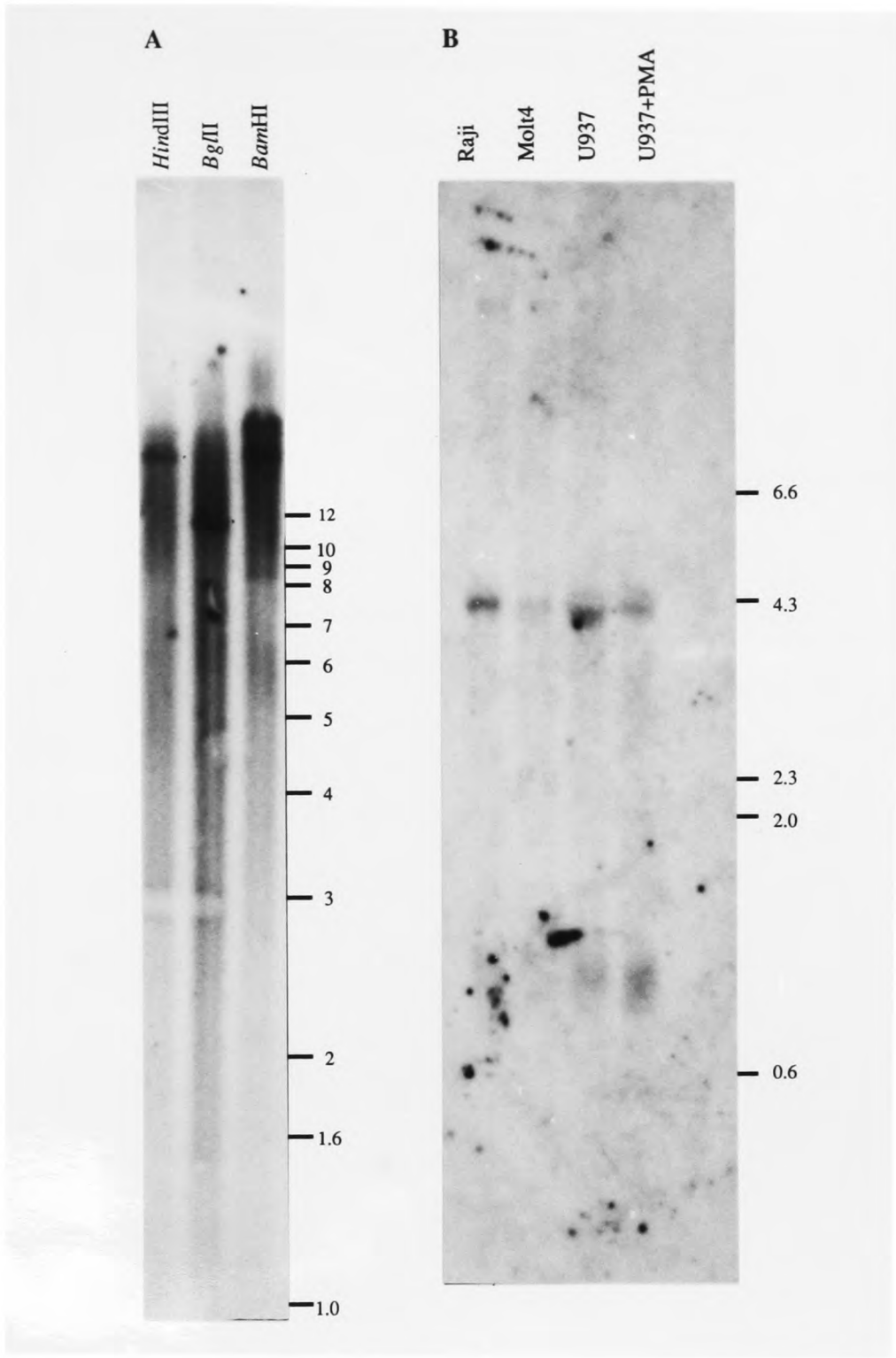


Figure 4.2 Detection of the G7a and G7b genes.
Southern blot of total genomic DNA from ICE5 (**A**) and Northern blot of total RNA isolated from the cell lines shown (**B**) hybridised with the ~9kb *Bgl*II/*Xho*I genomic DNA probe under competitive conditions. The probe contains single-copy sequences and detects two mRNA transcripts of ~0.9kb and ~4kb in size. Sizes are in kb.

*Bss*HII and *Sac*II, separated by ~500bp, that are cleaved in genomic DNA (Sargent *et al.* 1989a). An ~9kb *Bgl*II/*Xho*I DNA fragment containing the HTF-island was isolated from cosmid 10S (Figure 4.1) and hybridised, under competitive conditions by RD Campbell to a Southern blot of total genomic DNA digested with *Bam*HI, *Bgl*II and *Hin*dIII and to a Northern blot (Figure 4.2). These two results demonstrated that the probe contains single-copy sequences that hybridise to two mRNA species of ~0.9kb and ~4kb, present in all the cell lines examined (Figure 4.2). The two mRNA species have since been shown to be encoded by different genes. The ~4kb mRNA is transcribed from G7a, which encodes a valyl-tRNA synthetase (VARS2) (Hsieh and Campbell 1991). The other mRNA of ~0.9kb is transcribed from a novel gene, G7b. This novel gene has been sequenced at the cDNA and genomic DNA levels in order to define the protein product it encodes. This, in turn, may give some indication as to the function of the protein or to the possible role it may play in autoimmune disease aetiology. Antibodies have been raised against the predicted G7b protein product in an attempt to identify the protein by Western blot analysis.

4.2 ISOLATION AND CHARACTERISATION OF G7b cDNA CLONES

4.2.1 Isolation and identification of G7b cDNA clones

The ~9kb *Bgl*II/*Xho*I DNA fragment (Figure 4.1), isolated from cosmid 10S [section 2.4.3], which contains the potential HTF-island located ~15kb telomeric of HSP70–Hom, was used to screen a cDNA library constructed from mRNA derived from U937 cells (a gift from D Simmons, Institute of Molecular Medicine, Oxford), since both mRNA transcripts detected by this probe were present in that cell line (Figure 4.2).

A total of $5x10^5$ colonies were screened using the genomic DNA probe, competed with 125µg/ml placental DNA [section 2.8.1]. Sixty-nine, duplicating positives were identified and 18 out of 20 of these were successfully rescreened with the same probe [section 2.8.2]. The largest cDNA insert, the 3.6kb insert of clone G7a-1 (Hsieh and Campbell 1991), was excised using *Xho*I and isolated from LGT agarose [section 2.4.3]. This insert fragment was then used as a probe to screen the other cDNA inserts which had been excised from the vector, separated by agarose gel electrophoresis and

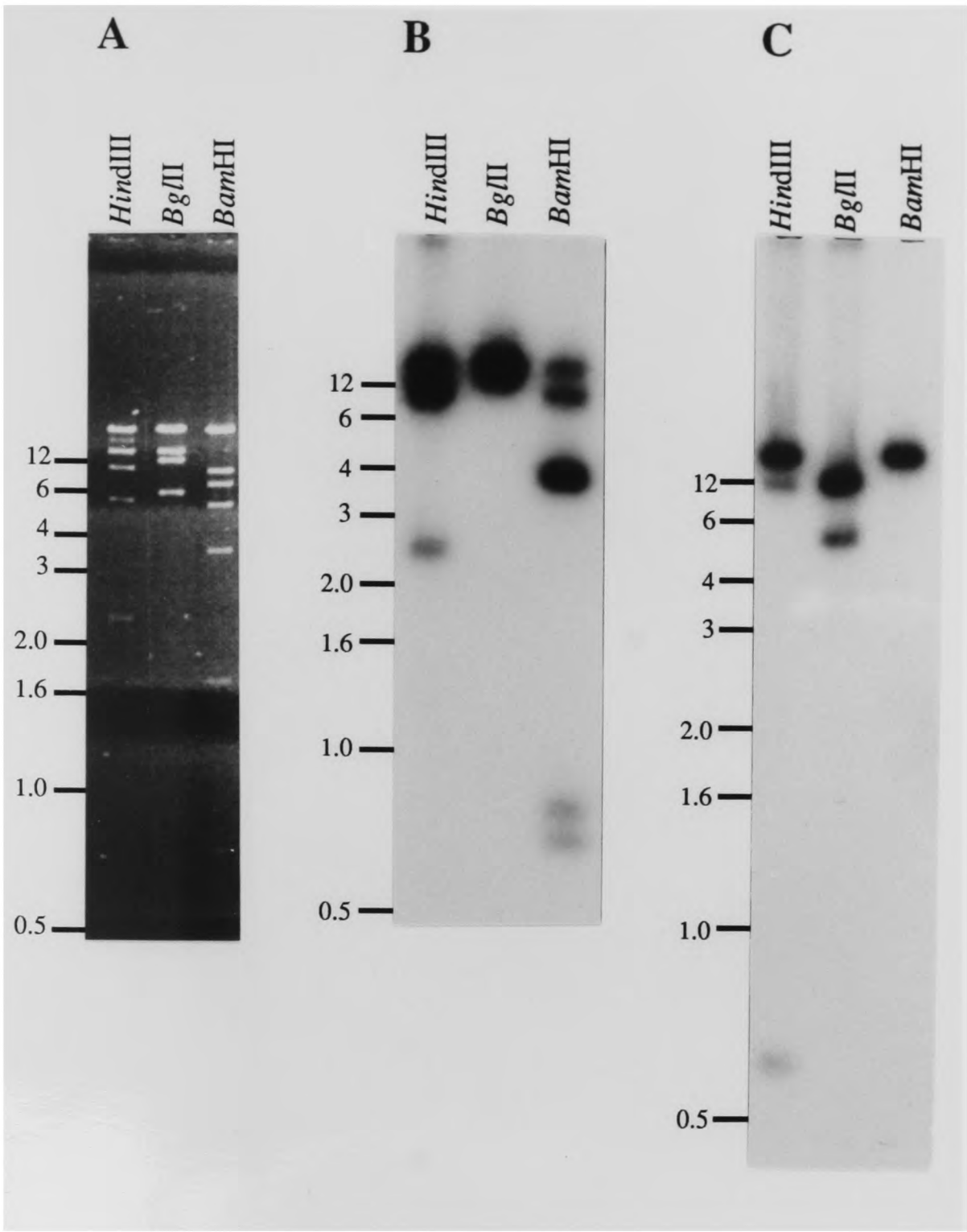
transferred onto nitrocellulose by Southern blotting [section 2.5.1]. As a result, fourteen clones, related to G7a-1, were identified. The largest cDNA insert from the remaining three clones, the ~0.9kb insert of clone G7b.1-7, was excised using *Xho*I, isolated from LGT agarose [section 2.4.3] and similarly used to show that these three clones were also related. The G7b.1-7 probe was also used to directly screen the remaining 49 positives not previously rescreened [section 2.8.2] and to screen the library again for additional cDNA clones [section 3.2.1]. Two G7b clones were identified from the 49 remaining positives. In addition, two new clones, G7b.2-a and G7b.3-a, were also isolated from the cDNA library and rescreened [section 2.8.2]. The insert sizes of the seven positive clones ranged from 0.6kb to 0.9kb.

4.2.2 Mapping the location of the cDNA inserts in cosmid 10S

To confirm that the two sets of cDNA clones are both encoded within cosmid 10S, the G7a-1 and G7b.1-7 inserts were used to probe Southern blots of cosmid 10S DNA digested with *Hin*dIII, *Bgl*II and *Bam*HI (Figure 4.3). The results demonstrated that the two cDNA inserts hybridise to two sets of DNA fragments from cosmid 10S that overlap in the region from which the original genomic DNA probe was derived (Figure 4.3). Subsequent mapping, using G7a cDNA inserts, revealed that the 5′ end of G7a lies within the HTF-island and that the gene extends in a telomeric direction towards G7 (Hsieh and Campbell 1991). Additional mapping revealed that G7b lies centromeric of the HTF-island [section 4.3.1].

4.2.3 Southern and Northern blot analysis

In order to determine the copy number, and to check the authenticity of the G7b.1-7 cDNA insert, the 0.9kb cDNA probe [section 4.2.1] was hybridised to Southern blots of cloned DNA from cosmid 12I (Figure 4.1), which extends centromeric of cosmid 10S, and total genomic DNA from the HLA homozygous cell line, ICE5, from which the cosmid library was originally constructed (Sargent *et al.* 1989a). The DNA was digested with the restriction enzymes *Bam*HI, *Bgl*II, *Hin*dIII, *Bst*EII, *Pst*I, *Sst*I, *Eco*RI, *Rsa*I, *Nco*I and *Pvu*II. The ~20kb *Bam*HI, ~11kb and 6kb *Bgl*II and ~20kb, 5kb and 0.5kb *Hin*dIII DNA fragments from cosmid 12I (Figure 4.4 A) detected by the probe were in agreement with those obtained by Sargent *et al.* (1989a; figure 4.1). However, upon probing the total genomic DNA Southern blot, the same restriction pattern



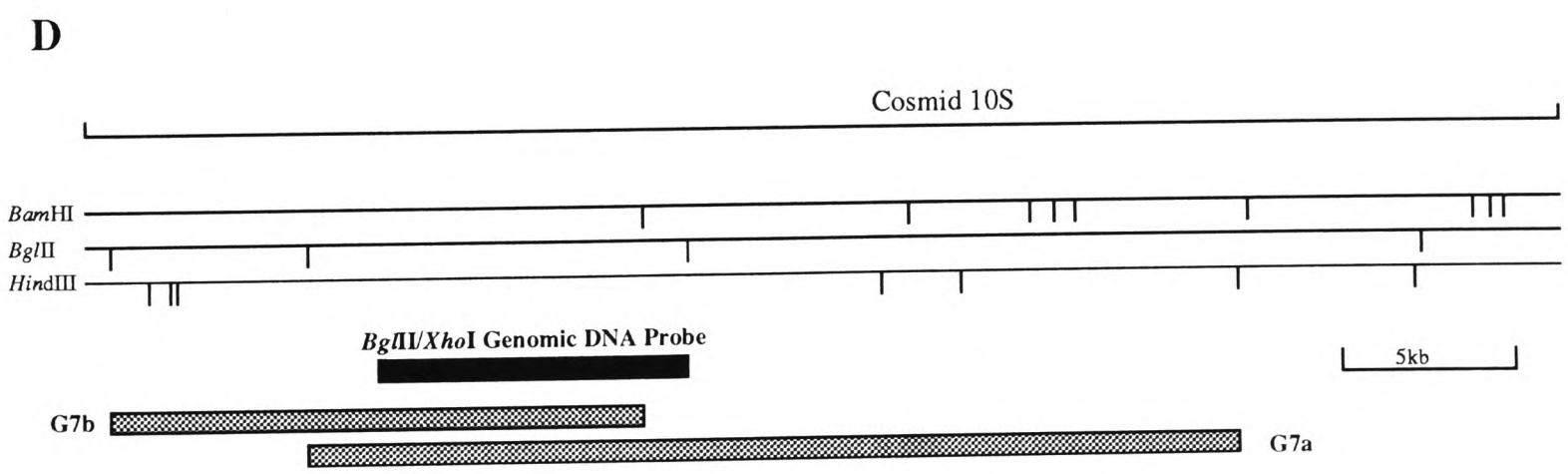


Figure 4.3 Mapping the G7a and G7b cDNA clones within cosmid 10S.
A Cosmid 10S DNA digested with restriction enzymes and fractionated on a 0.8% (w/v) agarose gel. Southern blots of digested cosmid 10S DNA hybridised with the G7a -1 (**B**) and G7b.1-7 (**C**) cDNA clones. Sizes are in kb. **D** The extent of the cosmid 10S insert and the positions of restriction sites in this region are shown. The ~9kb *Bgl*II/*Xho*I genomic DNA fragment is depicted as a filled box and the regions hybridised by the cDNA inserts are depicted by shaded boxes.

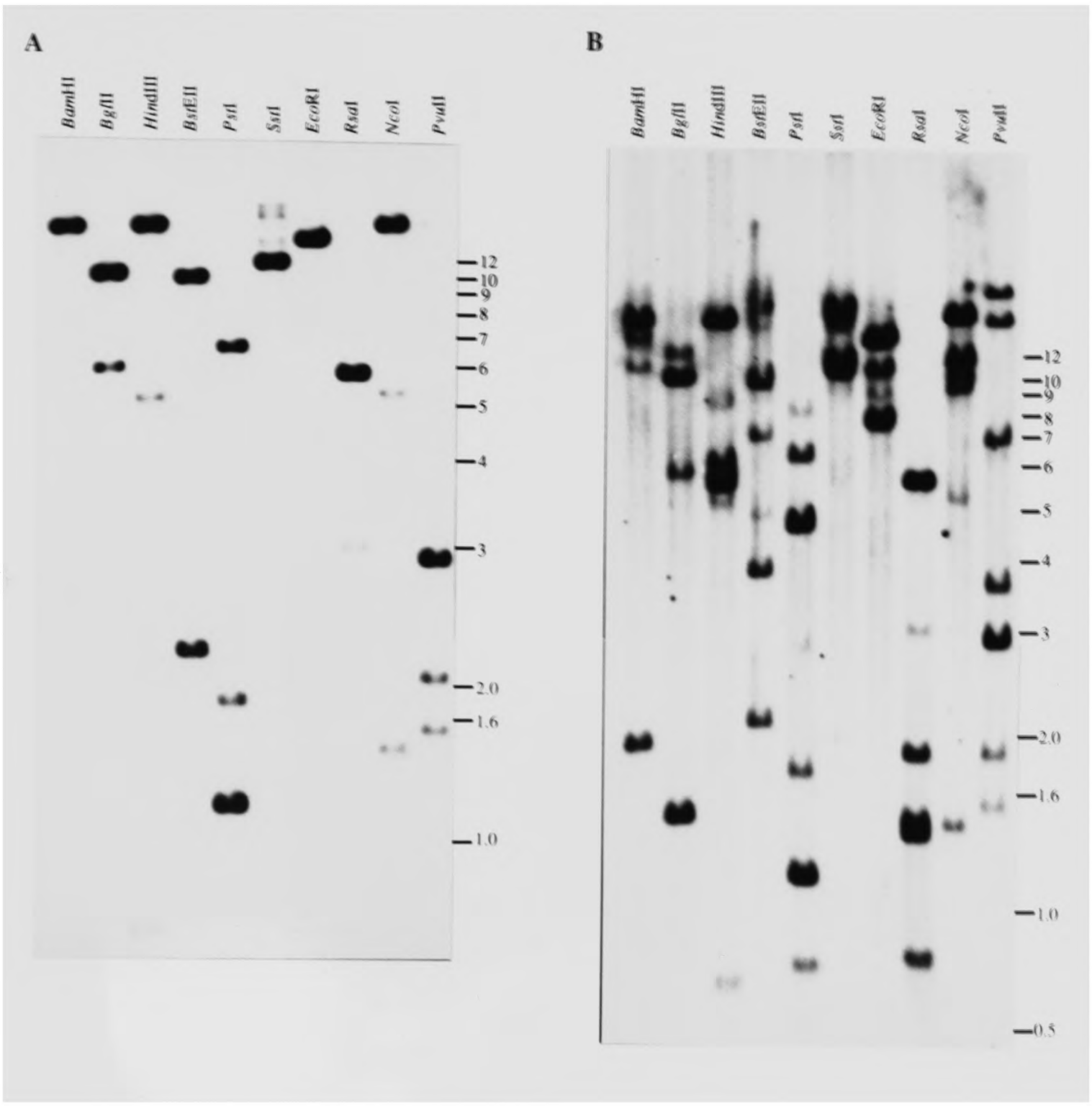


Figure 4.4 Southern blot analysis of G7b.

Southern blots of cosmid 12I DNA (**A**) and ICE5 total genomic DNA (**B**) hybridised with the G7b.1-7 cDNA insert. Sizes are in kb

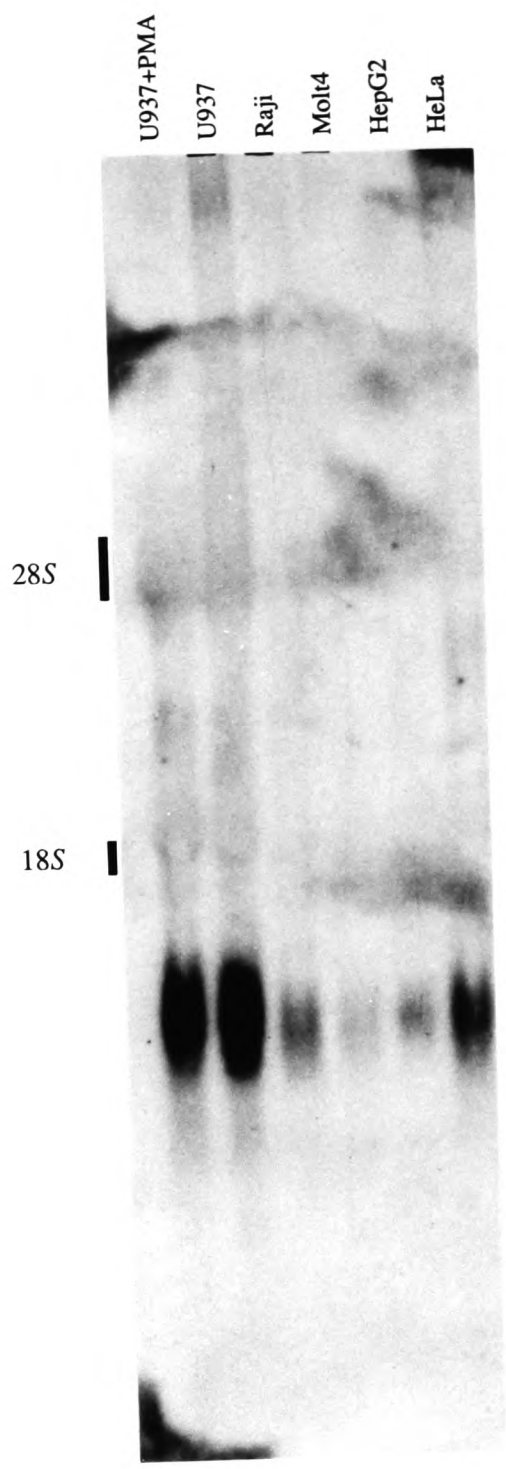


Figure 4.5 Northern blot analysis of G7b.
Total RNA (20μg) isolated from the cell lines shown was fractionated on a 1% (w/v) agarose - formaldehyde denaturing gel and transferred onto nitrocellulose. The Northern blot was hybridised with the G7b.1–7 cDNA insert. The positions of migration of the 18*S* and 28*S* rRNA are shown.

was observed except that additional bands were also present in all the digested samples (Figure 4.4 B) indicating that the cDNA insert does not correspond to a single-copy sequence. The number of additional bands varies from three in *Nco*I and *Sst*I digested DNA, to five in *Bst*EII digested DNA, suggesting that there may be up to three homologous sequences elsewhere in the human genome.

The G7b.1-7 probe was also hybridised to a Northern blot of total RNA prepared from cultured cell lines representing monocytes (U937), macrophages (U937 cells stimulated with Phorbol 12-myristate 13-acetate [PMA]), B lymphocytes (Raji), T lymphocytes (Molt4), liver (HepG2) and epithelium (HeLa). A mRNA of ~0.9kb was prominent in U937 cells and U937 cells stimulated with PMA (Figure 4.5), although it was also detected in the other cell lines at a lower level. However, on the original Northern blot, probed by the single copy genomic DNA probe (Figure 4.2 B), the level of this mRNA species was the same in all the cell lines and it is therefore possible that the increase is due to cross hybridisation to another mRNA species encoded by one of the other homologous sequences elsewhere in the human genome.

4.2.4 Nucleotide sequence analysis of the G7b cDNA clones

The G7b.1-7 clone was initially digested with a range of restriction enzymes to identify those that cleave within the insert. Only *Pst*I and *Sal*I were found to cut within the insert and both sites were found to lie towards the middle of the sequence (Figure 4.6). Using this information, the G7b.1-7 cDNA insert was excised by digestion with *Xho*I and cleaved using *Sal*I, and the two resulting, ~450bp, DNA fragments were separated from the CDM8 vector on a 3.5% (w/v) polyacrylamide gel [section 2.4.5] and recovered as described in section 2.4.6. The isolated DNA fragments were subcloned into *Sma*I cut M13mp10 [section 2.7] and sequenced using the M13 universal primer [section 2.9.3]. The sequence data obtained from subclones containing the fragments in different orientations were shown to overlap in the middle (Figure 4.6) so that the complete sequence of both fragments was obtained. Analysis of this sequence identified other restriction sites for *Eag*I, *Nci*I, *Sac*II, *Pvu*II, *Bst*EII, *Sty*I and *Bst*NI within the original insert of G7b.1-7 (Figure 4.6). To confirm this sequence, the complete sequence of both strands of DNA was obtained by excising the G7b.1-7 insert using *Xho*I and then cleaving with *Nci*I and *Bst*EII to

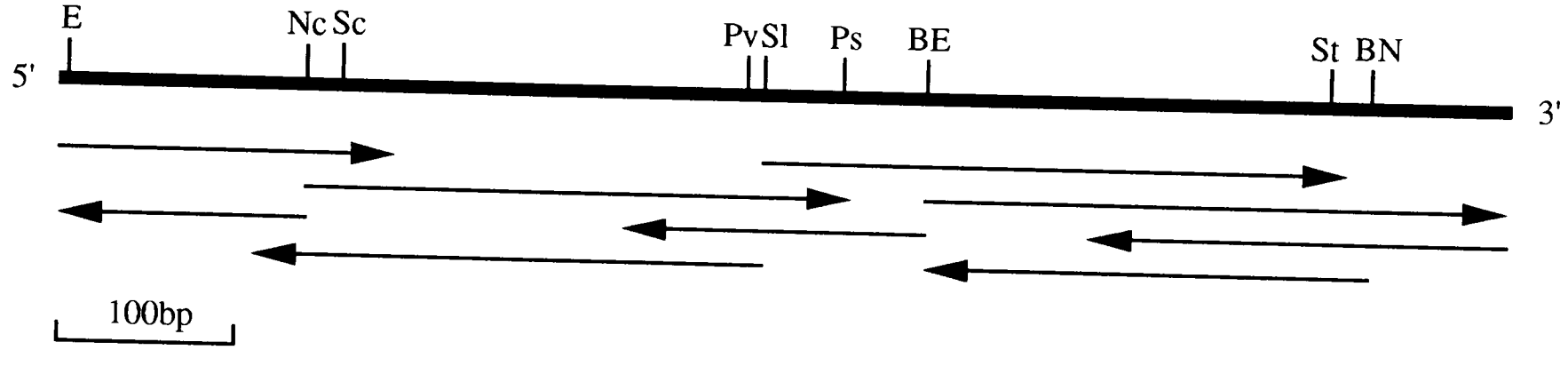


Figure 4.6 Sequencing strategy used for the G7b.1-7 clone.
The cDNA insert is shown as a horizontal bar with the positions of restriction sites for the enzymes *Eag*I (E), *Nci*I (Nc), *Sac*II (Sc), *Pvu*II (Pv), *Sal*I (Sl), *Pst*I (Ps), *Bst*EII (BE), *Sty*I (St) and *Bst*NI (BN) indicated. The arrows beneath the restriction map indicate the orientation and extent of the sequence obtained from each subclone.

generate DNA fragments of 140, 355 and 320bp. These DNA fragments were fractionated on a 3.5% (w/v) polyacrylamide gel [section 2.4.5], recovered [section 2.4.6] and subcloned into *Sma*I cut M13mp10 [section 2.7]. The subclones were sequenced using the M13 universal primer [section 2.9] and this completed the sequence of the forward strand (Figure 4.6). However, the presence of a 22bp poly(A) tail prevented the sequencing reactions at the 3′ end from reaching the *Bst*EII site. To complete the sequence, a 250bp *Bst*EII/*Bst*NI DNA fragment from clone G7b.1-7 was isolated, subcloned and sequenced (Figure 4.6).

The G7b.1-7 cDNA sequence is 837bp long (Figure 4.7), including a 22bp poly(A) tail which lies 12bp downstream of a polyadenylation signal, AATAAA, indicating that this corresponds to the 3′ end of G7b. To confirm whether any of the other cDNA clones that had been isolated contained inserts that extended further than G7b.1-7 at the 5′ end, all the remaining positive clones [section 4.2.1] were digested with *Eag*I and *Sal*I in combination with *Xho*I. Analysis of the digestion products by agarose gel electrophoresis revealed that clones G7b.2-a and G7b.3-a extended further than G7b.1–7. As a result, these two clones were digested with *SalI*/*Xho*I and the insert DNA fragments isolated, subcloned and sequenced, as described for G7b.1-7. Comparison of the G7b.1-7, G7b.2-a and G7b.3-a cDNA sequences (Figure 4.7) revealed that clone G7b.3-a extends G7b.1-7 at the 5′ end by 49bp to yield 864bp of sequence.

The 3′ end of the G7b.2-a and G7b.3-a sequences contain poly(A) tails of 21bp and 46bp, respectively. In addition, clone G7b.3-a includes a 31bp deletion, 240bp from the 5′ end of G7b.1-7 (Figure 4.7), which was later shown to correspond to one of the exons of G7b [section 4.3.3]. This deletion produces a frameshift in the coding sequence which alters the putative G7b protein product.

4.2.5 Mapping the 5′ end of G7b mRNA by RNase mapping

Genomic DNA sequence analysis of G7b [section 4.3.2] revealed that the most 5′ end of the cDNA sequence lies within a 1.4kb *Bgl*II/*Sal*I DNA fragment. A 520bp *Sty*I/*Pvu*II DNA fragment, from this region, containing exon I of G7b and 300bp of upstream sequence (Figure 4.14), was subcloned into *Hin*cII cut pGEM-3Zf. This clone was used to determine the extent of the 5′ exon by RNase mapping, as described in section 2.10.2. An antisense RNA probe was transcribed, from the

```
G7B.3-a  GGCTTCCTCCTCCCCTCTAC TCTACCCAGCTTGCGCTCCC CAGCCGCAAGTCGGCCGCGC TTTGCCCGTCAGCGCTTGGA GCTTTCTGCGTCGCTTCCCG  100
G7B.1-7                                          GTCGGCCGCGC TTTGCCCGTCAGCGCTTGGA GCTTTCTGCGTCGCTTCCCG  100
G7B.2-a                                                      CCCGTCAGCGCTTGGA GCTTTCTGCGTCGCTTCCCG  100

G7B.3-a  CTGCGCCTGCGCGGTCCCGC CTCGCCCCACGCGCGGGCTC GCGCTTCGGTTTCCCCAGAC CTGCTCGCAGCACCCTGCTG TCTTCCCGGTCCGGCCCGCT  200
G7B.1-7  CTGCGCCTGCGCGGTCCCGC CTCGCCCCACGCGCGGGCTC GCGCTTCGGTTTCCCCAGAC CTGCTCGCAGCACCCTGCTG TCTTCCCGGTCCGGCCCGCT  200
G7B.2-a  CTGCGCCTGCGCGGTCCCGC CTCGCCCCACGCGCGGGCTC GCGCTTCGGTTTCCCCAGAC CTGCTCGCAGCACCCTGCTG TCTTCCCGGTCCGGCCCGCT  200

G7B.3-a  GCCCGCGGCGCCAGCACCAT GCTCTTCTATTCTTTTTTCA AGTCCCTTGTGGGCAAGGAT GTGGTCGTGGAACTAAAGAA TGACCTGAG-----------  300
G7B.1-7  GCCCGCGGCGCCAGCACCAT GCTCTTCTATTCTTTTTTCA AGTCCCTTGTGGGCAAGGAT GTGGTCGTGGAACTAAAGAA TGACCTGAGCATCTGTGGAA  300
G7B.2-a  GCCCGCGGCGCCAGCACCAT GCTCTTCTATTCTTTTTTCA AGTCCCTTGTGGGCAAGGAT GTGGTCGTGGAACTAAAGAA TGACCTGAGCATCTGTGGAA  300

G7B.3-a  -------------------- TATCTCAACATCAAACTAAC TGACATCAGTGTCACAGACC CTGAGAAATACCCTCACATGT TATCAGTGAAGAACTGCTT  400
G7B.1-7  CCCTCCATTCTGTGGATCAG TATCTCAACATCAAACTAAC TGACATCAGTGTCACAGACC CTGAGAAATACCCTCACATGT TATCAGTGAAGAACTGCTT  400
G7B.2-a  CCCTCCATTCTGTGGATCAG TATCTCAACATCAAACTAAC TGACATCAGTGTCACAGACC CTGAGAAATACCCTCACATGT TATCAGTGAAGAACTGCTT  400

G7B.3-a  CATTCGGGGCTCAGTGGTCC GATACGTGCAGCTGCCAGCA GATGAGGTCGACACACAGTT GCTACAGGATGCGGCAAGGA AGGAAGCCCTGCAGCAGAAA  500
G7B.1-7  CATTCGGGGCTCAGTGGTCC GATACGTGCAGCTGCCAGCA GATGAGGTCGACACACAGTT GCTACAGGATGCGGCAAGGA AGGAAGCCCTGCAGCAGAAA  500
G7B.2-a  CATTCGGGGCTCAGTGGTCC GATACGTGCAGCTGCCAGCA GATGAGGTCGACACACAGTT GCTACAGGATGCGGCAAGGA AGGAAGCCCTGCAGCAGAAA  500

G7B.3-a  CAGTGATGGCTCCTCCTCCT CTTCCCCTCCCTCTTTCATT GGTGACCCATAACCCCAAGT CCCAGCCCAGAACCCCTAAC CCCCAATACTTGAAGGGGTT  600
G7B.1-7  CAGTGATGGCTCCTCCTCCT CTTCCCCTCCCTCTTTCATT GGTGACCCATAACCCCAAGT CCCAGCCCAGAACCCCTAAC CCCCAATACTTGAAGGGGTT  600
G7B.2-a  CAGTGATGGCTCCTCCTCCT CTTCCCCTCCCTCTTTCATT GGTGACCCATAACCCCAAGT CCCAGCCCAGAACCCCTAAC CCCCAATACTTGAAGGGGTT  600

G7B.3-a  TTGTTTTTTTACTAATGATG GTTTTGTGGGTTTTTTTTAA GGGATGAGTGGATGAGAGGA GTAATAGGGAACAGCTATCC TCTCTTGAGAAGGGGAGGAT  700
G7B.1-7  TTGTTTTTTTACTAATGATG GTTTTGTGGGTTTTTTTTAA GGGATGAGTGGATGAGAGGA GTAATAGGGAACAGCTATCC TCTCTTGAGAAGGGGAGGAT  700
G7B.2-a  TTGTTTTTTTACTAATGATG GTTTTGTGGGTTTTTTTTAA GGGATGAGTGGATGAGAGGA GTAATAGGGAACAGCTATCC TCTCTTGAGAAGGGGAGGAT  700

G7B.3-a  AAGTAGGCTGGGAAACTTCA AAGCCTTCCCAGTCCCCAGC ACCTGCCTTTCTCACTACTT CTCTGGAGATGGTAGGAGAG TTTCCTAGGTCTTTCCAGGG  800
G7B.1-7  AAGTAGGCTGGGAAACTTCA AAGCCTTCCCAGTCCCCAGC ACCTGCCTTTCTCACTACTT CTCTGGAGATGGTAGGAGAG TTTCCTAGGTCTTTCCAGGG  800
G7B.2-a  AAGTAGGCTGGGAAACTTCA AAGCCTTCCCAGTCCCCAGC ACCTGCCTTTCTCACTACTT CTCTGGAGATGGTAGGAGAG TTTCCTAGGTCTTTCCAGGG  800

G7B.3-a  CAGCATGTGATTCATTTGGG GATGGAAGGAATCTGTCCCG CATCGGGAATAAAATTTATG ATGC (A)22
G7B.1-7  CAGCATGTGATTCATTTGGG GATGGAAGGAATCTGTCCCG CATCGGGAATAAAATTTATG ATGC (A)21
G7B.2-a  CAGCATGTGATTCATTTGGG GATGGAAGGAATCTGTCCCG CATCGGGAATAAAATTTATG ATGC (A)46
```

Figure 4.7 Alignment of G7b cDNA insert nucleotide sequences.

The G7b cDNA insert sequences are aligned and the sequence deleted in clone G7b.3-a is denoted by dashes. The polyadenylation signal AATAAA at the 3' end is underlined.

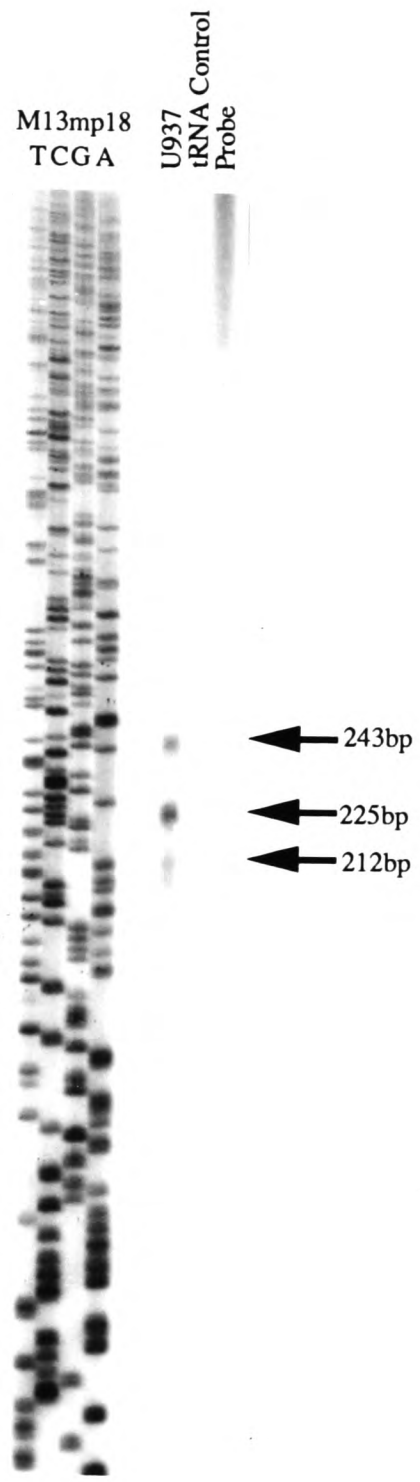


Figure 4.8 RNase mapping analysis of the 5' end of G7b mRNA.
An antisense RNA probe was transcribed from a 520bp *StyI*/*Pvu*II genomic DNA fragment which contained exon I of G7b and ~300bp of upstream sequence using T7 RNA polymerase.The probes were hybridised with 20µg of total RNA isolated from U937 cells and 40µg of tRNA. Following RNase digestion samples were electrophoresed on a 6% (w/v) acrylamide/7M urea denaturing gel with an M13mp18 sequence ladder as size markers. Specific digestion products of 243, 225 and 212bp in the track containing probes hybridised with RNA isolated from U937 cells are indicated.

*Hin*dIII cut pGEM clone containing the 520bp *Sty*I/*Pvu*II insert, using the bacteriophage T7 promoter in the vector. The antisense riboprobe was hybridised to total U937 RNA and then digested with RNase A and RNase T_1 to remove single-stranded RNA. The specific RNA:RNA hybrid products (Figure 4.8), not present in the tRNA control, were 212bp, 225bp and 243bp in size. The largest product, of 243bp, suggested that the cDNA sequence could be extended by 22bp so that the longest mRNA sequence is apparently 886 nucleotides. This is in good agreement with the size of the mRNA species detected by Northern blot analysis (Figure 4.5).

4.2.6 G7b mRNA structure

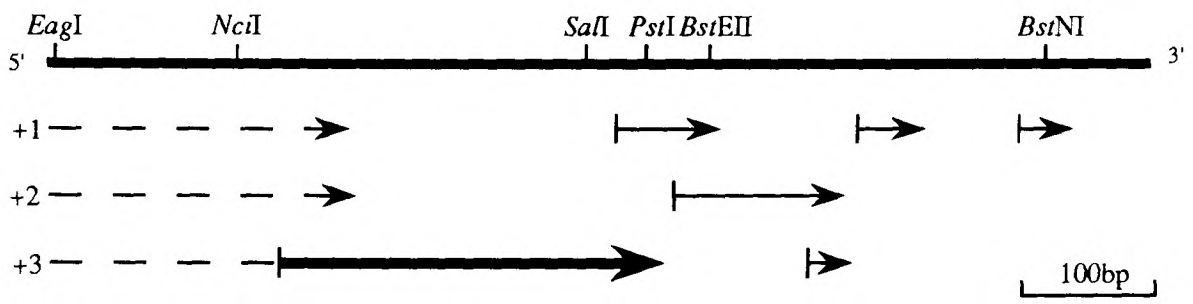
Translation of the cDNA sequence data in three phases revealed the potential ORFs in the G7b cDNA sequence (Figure 4.9). The first methionine codon (ATG) is located 219bp from the start of the cDNA sequence (Figure 4.9) and conforms to the start consensus, $^{A}/_{G}$NCAUG, of Cavener and Ray (1991) and Kozak (1991). However, although this is the first methionine codon, there are no stop codons upstream of this position, or in the 22bp extending 5′ from this sequence (Figure 4.15). The ORF starting with the first methionine extends for 285bp and encodes a 95 amino acid residue polypeptide (Figure 4.9 C). Therefore, including the 22bp extension derived from the RNase mapping result, the G7b mRNA consists of a 5′ untranslated region of 240bp, a 3′ untranslated region of 361bp containing a polyadenylation signal and an ORF of 285bp.

The deletion of one exon in clone G7b.3-a [section 4.2.4] alters the ORF to produce a truncated putative polypeptide of 26 amino acids (Figure 4.9 B).

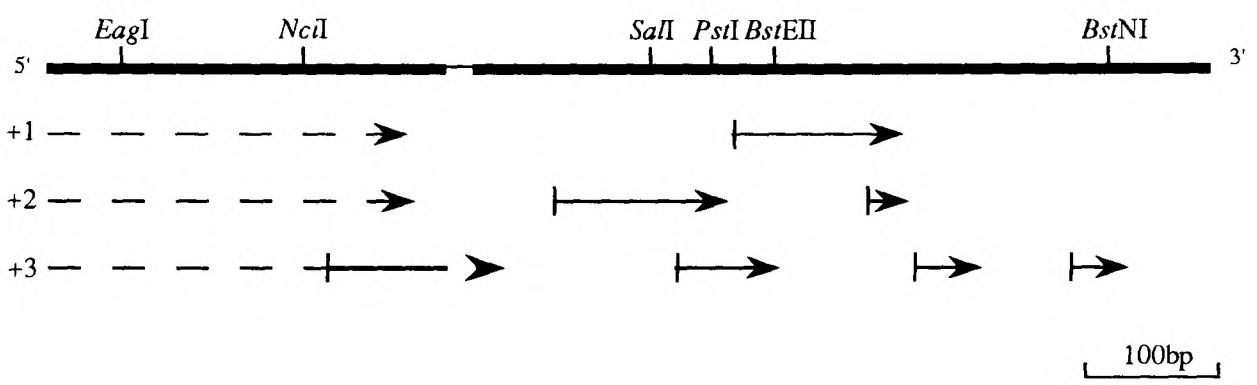
4.3 DETERMINATION OF G7b GENE STRUCTURE

The presence of a deletion in one of the cDNA clones necessitated confirmation of the sequence of the G7b exons. As a result, the G7b gene structure was determined in order to confirm the integrity of the cDNA sequence. Also, in the absence of a single putative ORF in the G7b cDNA sequence, confirmation of the nucleotide sequence could help to determine the derived amino sequence.

A



B



C

```
GGCTTCCTCCTCCCCTCTAC TCTACCCAGCTTGCGCTCCC CAGCCGCAAGTCGGCCGCGC TTTGCCCGTCAGCGCTTGGA GCTTTCTGCGTCGCTTCCCG  100

CTGCGCCTGCGCGGTCCCGC CTCGCCCCACGCGCGGGCTC GCGCTTCGGTTTCCCCAGAC CTGCTCGCAGCACCCTGCTG TCTTCCCGGTCCGGCCCGCT  200

                   M   L   F   Y   S   F   F   K   S   L   V   G   K   D   V   V   V   E   L   K   N   D   L   S   I   C   G    27
GCCCGCGGCGCCAGCACCAT GCTCTTCTATTCTTTTTTCA AGTCCCTTGTGGGCAAGGAT GTGGTCGTGGAACTAAAGAA TGACCTGAGCATCTGTGGAA  300

T   L   H   S   V   D   Q   Y   L   N   I   K   L   T   D   I   S   V   T   D   P   E   K   Y   P   H   M   L   S   V   K   N   C   F   61
CCCTCCATTCTGTGGATCAG TATCTCAACATCAAACTAAC TGACATCAGTGTCACAGACC CTGAGAAATACCCTCACATG TTATCAGTGAAGAACTGCTT  400

  I   R   G   S   V   V   R   Y   V   Q   L   P   A   D   E   V   D   T   Q   L   L   Q   D   A   A   R   K   E   A   L   Q   Q   K
CATTCGGGGCTCAGTGGTCC GATACGTGCAGCTGCCAGCA GATGAGGTCGACACACAGTT GCTACAGGATGCGGCAAGGA AGGAAGCCCTGCAGCAGAAA  500

  Q   *   95
CAGTGATGGCTCCTCCTCCT CTTCCCCTCCCTCTTTCATT GGTGACCCATAACCCCAAGT CCCAGCCCAGAACCCCTAAC CCCCAATACTTGAAGGGGTT  600

TTGTTTTTTTACTAATGATG GTTTTGTGGGTTTTTTTTAA GGGATGAGTGGATGAGAGGA GTAATAGGGAACAGCTATCC TCTCTTGAGAAGGGGAGGAT  700

AAGTAGGCTGGGAAACTTCA AAGCCTTCCCAGTCCCCAGC ACCTGCCTTTCTCACTACTT CTCTGGAGATGGTAGGAGAG TTTCCTAGGTCTTTCCAGGG  800

CAGCATGTGATTCATTTGGG GATGGAAGGAATCTGTCCCG CATCGGGAATAAAATTTATG ATGC  864
```

Figure 4.9 Open reading frames in the G7b inserts.
The G7b cDNA inserts G7b.1-7 (**A**) and G7b.3-a (**B**) are shown as horizontal lines with a narrow line in G7b.3-a at the position of the 31bp deletion. ORFs are represented as arrows in the phase in which they occur. Methionine codons at the 5' ends are shown by vertical lines on the arrows and ORFs at the 5' end without methionine codons are shown by arrows with dashed lines. The longest continuous ORF is marked by a bold arrow.
C The derived amino acid sequence of the longest ORF is shown above the cDNA sequence. The polyadenylation signal AATAAA is underlined.

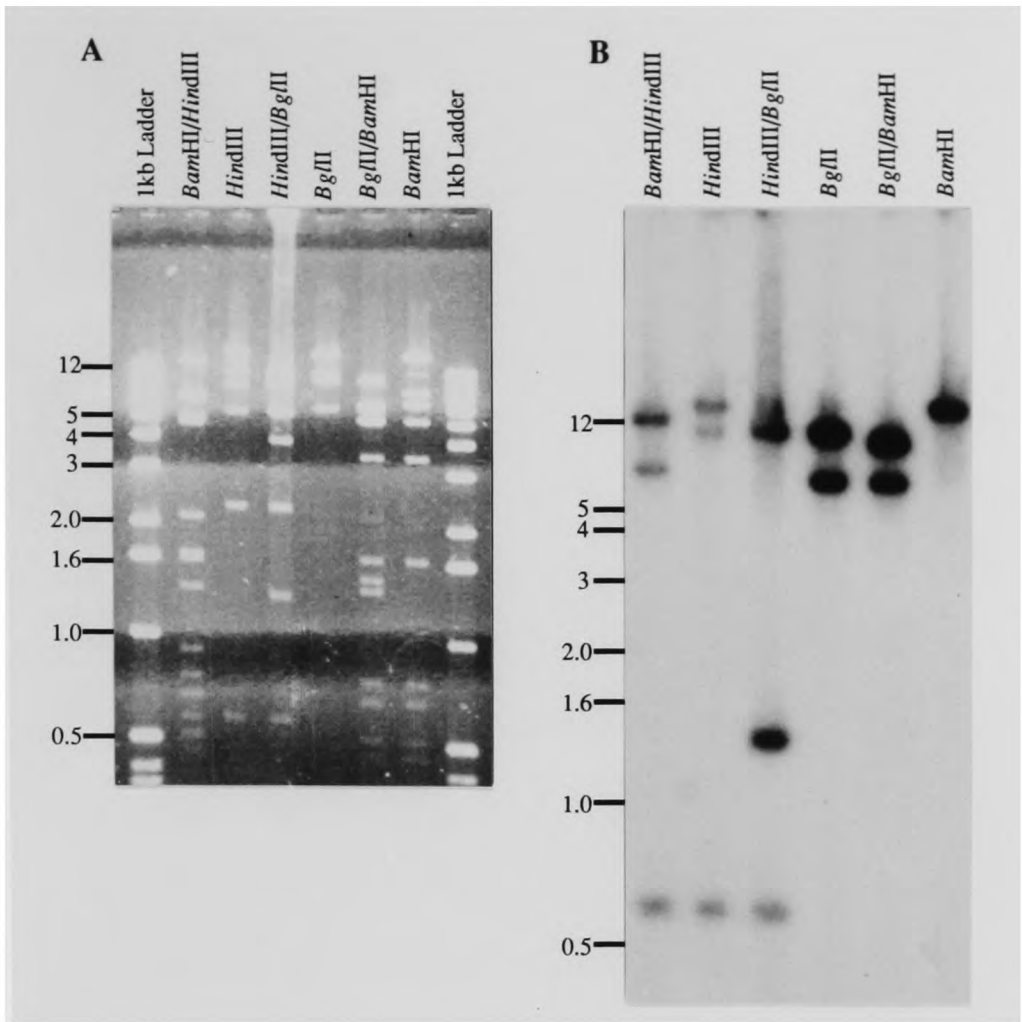


Figure 4.10 Mapping the G7b gene within cosmid 10S.
A Digested cosmid 10S DNA fractionated on a 0.8% (w/v) agarose gel. **B** Southern blot of cosmid 10S DNA hybridised with the G7b.1-7 cDNA insert. Sizes are in kb.

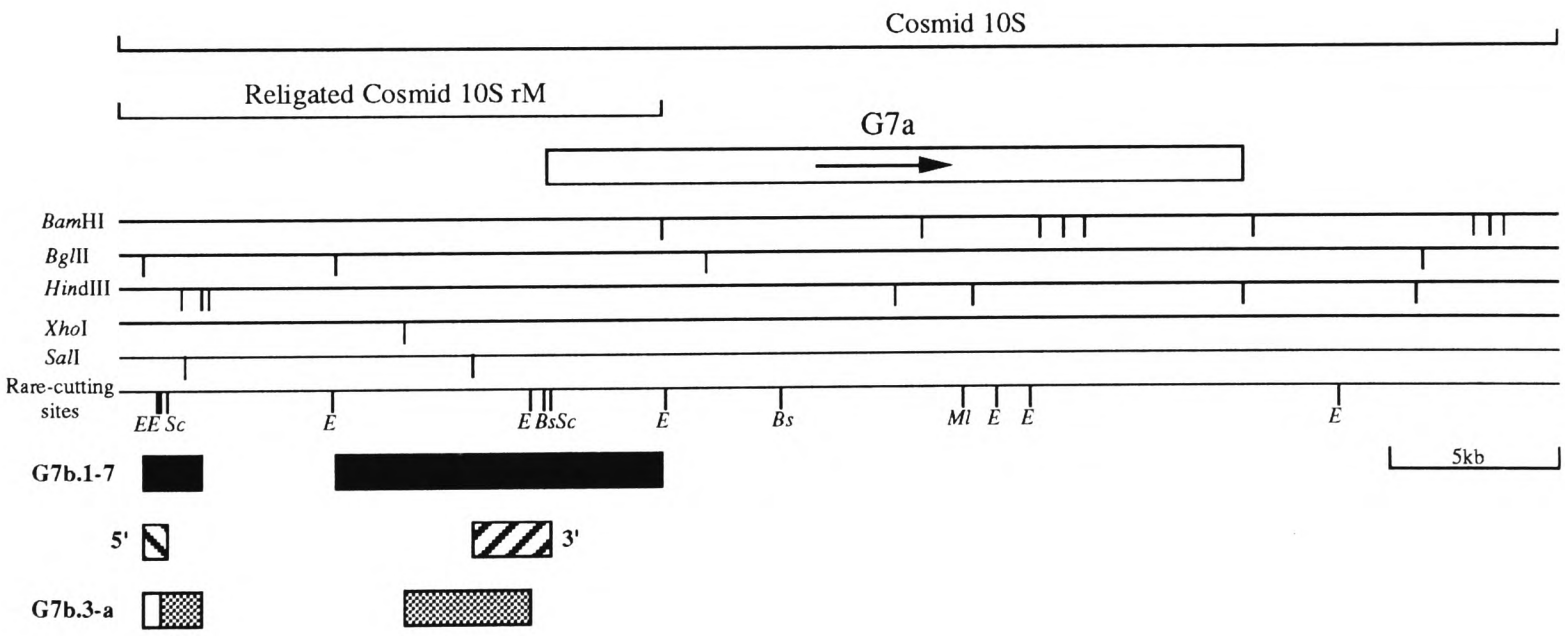


Figure 4.11 Map of G7b cDNA insert hybridisations within genomic DNA.
The upper part of this figure shows the cosmid 10S and religated cosmid 10S rM inserts. The location of the G7a gene is depicted as an open box with an arrow to show its direction of transcription. The positions of restriction sites in the inserts including sites for the rare-cutting enzymes *Eag*I (E), *Sac*II (Sc), *Bss*HII (Bs) and *Mlu*I (Ml) are shown. The regions hybridised by the G7.1-7 cDNA insert in figure 4.10 are shown as filled boxes. The regions hybridised by the 5' *Xho*I/*Nci*I and 3' *Bst*EII/*Xho*I cDNA fragments in figure 4.12 are shown as separate hatched boxes. The regions hybridised by the G7b.3-a cDNA insert in figure 4.13 are shown as shaded boxes with the 5' flanking region shown as an open box.

4.3.1 Mapping of the G7b gene

To map the location of G7b in finer detail, cosmid 10S DNA digested with *Bam*HI, *Bgl*II and *Hin*dIII, singly and in pairwise combinations, was fractionated on an agarose gel, transferred onto nitrocellulose by Southern blotting [section 2.5.1] and probed with the G7b.1-7 insert (Figure 4.10). The cDNA insert hybridised to two distinct regions of cosmid 10S that lie at least 4kb apart and span up to 15.5kb (Figure 4.11). This raised the possibility that there may be two homologous sequences between HSP70-Hom and G7a that hybridise to the cDNA insert.

To reduce the size of the cosmid insert in order to assist in the mapping of the region and in subcloning genomic DNA fragments, cosmid 10S was religated [section 2.7.4] at two *Bam*HI sites telomeric of G7b in the genomic DNA insert. This reduces the genomic DNA insert of cosmid 10S from 42.5kb to 18kb in the religated cosmid, 10S rM.

In order investigate the possibility that the G7b locus may be duplicated within the region between G7a and HSP70-Hom, the 5′ *Xho*I/*Nci*I and 3′ *Bst*EII/*Xho*I cDNA fragments [section 4.2.4] were radiolabelled [section 2.6.1] and each hybridised to a Southern blot of religated cosmid 10S rM DNA, digested with *Sal*I/*Sac*II, *Sal*I/*Xho*I, *Sal*I/*Bgl*II and *Hin*dIII/*Xho*I. The results confirm that the G7b cDNA is encoded by a single locus (Figure 4.12), that spans up to 12kb, between HSP70-Hom and G7a (Figure 4.11). In addition, the use of specific probes, from the 5′ and 3′ ends of the cDNA insert, revealed that the 5′ end of the gene lies adjacent to the 3′ end of HSP70-Hom and that the 3′ end lies adjacent to the 5′ end of G7a (Figure 4.11), in a head-to-tail arrangement. The 3′ limit of G7b also lies centromeric of the *Sac*II site at the 5′ end of G7a, confirming that the two genes are distinct and encode different mRNA species.

Additional mapping, using the G7b.3-a cDNA insert, was carried out to further refine the location of G7b exons. The cDNA probe was hybridised to a Southern blot of cosmid 10S rM DNA, digested with *Sal*I/*Bgl*II, *Sal*I/*Sac*II, *Sal*I/*Eag*I, *Xho*I/*Eag*I *Bgl*II/*Eag*I and *Bgl*II/*Sac*II combinations. The hybridisation pattern obtained (Figure 4.13), and the original mapping data [section 4.2.2], revealed that the G7b exons are encoded within two regions of genomic DNA, up to 2kb and 4kb in size, that lie at least 6kb apart (Figure 4.11).

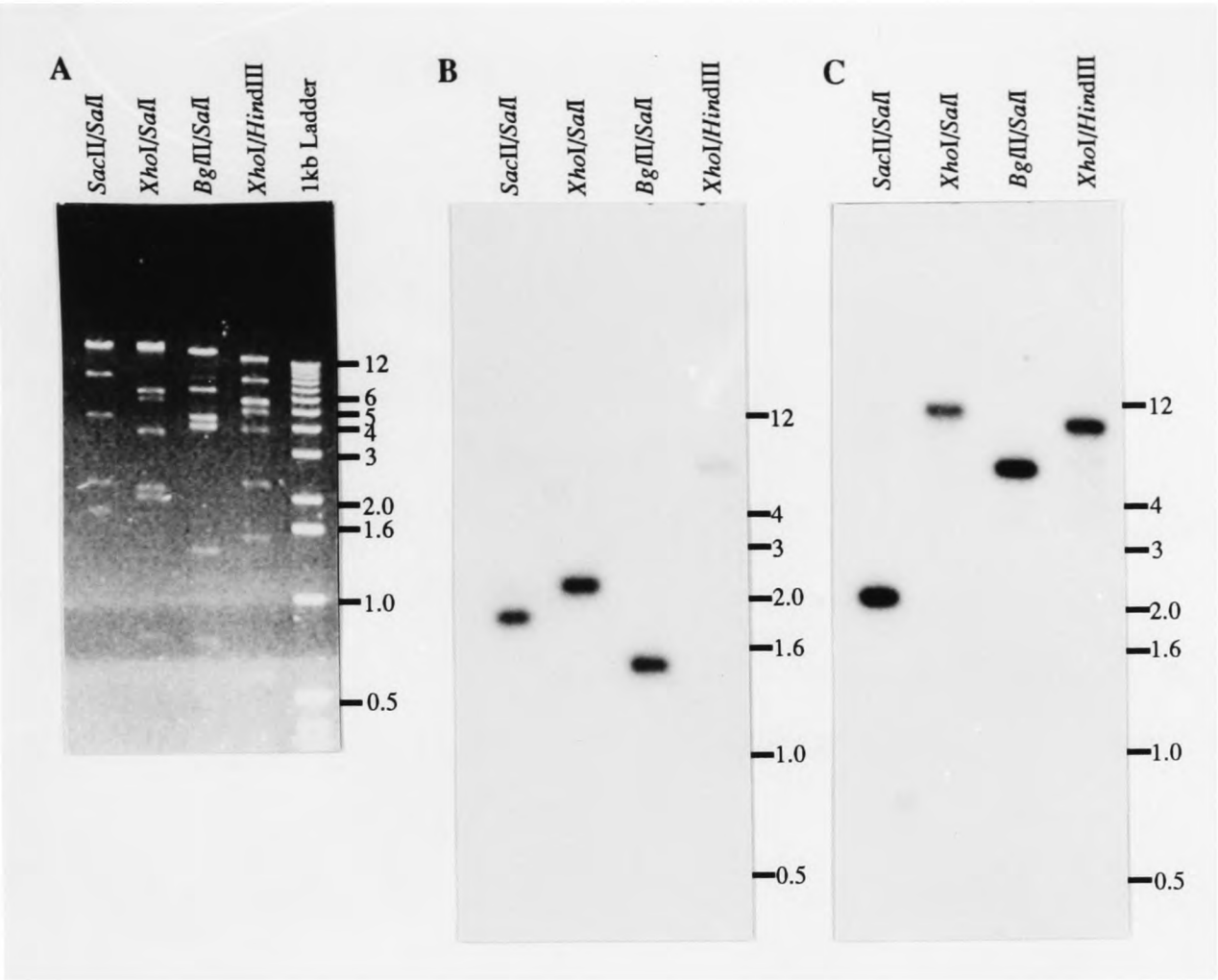


Figure 4.12 Mapping the orientation of the G7b gene.
A Digested cosmid subclone 10S rM DNA fractionated on a 0.8% (w/v) agarose gel. Southern blots of cosmid subclone 10S rM DNA hybridised with 5' *Xho*I/*Nci*I (**B**) and 3' *Bst*EII/*Xho*I (**C**) cDNA fragments are shown. Sizes are in kb.

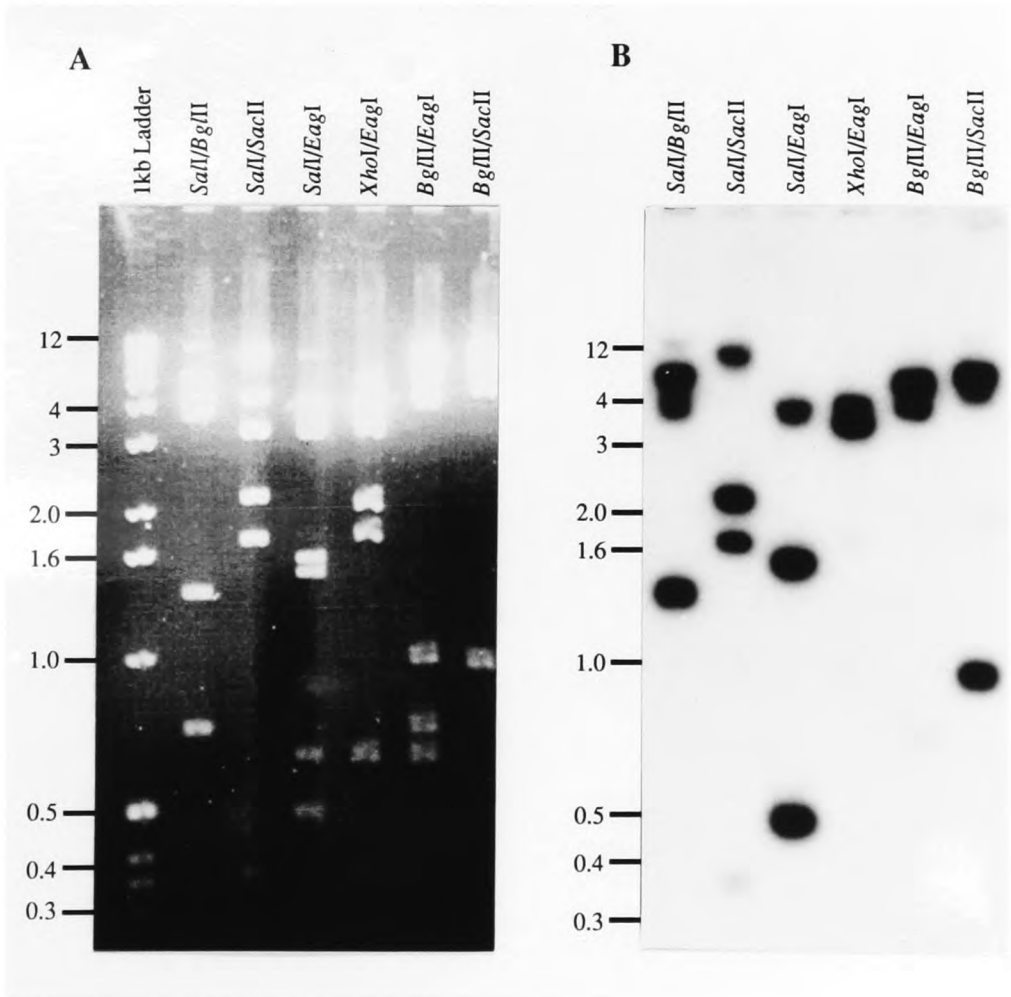


Figure 4.13 Mapping of the G7b gene within cosmid subclone 10S rM.
A Digested cosmid subclone 10S rM DNA fractionated on a 0.8% (w/v) agarose gel. **B** Southern blot of cosmid subclone 10S rM DNA hybridised with the G7b.3-a cDNA insert is shown. Sizes are in kb.

4.3.2 Nucleotide sequence analysis of G7b exons

4.3.2.1 *Sequencing the 5′ region of G7b*

At the 5′ end, the coding region of G7b lies within an ~1.4kb *Bgl*II/*Sal*I DNA fragment and an ~0.6kb *Hin*dIII DNA fragment, which overlap for ~100bp (Figure 4.14). The ~1.4kb *Bgl*II/*Sal*I DNA fragment was isolated from the genomic DNA subclone, 10S rM [section 2.4.5], and recovered as described in section 2.4.6. The DNA fragment was subcloned into *Hin*cII cut pGEM-3Zf and sequenced using a series of further subcloning steps (Figure 4.14) to generate templates for use in sequencing reactions [section 2.9.4] primed with the M13 universal primer. At one position a compression in the DNA sequence was resolved using cycle sequencing [section 2.9.5].

The ~600bp *Hin*dIII DNA fragment was excised from the subclone 10S rM and cloned directly into *Hin*cII cut pBC. Single-stranded DNA was prepared from the phagemid [section 2.9.2] and sequenced using the M13 universal primer [section 2.9.3] (Figure 4.14).

4.3.2.2 *Sequencing the 3′ region of G7b*

Within the 4kb of DNA to which the 3′ end of G7b had been mapped [section 4.3.1], few restriction sites had previously been identified. To obtain a more detailed restriction map of this region, the restriction sites known to cleave within the cDNA insert, and also a variety of other restriction sites, were mapped within the subclone 10S rM DNA. Approximately 500bp *Eco*RI/*Pvu*II and ~750bp *Hin*cII DNA fragments, flanking the *Sal*I site which is common to both the G7b cDNA and genomic DNA sequences, were subcloned directly from the subclone 10S rM into *Hin*cII cut pBC. Using the T3 pBluescript primer and the M13 universal primer, both inserts were sequenced using double-stranded DNA templates [section 2.9.4] (Figure 4.14). To complete the ~750bp *Hin*cII DNA fragment sequence, the clone was religated at the *Stu*I site, within the insert, and unique restriction sites in the vector polylinker sequence. The double-stranded templates generated were then sequenced using the T3 pBluescript primer and the M13 universal primer to complete the exonic sequence of G7b (Figure 4.14).

4.3.3 G7b gene structure

By comparison of the G7b cDNA sequence data [section 4.2.4] with the G7b genomic

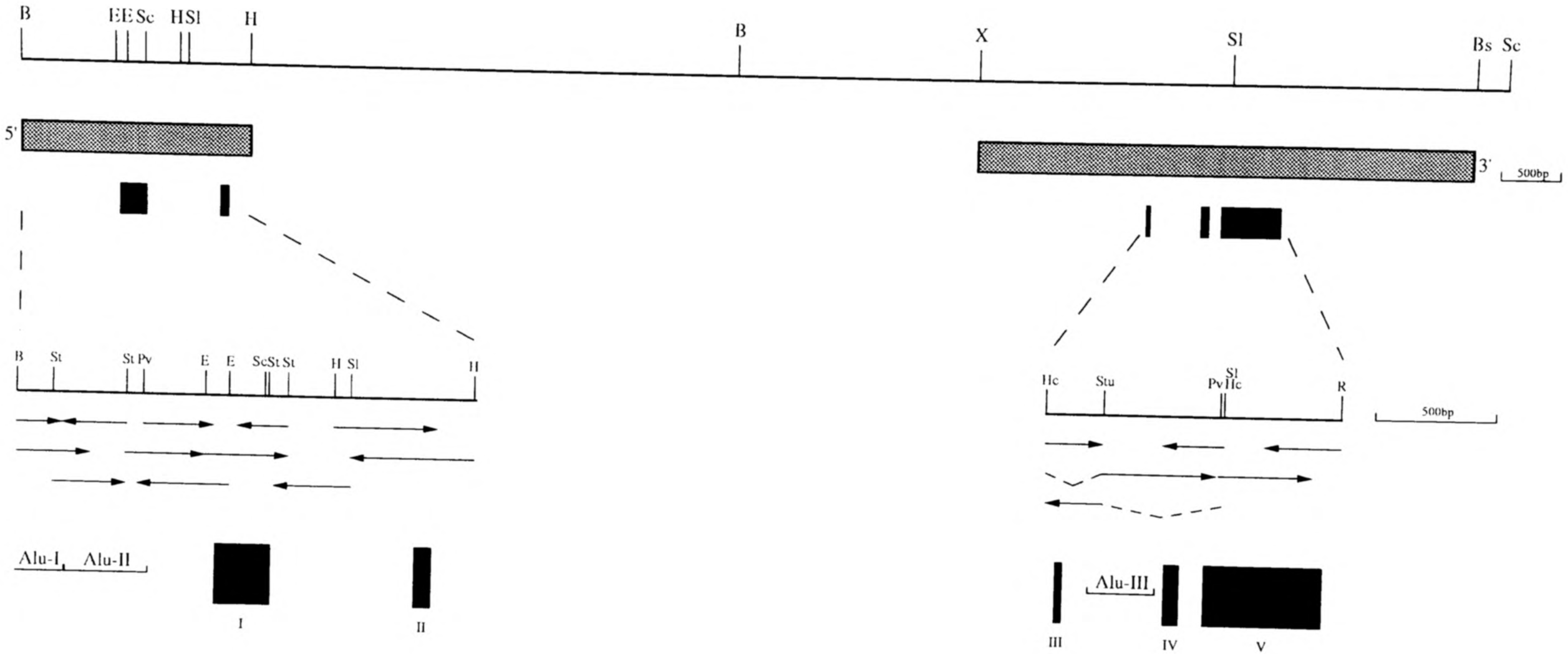


Figure 4.14 Molecular map of the organisation of the G7b gene.
The G7b gene was mapped by hybridisation to the regions shown by shaded boxes using restriction enzymes *Bgl*II (B), *Eag*I (E), *Sac*II (Sc), *Hin*dIII (H), *Sal*I (Sl), *Xho*I (X) and *Bss*HII (Bs). The regions hybridised were characterised in greater detail and additional restriction sites for *Sty*I (St), *Pvu*II (Pv), *Hin*cII (Hc), *Stu*I (Stu) and *Eco*RI (R) were mapped.The expanded regions were subcloned and sequenced to locate the G7b exons. The arrows beneath the restriction map indicate the orientation and extent of the sequence obtained from subcloned fragments. Deletions generated by the religation of subclones are shown by dashed lines. The G7b exons I-V are shown as filled boxes and the positions of *Alu* repeat sequences are indicated.

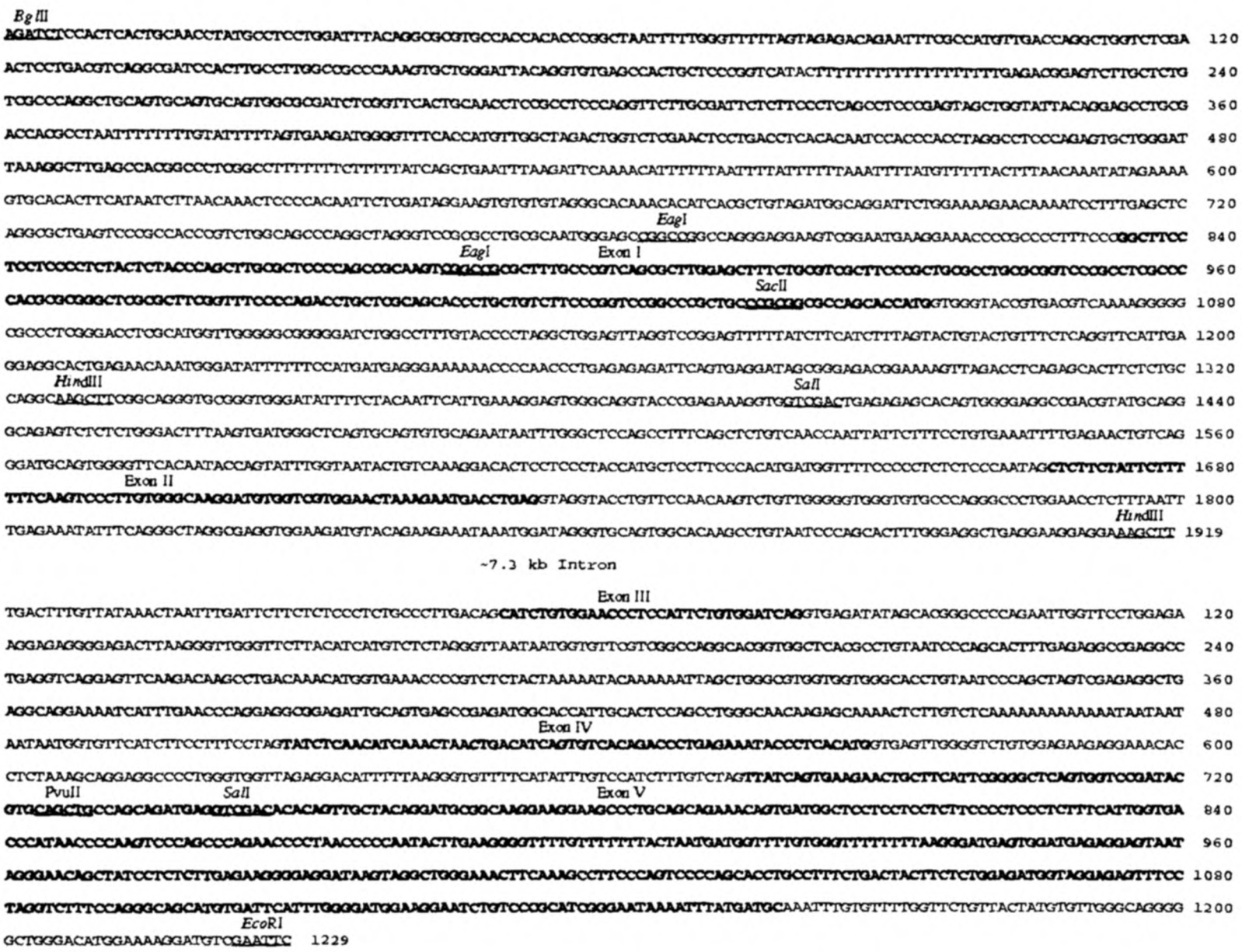


Figure 4.15 Nucleotide sequence of the G7b exons.

The G7b exons are shown in blue and numbered as in figure 4.14. ***Alu*** repeats are shown in red. Restriction enzyme recognition sequences are underlined.

DNA sequence data [section 4.3.2], the genomic organisation of G7b was defined (Figure 4.15). The G7b gene is organised into five exons, I to V, of 243, 68, 31, 60 and 484bp, respectively, that span ~10kb and are clustered into two groups. At the 5′ end, exons I and II lie 600bp apart, whilst at the 3′ end exons III and IV are separated by 227bp and exons IV and V are 106bp apart. These two groups of exons are separated by an ~8kb intron. The putative G7b ORF starts in exon I and extends into exon V [section 4.2.6]. There is no variation between the sequence derived from the cDNA clones and the genomic DNA, except for the 31bp deletion observed in cDNA clone G7b.3-a [section 4.2.4] which corresponds to exon III. All the exon-intron boundaries were sequenced and all the splice sites conform to the GT/AG rule (Krainer and Maniatis 1988) (Figure 4.16).

4.3.4 Identification of an additional HTF-island at the 5′ end of G7b

Analysis of the CpG content of 2kb of genomic DNA sequence, encoding exons I and II of the G7b gene and extending ~800bp upstream of the 5′ end of the cDNA sequence, was carried out using the GCG package CPGPLOT program. The result revealed the presence of a CpG-rich sequence characteristic of an HTF-island at the 5′ end of G7b, between nucleotides 650 and 950 (Figure 4.17). This region contains two *Eag*I sites within the putative island and a *Sac*II site, clustered about exon I (Figure 4.15), which are also cleaved in genomic DNA (Sargent *et al.* 1989a), confirming that this CpG-rich sequence is probably unmethylated in uncloned genomic DNA. Therefore, G7b is probably transcribed from a previously undefined HTF-island which is located ~2.0kb from the 3′ end of HSP70-Hom.

4.3.5 Upstream sequence analysis of G7b

Genomic DNA sequence analysis of G7b included ~800bp of sequence upstream of the 5′ end of the cDNA start sites (Figure 4.15). Within this sequence there are two *Alu* repeat sequences [section 4.3.6] in addition to the 5′ end of the HTF-island [section 4.3.4]. This region was analysed for the presence of putative transcription factor binding sites, by comparison with the consensus elements reported by Locker and Buzard (1990) and Jones *et al.* (1988), and the results are shown in figure 4.18. This region does not contain any CAAT or TATA box promoter sequence elements, which are normally located at specific positions relative to the transcription start site

```
                        M      aa 1                                                    aa 2      L  F  Y
Exon I    CAGCACCATG gtgggtaccg  intron 1 612bp  tggttttccccctctctcccaatag CTCTTCTATT  Exon II   68bp

          N  D  L  S    aa 24                                                   aa 25      I  C  G
Exon II   ATGACCTGAG gtaggtacct  intron 2 7.5kb  tcttctctccctctgcccttgacag CATCTGTGGA  Exon III  31bp

            V  D  Q     aa 34                                                   aa 35      Y  L  N
Exon III  TGTGGATCAG gtgagatata  intron 3 427bp  aatggtgttcatcttcctttcctag TATCTCAACA  Exon IV   60bp

            P  H  M     aa 54                                                   aa 55      L  S  V
Exon IV   CCCTCACATG gtgagttggg  intron 4 107bp  ttcatatttgtccatctttgtctag TTATCAGTGA  Exon V    487bp
```

Figure 4.16 Exon-intron boundaries in the G7b gene.

The G7b exon sequences at the exon-intron boundaries are shown in upper case and the intron sequences, including the polypyrimidine tracts at the 3' ends, flanking the G7b exons are shown in lower case. The invariant **ag** and **gt** sequences at the splice sites are shown in bold and the sizes of the exons and introns are included. The derived amino acid sequences at the splice sites and their positions in the protein sequence are shown above the nucleotide sequence.

aa - amino acid.

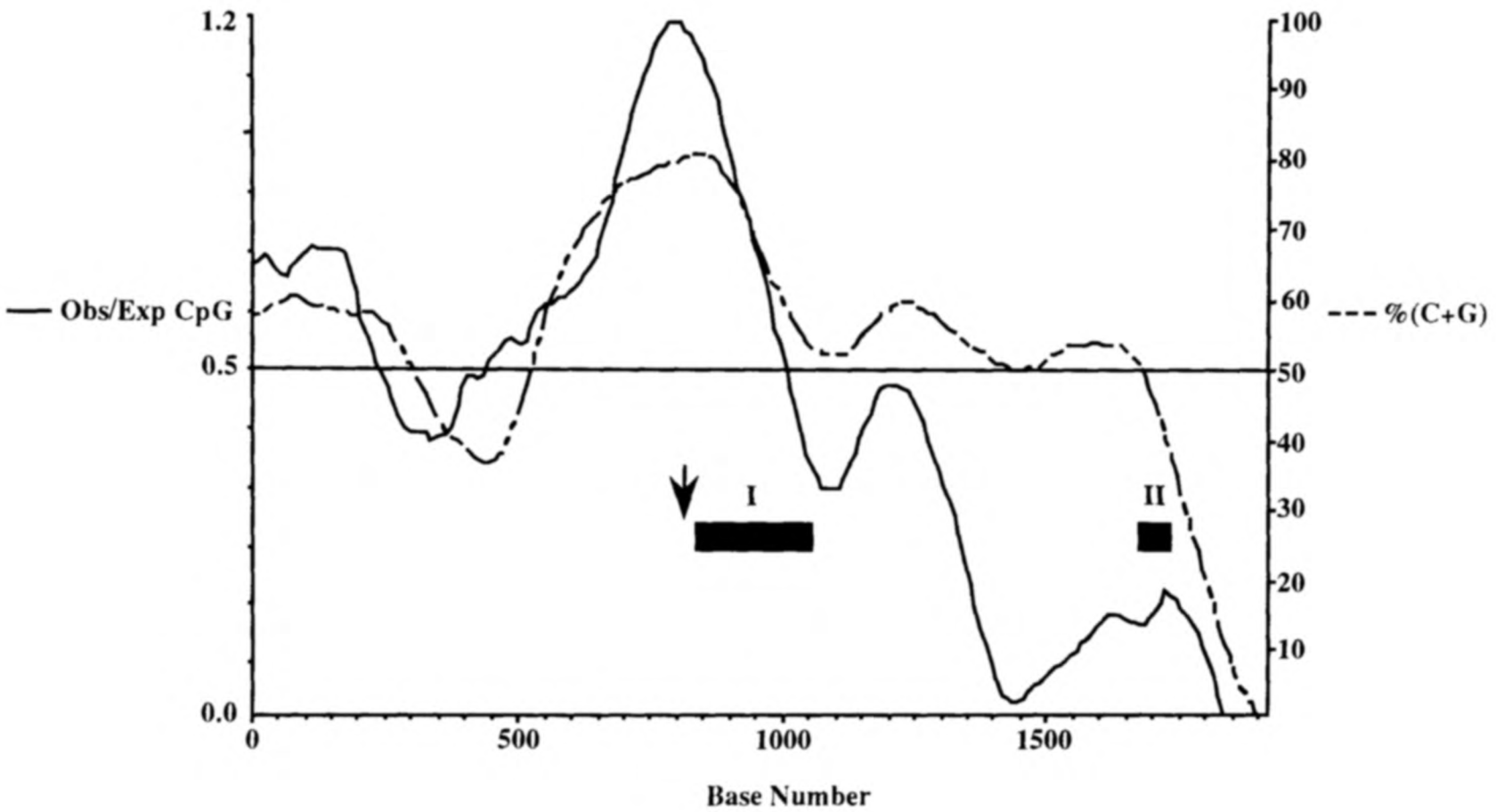


Figure 4.17 Gardiner-Garden Plot of the 5' end of G7b.
The CpG content of the nucleotide sequence from the 5' contig (Figure 4.15) analysed by the CPGPLOT program, using a 200 base window and shift increment of 10, is shown. Exons I and II of the G7 gene are shown as filled boxes and the 5' extent of the RNase mapping products is shown as a vertical arrow.

Obs/Exp is (No. of CpG dinucleotides/No. of C x No. of G) x N
where N is the window size. (Gardiner-Garden and Frommer 1987)

AGATCTCCACTCACTGCAACCTATGCCTCCTGGATTTACAGGCGCGTGCCACCACACCCGGCTAATTTTTGGGTTTTTAGTAGAGACAGAATTTCGCCATGTTGACCAGGCTGGTCTCGA 120

ATF/CRE
ATF/CRE Sp1
ACTCCTGACCTCAGGCGATCCACTTGCCTTGGCCGCCCAAAGTGCTGGGATTACAGGTGTGAGCCACTGCTCCCGGTCATACTTTTTTTTTTTTTTTTTTTTGAGACGGAGTCTTGCTCTG 240

AP2
TCGCCCAGGCTGCAGTGCAGTGCAGTGGCGCGATCTCGGTTCACTGCAACCTCCGCCTCCCAGGTTCTTGCGATTCTCTTCCCTCAGCCTCCCGAGTAGCTGGTATTACAGGAGCCTGCG 360

GC Box TFIID E11a NF-kB AP3 AP2
ACCACGCCTAATTTTTTTTGTATTTTTAGTGAAGATGGGGTTTCACCATGTTGGCTAGACTGGTCTCGAACTCCTGACCTCACACAATCCACCCACCTAGGCCTCCCAGAGTGCTGGGAT 480

TAAAGGCTTGAGCCACGGCCCTCGGCCTTTTTTTCTTTTTATCAGCTGAATTTAAGATTCAAAACATTTTTTAATTTTATTTTTTAAATTTTATGTTTTTACTTTAACAAATATAGAAAA 600

NF-E1 PEA3 EF-1A GR/PR
GTGCACACTTCAATAATCTTAACAAACTCCCCACAATTCTCGATAGGAAGTGTGTGTAGGGCACAAACACATCACGCTGTAGATGGCAGGATTCTGGAAAAGAACAAAATCCTTTGAGCTC 720

AP2 PEA3 NF-kB Sp1
Sp1 AP2 EF-1A GC Box
AGGCGCTGAGTCCCGCCCACCCCTCTGGCAGCCCAGGCTAGGGTCCGCGCCTGCGCAATGGGAGCCGGCCGGCCAGGGACGAAGTCGGAATGAAGCAAACCCCGCCCCTTTCCCGGCTTCC 840

Exon I
TCCTCCCCTCTACTCTACCCAGCTTGCGCTCCCCAGCCGCAAGTCGGCCGCGCTTTGCCCGTCAGCGCTTGGAGCTTTCTGCGTCGCTTCCCGCTGCGCCTGCGCGGTCCCGCCTCGCCC 960

CACGCGCGGGCTCGCGCTTCGGTTTCCCCAGACCTGCTCGCAGCACCCTGCTGTCTTCCCGGTCCGGCCCGCTGCCCGCGGCGCCAGCACCATGGTGGGTACCGTGACGTCAAAAGGGGG 1080

Figure 4.18 Analysis of the 5' end of the G7b gene.
Exon I of the G7b gene is shown in blue and the HTF-island shown in figure 4.17 is shown in pink.Potential binding sites for transcription factors are shown in green and those oriented 5' to 3' are underlined whilst those oriented 3' to 5' are overlined. *Alu* repeat sequences are shown in red. † indicates the 5' limits of RNase mapping products.

of a gene. However, sequences for multiple potential GC boxes, Sp1 and AP-2 binding sites and other transcription factor binding sites were identified, clustered around the 5′ end of the cDNA sequences and the putative transcription start site, and extending upstream into the *Alu* repeat elements (Figure 4.18). Although the potential transcription start sites have been identified by RNase mapping, it is not clear which transcription factors may be involved in the control of G7b transcription.

4.3.6 *Alu* repeat sequences in G7b

Using the genomic DNA sequence data obtained in section 4.3.2, divided into overlapping 500bp fragments, the GENEMBL DNA database was screened using the GCG package FASTA program. The results identified two complete repeats and another partial repeat sequence from the *Alu* family of SINES [section 3.3.3.2]. One of the complete repeat sequences, III, is located in the third intron, between exons III and IV, oriented 5′ →3′ within G7b, and contains a mixed poly(A) tract (Economou *et al.* 1990) of thirteen adenines followed by four $(TA_2)_n$ repeats. The other complete repeat, II, lies ~300bp upstream of the transcription start site of G7b, oriented towards HSP70–Hom with a simple, type I (Economou *et al.* 1990) poly(A) tract of nineteen adenines. In addition, a partial *Alu* repeat, I, which contains a 40bp deletion lies adjacent to the 3′ end of the second, complete repeat, in the 5′ untranscribed flanking sequence, oriented in the same direction, towards the 3′ end of HSP70-Hom (Figure 4.15).

These *Alu* repeat sequences were aligned with the consensus *Alu* repeat sequence of Jurka and Smith (1988) using the CLUSTAL V program (Higgins and Sharp 1988) (Figure 4.19). These three *Alu* repeats associated with G7b, like those identified in G7, appear to belong to the class II (Britten *et al.* 1988) or the consensus, branch *a* of the S subfamily, which is the most common (~55%) type of *Alu* repeat (Jurka and Smith 1988).

4.3.7 RFLP analysis of G7b

In order to identify RFLPs associated with the G7b gene that may be used for linkage analysis to determine whether the gene is associated with any inherited disorder, RFLP analysis was carried out for G7b.

To search for RFLPs, Southern blots of total genomic DNA from ten, HLA-typed,

```
ALU  GGCCGGGC-GGGTGGCTCAC GCCTGTAATCCCAGCACTTT GGGAGGCCGAGGCGGGCGGA TCACCTGAGGTCAGGAGTTC
III  GGCCAGGCACGGTGGCTCAC GCCTGTAATCCCAGCACTTT GAGAGGCCGAGGC------- ----CTGAGGTCAGGAGTTC
II   GCCCGAGGGCCGTGGCTCAA GCCTTTAATCCCAGCACTCT GGGAGGCCTAGGTGGGTGGA TTGTGTGAGGTCAGGAGTTC
I    GACCGGGAGCAGTGGCTCAC ACCTGTAATCCCAGCACTTT GGGCGGCCAAGGCAAGTGGA TCGCCTGACGTCAGGAGTTC

ALU  GAGACCAGCCTGGCCAACAT GGTGAAACCCCGTCTCTACT AAAAATACAAAAA-TTAGCC GGGCGTGGTGGCGCGCGCCT
III  AAGACAAGCCTGACAAACAT GGTGAAACCCCGTCTCTACT AAAAATACAAAAAATTAGCT GGGCGTGGTGGTGGGCACCT
II   GAGACCAGTCTAGCCAACAT GGTGAAACCCCATCTTCACT AAAAATACAAAAAA--AATT AGGCGTGGTCGCAGGCTCCT
I    GAGACCAGCCTGGTCAACAT GGCGAAATTCTGTCTCTACT AAAAACCCAAAAA-TTAGCC GGGTGTGGTGGCACGCGCCT

ALU  GTAATCCCAGCTACTCGGGA GGCTGAGGCAGGAGAATCGC TTGAACCCGGGAGGCGGAGG TTGCAGTGAGCCGAGATCGC
III  GTAATCCCAGCTAGTCGAGA GGCTGAGGCAGGAAAATCAT TTGAACCCAGGAGGCGGAGA TTGCAGTGAGCCGAGATGGC
II   GTAATACCAGCTACTCGGGA GGCTGAGGGAAGAGAATCGC AAGAACCTGGGAGGCGGAGG TTGCAGTGAACCGAGATCGC
I    GTAA---------------- -------------------- ----ATCCAGGAGGCATAGG TTGCAGTGAGTGGAGATC--

ALU  GCCACTGCACTCCA-----G CCTGGGCGACA-GAGCGAGA CTCC-GTCTC A-rich tract
III  ACCATTGCACTCCA-----G CCTGGGCAACAAGAGCAAAA CTCTTGTCTC AAAAAAAAAAAAA (TAA)4
II   GCCACTGCACTGCACTGCAG CCTGGGCGACA-GAGCAAGA CTCC-GTCTC AAAAAAAAAAAAAAAAAAA
I    ------------------- -------------------- ----------
```

Figure 4.19 Alignment of the G7b *Alu* repeat sequences with the consensus sequence.
The *Alu* repeat sequences obtained from the G7b genomic sequence (I-III) are aligned with the consensus sequence (ALU) of Jurka and Smith (1988). The alignment was produced by the CLUSTAL program (Higgins and Sharp 1988) and adjusted manually.

unrelated individuals (Table 3.2), digested with restriction enzymes *Bam*HI, *Bgl*II, *Hin*dIII, *Bst*EII, *Nco*I, *Pvu*II, *Pst*I, *Sst*I, *Eco*RI, *Rsa*I, *Taq*I and *Msp*I, were hybridised with the G7b.1-7 cDNA probe. The only polymorphism, detected in DNA digested with *Pvu*II, was present only in DNA isolated from the HLA homozygous cell line, ICE5, and was associated with the additional, cross-hybridising DNA fragments [section 4.2.3]. Therefore, the restriction enzymes used do not generate any polymorphic DNA fragments associated with G7b.

4.4 CHARACTERISATION OF THE G7b GENE PRODUCT

4.4.1 Features of the putative G7b polypeptide sequence

The longest ORF identified in the G7b cDNA sequences [section 4.2.4], confirmed by the genomic DNA exon sequence data [section 4.3.2], starts with the first methionine codon at the 5′ end, and encodes a putative polypeptide of 95 amino acids (Figure 4.20 A), with a M_r of 10835. At the N-terminus, the predicted G7b protein does not contain a hydrophobic leader peptide (von Heijne 1988) and a hydrophobicity plot (Figure 4.20 B) indicates that there are no significant hydrophobic regions that may serve as transmembrane segments. In addition, the polypeptide sequence does not include any putative N–linked glycosylation sites ($NX^S/_T$) or highly charged regions (Figure 4.20 A). The protein sequence does, however, include two cysteine residues which would remain as reduced, thiol groups in an intracellular environment (Barbosa *et al.* 1989). A DIAGON plot (Devereux *et al.* 1984) of the predicted G7b protein sequence revealed the absence of any repeated sequences.

4.4.2 G7b protein sequence similarity searches

The 95 amino acid translation product from the predicted G7b ORF (Figure 4.20 A) was used to screen the SwissProt and NBRF protein databases using the GCG package FASTA program [section 2.12.1]. The results revealed similarities with the human small nuclear ribonucleoprotein Sm-D autoantigen (Rokeach *et al.* 1988) and with short stretches of sequence from several members of the annexin protein family. However, since the G7b cDNA sequence starts with ORFs in all three phases (Figure 4.9), the remaining two translated sequences were also used to screen the same

A

```
M L F Y S F F K S L  V G K D V V V E L K  N D L S I C G T L H  30
S V D Q Y L N I K L  T D I S V T D P E K  Y P H M L S V K N C  60
F I R G S V V R Y V  Q L P A D E V D T Q  L L Q D A A R K E A  90
L Q Q K Q *  95
```

B

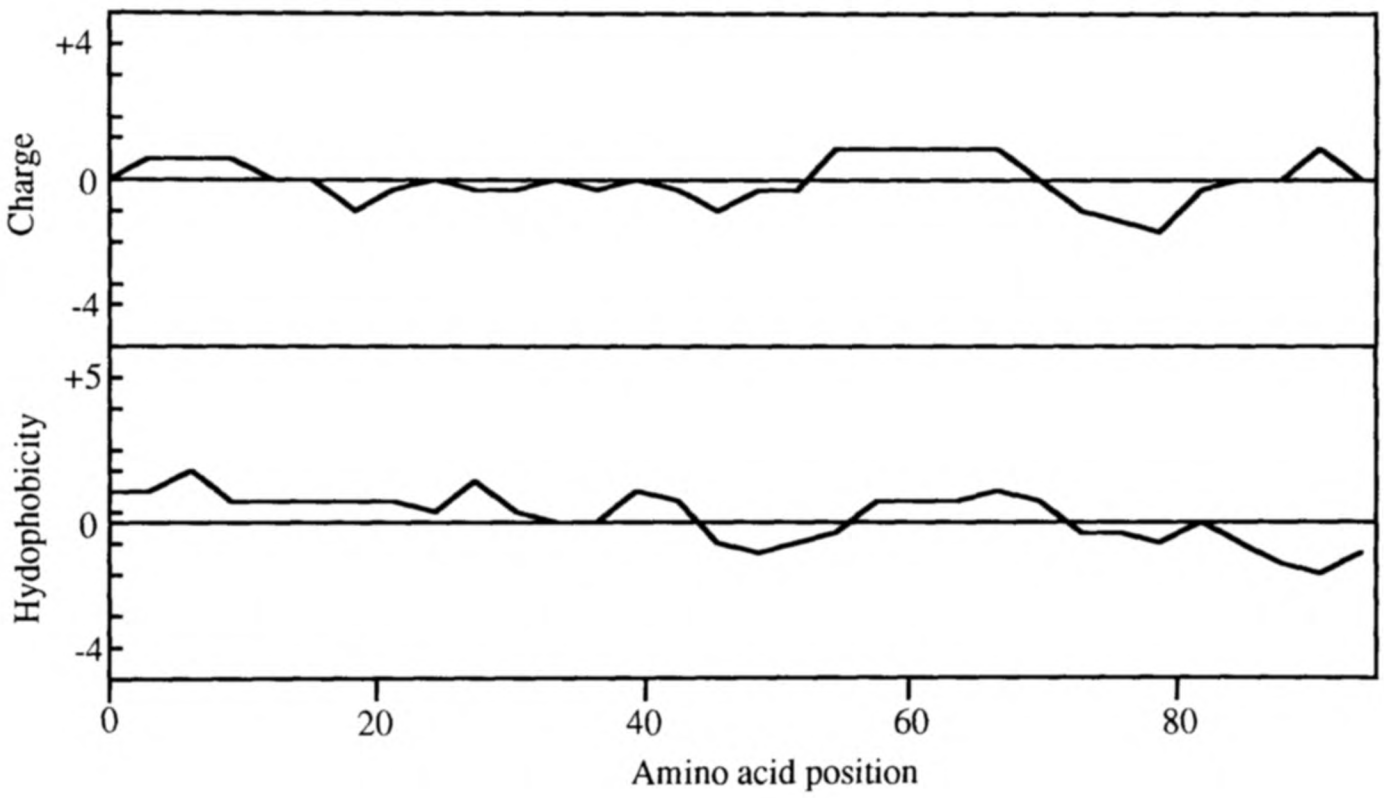


Figure 4.20 Analysis of the derived amino acid sequence from the G7b cDNA sequence. **A** The amino acid sequence translated from the G7b cDNA clones is shown, with an asterisk (*) indicating the stop codon. **B** Hydrophobicity and charge profiles of the putative G7b protein sequence were generated using the Protein Interpretation Program (PIP) using the default parameters.

protein databases. No significant similarity with any protein sequence was detected with these translated sequences.

The putative G7b polypeptide shares 29.5% amino acid sequence identity with Sm-D, starting with both initiation methionines, over the whole protein sequence, rising to 43% when chemically equivalent amino acid substitutions are taken into account. However, Sm-D contains nine Gly-Arg repeats at the C-terminus, which are absent in G7b (Figure 4.21 A). Sm-D is one of the core proteins of the small nuclear ribonucleoprotein complexes implicated in RNA processing and the presence of antibodies to the Sm-D polypeptide antigen are closely associated with SLE.

The Sm-D and putative G7b protein sequences were compared using the ALIGN program (Dayhoff *et al.* 1983) to determine the statistical significance of the sequence identity. The ALIGN analysis was carried out using the 250 PAMs Mutation Data Matrix (Dayhoff *et al.* 1978) with a bias of 6 added to each term and a break penalty of 6. One hundred random runs were performed to calculate a mean random score. The ALIGN score consists of the difference between the real score (R), of 684, and the mean score (M), of 571.16, divided by the standard deviation (SD), 12.18; *i.e.* (R-M)/SD. This score, of 9.26, in SD units, represents the SD of the pairwise score from the mean random score. Since this value is greater than 6, it indicates that the sequence similarity shown in figure 4.21 A is very significant.

Amino acids 7–24 of the predicted G7b polypeptide sequence share 40-50% identity, over 18 amino acids, with human (Ahn *et al.* 1988; Tait *et al.* 1988), pig (Weber *et al.* 1987) and bovine (Hamman *et al.* 1988) annexin IV, human (Grundmann *et al.* 1988), rat (Pepinsky *et al.* 1988) and chicken (Pfaeffle *et al.* 1990) annexin V, human (Crompton *et al.* 1988) and mouse (Moss *et al.* 1988) annexin VI, human annexin VII (Burns *et al.* 1989), human annexin VIII (Hauptmann *et al.* 1989) and annexin IX from *Drosophila melanogaster* (Johnston *et al.* 1990). This rises to 80-90% similarity when conservative substitutions are taken into account. To determine the statistical significance of the sequence identity, pairwise comparison of these sequences was carried out using the ALIGN program (Dayhoff *et al.* 1983) using the 250 PAMs Mutation Data Matrix (Dayhoff *et al.* 1978) with a bias of 6 added to each term and a break penalty of 6. Using one hundred random runs to calculate mean random scores, ALIGN scores between 6.0 and 8.2 were obtained. Although the similarity is over a short region of sequence, these scores are potentially significant and indicate

A

```
G7b   MLFYSFFKSLVGKDVVVELK NDLSICGTLHSVDQYLNIK LTDISVTDPEKYPHMLSVKNC  60
      | :  |:  |    | :||| |   : ||:  ||  :| : |  : :|   : |  |
Sm-D  MKLVRFLMKLSHETVTIELK NGTQVHGTITGVDVSMNTH LKAVKMTLKNREPVQLETLS-  59

G7b   FIRGSVVRYVQLPADEVDTQ LLQDAARKEALQQKQ  95
       |||  :||  ||         || |:  |    :
Sm-D  -IRGNNIRYFILPDSLPLDT LLVDVEPKVKSKKREAVAGR GRGRGRGRGRGRGRGRGGPR 118
```

B

Human annexin VIII	63-80	FKAQFGKDLTETLKSELS
G7B	7-24	FKSLVGKDVVVELKNDLS
Human annexin VII	205-222	FKTSYGKDLIKDLKSELS
Drosophila annexin IX	35-52	FKTSYGKDLISDLKSELG
Human annexin VI	61-78	YKSLYGKDLIADLKYELT
Chicken annexin V	57-74	FKTLFGRDLVDDLKSELT
Rat annexin V	55-72	FKTLFGRDLVNDMKSELT
Human annexin V	56-73	FKTLFGRDLLDDLKSELT
Bovine annexin IV	55-72	YKTTIGRDLMDDLKSELS
Pig annexin IV	55-72	YKSTIGRDLLDDLKSELS
Human annexin IV	55-72	YKSTIGRDLIDDLKSELS
Mouse annexin VI	404-421	FKSHFGRDLMADLKSEIS
		.*. *.*.. ..* ...

Figure 4.21 G7b protein sequence similarities.

A Alignment of the G7b derived amino acid sequence with the sequence of small nuclear ribonucleoprotein Sm-D using the ALIGN program (Dayhoff et al. 1983). Identical amino acids are shown by a line between the sequences (|) and conservative substitutions are shown by a colon (:). Breaks in the sequence to maximise the alignment are shown by dashes. The sequences share 29.5% sequence identity and 43% similarity when chemically equivalent substitutions are taken into account.

B Alignment of 18 amino acids from the G7b derived amino acid sequence with similar sequences from members of the annexin family of proteins generated using the CLUSTAL V program (Higgins and Sharp 1988). Identical amino acids are shown by an asterisk (*) below the sequence and chemically equivalent substitutions present in more than ten of the sequences are shown by a dot (.). The sequences share 39-50% identity with the G7b sequence and up to 95% similarity when chemically equivalent substitutions are taken into account.

that this part of the G7b protein sequence may have a similar structure to that of the annexin sequences. An alignment of these sequences using the CLUSTAL V program (Higgins and Sharp 1988) is shown in figure 4.21 B. Recently, the known annexin sequences were aligned (Huber *et al.* 1992) and related to the three dimensional structure of annexin V (Huber *et al.* 1990). The region that the G7b sequence shares similarity with spans 18 residues in domain I. These residues are the last four of helix 1C, four residues of turn and ten residues that constitute helix 1D (Huber *et al.* 1992). This region lies on the surface of the structure (Huber *et al.* 1990) on the edge of the convex face that is believed to bind to the cell membrane (Huber *et al.* 1992).

4.5 EXPRESSION OF THE G7b POLYPEPTIDE

4.5.1 Expression and isolation of G7b fusion protein

In order to confirm that G7b is expressed at the protein level, it is necessary to raise antibodies that detect the G7b polypeptide and towards this aim, the putative G7b polypeptide has been expressed using the *E.coli* expression vector, pGEX-2T system (Smith and Johnson 1988) [section 3.7.1].

Using the G7b.1-7 cDNA clone, a 490bp *Eag*I/*Bst*EII DNA fragment was excised, isolated on a 3.5% (w/v) polyacrylamide gel [sections 2.4.5 & 2.4.6], subcloned into *Sma*I cut pGEX-2T [section 2.7] and transformed into *E.coli* NM554 competent cells. This generated a pGEX-2T subclone that contained the ORF that encodes Sj26 followed by 56 amino acid residues from the cDNA 5′ untranslated region with the putative 95 amino acid G7b polypeptide at the C-terminus. Expression of this fusion protein was induced by the addition of IPTG [section 2.12.5] and the resultant 42.5kDa protein product was detected by SDS-PAGE of total *E.coli* cell lysate (Figure 4.22A).

The fusion protein was isolated from the cell lysate by absorption onto glutathione agarose beads [section 2.12.6] and the yield was estimated by SDS-PAGE analysis (Figure 4.22 B). The fusion protein was cleaved by thrombin digestion on the beads [section 2.12.7] and samples of both the supernatant and the glutathione agarose beads, removed during the digestion, were analysed by SDS-PAGE [section 2.12.2]. The results showed that the Sj26-G7b fusion protein was cleaved by thrombin, but that the 16.5kDa G7b polypeptide was apparently insoluble and

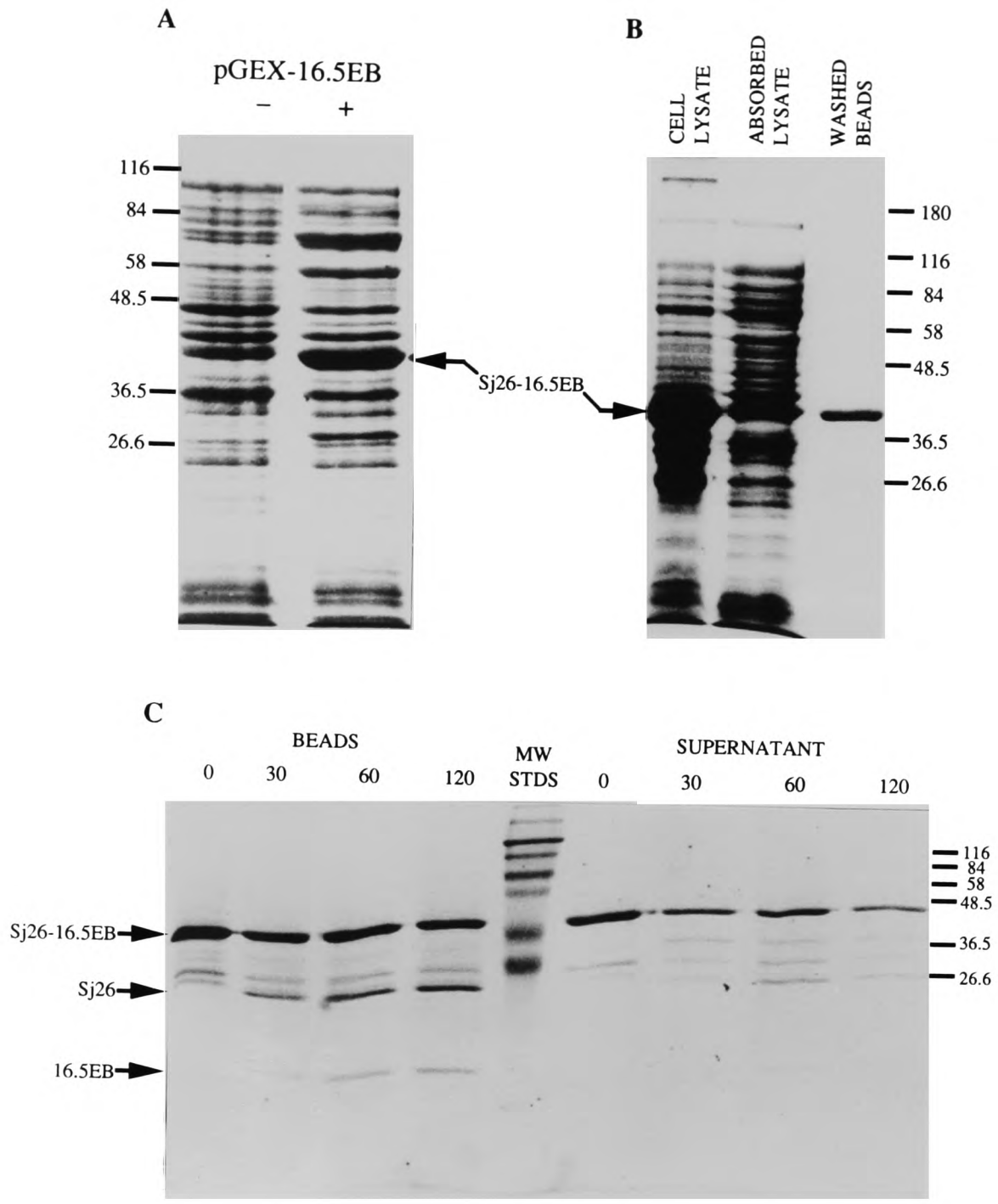


Figure 4.22 Expression of G7b fusion protein.

A Induction of expression of the 16.5kDa product from the 490bp *Eag*I/*Bst*EII G7b cDNA fragment with Sj26 as the Sj26-16.5EB fusion protein by IPTG. Total *E.coli* cell lysates fractionated on a 10% (w/v) acrylamide gel by SDS-PAGE are shown following induction by IPTG (+) or in the absence of IPTG (-). **B** Isolation of the Sj26-16.5EB fusion protein by absorption onto glutathione agarose beads. The lysate was removed by washing the beads to isolate the fusion protein absorbed onto the glutathione agarose beads which was visualised by SDS-PAGE analysis on a 10% acrylamide gel. The ~42.5kDa Sj26-16.5EB fusion protein is indicated.

C The Sj26-16.5EB fusion protein isolated on glutathione agarose beads was cleaved with thrombin at 25°C. Aliquots were removed 0, 30, 60 and 120 min after addition of thrombin and the beads and supernatant were analysed separately by SDS-PAGE on a 15% (w/v) acrylamide gel. The fusion protein and the cleavage products Sj26 and 16.5EB are indicated. None of the 16.5kDa G7b polypeptide is visible in the supernatant. Sizes are in kDa.

remained bound to the beads (Figure 4.22 C).

Since direct cleavage of the fusion protein did not release the G7b polypeptide, the fusion protein was eluted from the glutathione agarose beads using reduced glutathione [section 2.12.8] and then digested with thrombin. Analysis of these digestion products by SDS-PAGE revealed that the G7b polypeptide was degraded following thrombin cleavage of the fusion protein. As a result, the intact fusion protein was isolated from the glutathione agarose beads for use as an antigen to raise antibodies (Figure 4.23 A). Fusion protein yields were estimated by comparison with known amounts of ovalbumin (molecular weight 45kDa) on SDS-PAGE (Figure 4.23 B) [section 2.12.4]. Yields of ~200µg of fusion protein/l of bacterial culture were achieved; however, this is much less than the optimal ~15mg/l cited for this expression system (Smith and Johnson 1988).

4.5.2 Preparation of G7b antibodies and Western blot analysis

4.5.2.1 *ELISA analysis*

To raise antibodies, ~100µg of purified Sj26-G7b fusion protein [section 4.5.1] was used to immunise each of two rabbits [section 2.13.1]. The rabbits were each immunised three times, on days 1, 14 & 42, and blood samples were taken on days 1 (prebleed), 28 & 50. The antisera recovered [section 2.13.2] and the purified antibodies [section 4.5.2.3] were characterised by ELISA assays as described in section 2.14.1. Microtitre wells were coated with ~2µg of Sj26-G7b fusion protein, and sera from the prebleed and subsequent samples were added at dilutions between 10^{-2} and 10^{-6}. The measurement of the OD_{405} allowed the titre of the antibodies to be determined. Figure 4.25 shows the OD_{405} measurements for the serum samples tested. For both rabbits there was a negligible response in the pre-immune sera. Both immune sera showed responses that decreased upon dilution and became negligible at a dilution of 10^{-6} (Figure 4.25). In addition, there was no significant difference between the titre of the sera taken after the second and third immunisations (Figure 4.25). The antiserum from one rabbit, 4848, had a slightly higher titre and lower response in the pre-immune serum than those from the other rabbit, 4847, therefore, this antiserum was selected for Western blot analysis.

4.5.2.2 *Western blot analysis of* Sj26-G7b *fusion protein*

In order to confirm that the immune sera from rabbit 4848 contained antibodies

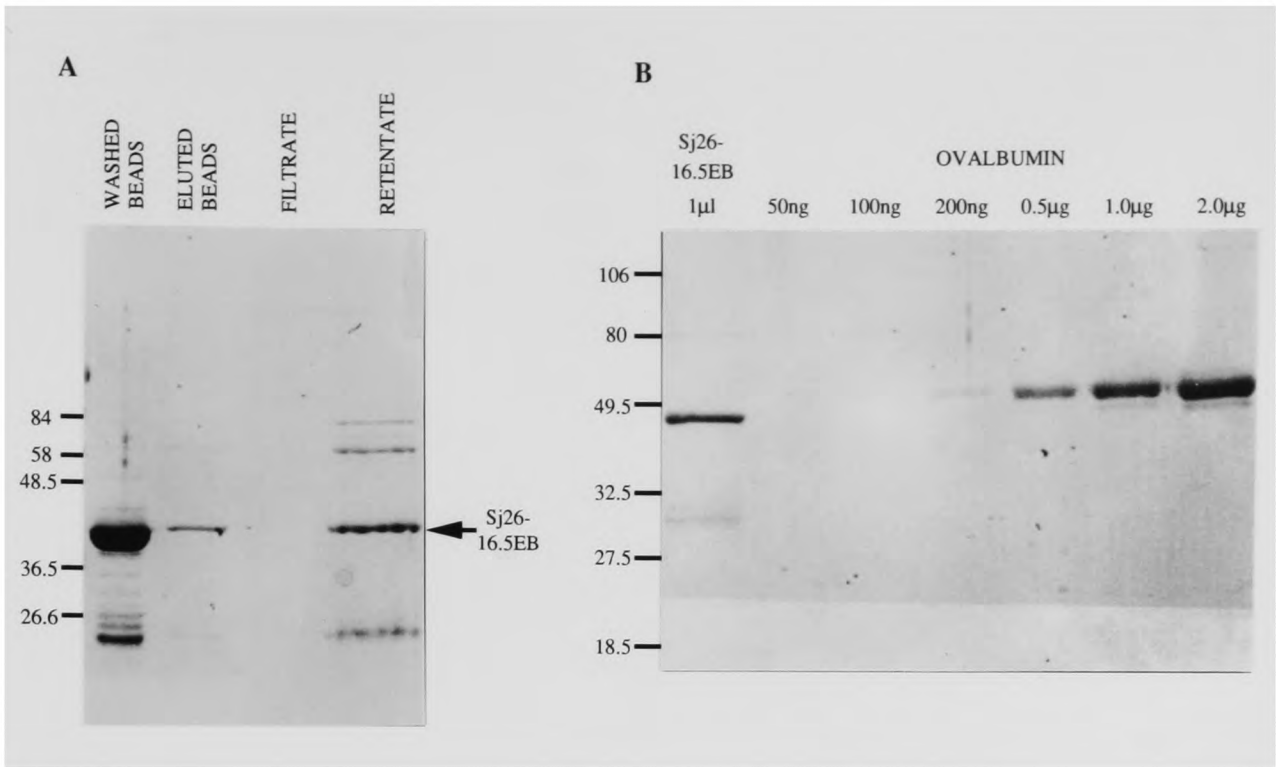


Figure 4.23 Isolation of the G7b Sj26-16.5EB fusion protein.
A The Sj26-16.5EB fusion protein isolated on glutathione agarose beads was eluted off using reduced glutathione. The eluted fusion protein was concentrated by centrifuging in a Centricon-30 column. The filtrate and retentate were analysed by SDS-PAGE on a 10% (w/v) acrylamide gel to confirm the presence of the fusion protein in the concentrated sample before quantitation of the protein sample.The concentration of the fusion protein in the samples was estimated by comparison with known amounts of ovalbumin (molecular weight 45kDa) by SDS-PAGE on 10% (w/v) acrylamide gels (**B**).Sizes are in kDa.

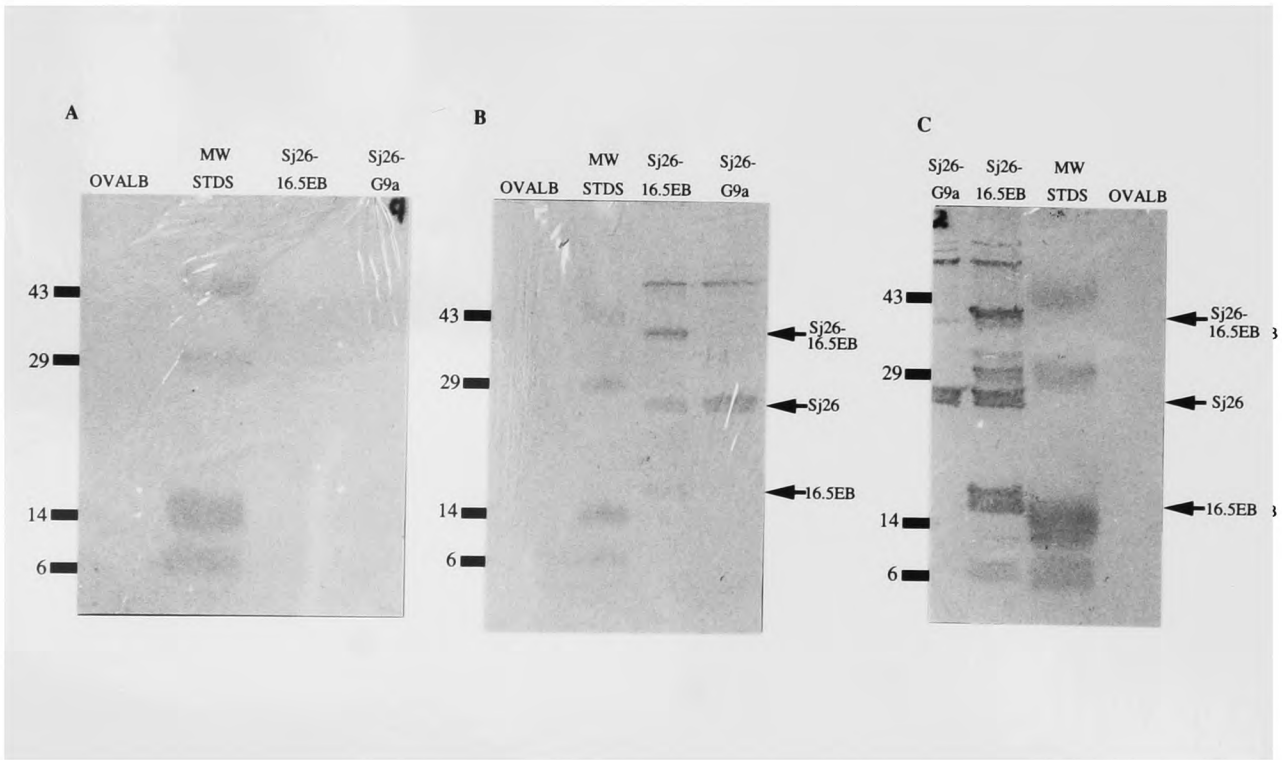


Figure 4.24 Characterisation of anti-Sj26-16.5EB antibodies.
Western blots of the Sj26-16.5EB fusion protein cleaved with thrombin, and negative controls of ovalbumin (OVALB) and Sj26-G9a fusion protein cleaved with thrombin, fractionated on 15% (w/v) acrylamide gels and transferred onto nitrocellulose were used to characterise the anti-Sj26-16.5EB antibodies.The pre-immune serum (**A**) and antisera collected after the second (**B**) and third (**C**) immunisations of rabbit 4848 were used to detect antigen on the Western blots.The pre-immune serum did not detect any antigen. The immune sera detected the 16.5EB antigen and Sj26 from both Sj26-16.5EB and Sj26-G9a. Sizes are in kDa.

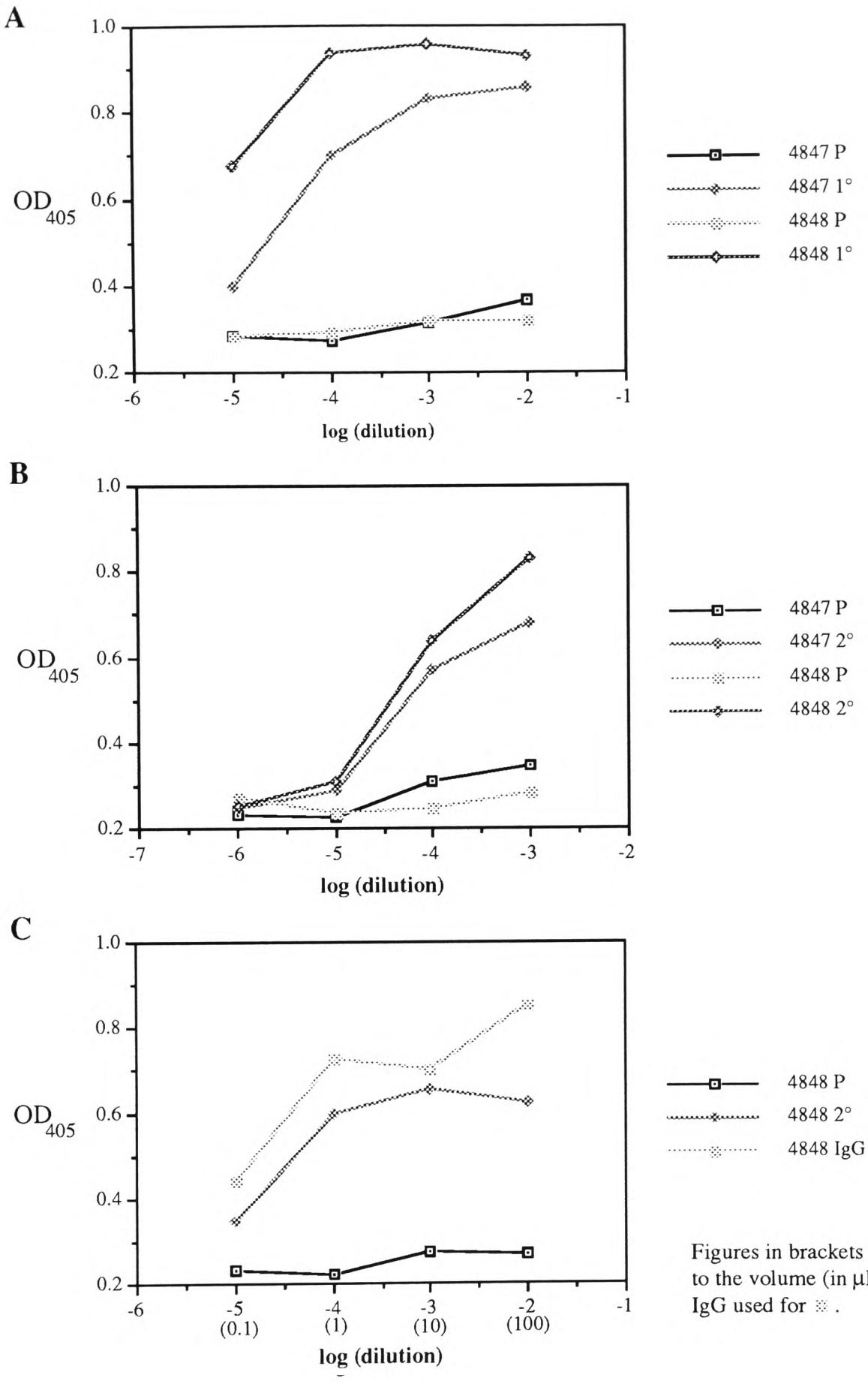


Figure 4.25 Characterisation of G7b antisera by ELISA.
The presence of antibodies directed against the Sj26-16.5EB fusion protein was determined by ELISA. Microtitre wells were incubated with 2µg of the G7b fusion protein before dilutions of the sera were added. Bound antibodies were detected by alkaline phosphatase conjugated anti-rabbit IgG and subsequent incubation for 1 h with substrate solution. The OD_{405} values indicate the degree of binding. **A** Comparison of the first (1°) immune sera with the pre-immune sera. **B** Comparison of the second (2°) immune sera with the pre-immune sera. **C** Comparison of the purified antibodies (IgG) with the second immune serum. A separate key is shown at the side of each graph.

against the G7b polypeptide expressed in pGEX, a set of Western blots were prepared. Fusion proteins Sj26-G7b and Sj26-G9a (provided by CM Milner) were cleaved with thrombin and loaded, with ovalbumin as a negative control, onto a 12% (w/v) SDS–polyacrylamide gel and separated by SDS-PAGE [section 2.12.2]. The proteins were then transferred by Western blotting onto nitrocellulose membranes [section 2.12.3]. The Western blots were incubated with sera, at 1:200 dilution, and bound antibodies were detected by subsequent incubation of the blots with anti-rabbit IgG, followed by NBT/BCIP substrate [section 2.14.2].

No signal was detected with the pre-immune serum (Figure 4.24 A). However, both of the immune sera from rabbit 4848 detected the G7b polypeptide in addition to Sj26 from the Sj26-G7b and Sj26-G9a fusion proteins. Neither of the immune sera detected the G9a polypeptide or ovalbumin (Figure 4.24 B & C). Although the titre of antibodies in the immune sera appeared to be similar by ELISA analysis, Western blot analysis demonstrated that the sample collected after the third immunisation generated a stronger signal. In order to remove any non-specific antibodies, anti-G7b IgG was purified from the immune serum collected after the third immunisation from rabbit 4848, by affinity chromatography.

4.5.2.3.*Affinity chromatography purification of anti-*G7b *IgG*

Two 3.5ml columns were prepared by coupling cell lysates from NM554 cells expressing Sj26-G7b and Sj26-G9a (provided by CM Milner) to CNBr-activated Sepharose. Column 1 contained ~13mg/ml cell lysate, including Sj26-G9a, and column 2 contained ~10mg/ml cell lysate, including Sj26-G7b. 2.5ml of immune serum was diluted to 10ml in PBS/1mM EDTA and passed through column 1 to absorb out antibodies to Sj26 and *E.coli* proteins. From OD_{280} measurements of the serum before and after passage through column 1, ~30% of the protein was bound to the column.

The absorbed serum was then passed through column 2 to isolate the anti-G7b antibodies. Approximately 13mg of protein was absorbed onto column 2. The bound IgG was eluted using 3M $MgCl_2$ and, after dialysis, 6ml of IgG was recovered at ~108µg/ml. ELISA analysis was used to measure the titre of the purified antibodies (Figure 4.25 C).

4.5.2.4 *Western blot analysis of total* U937 *protein*

Western blots of sonicated U937 cell lysates were prepared [section 2.12.3] following

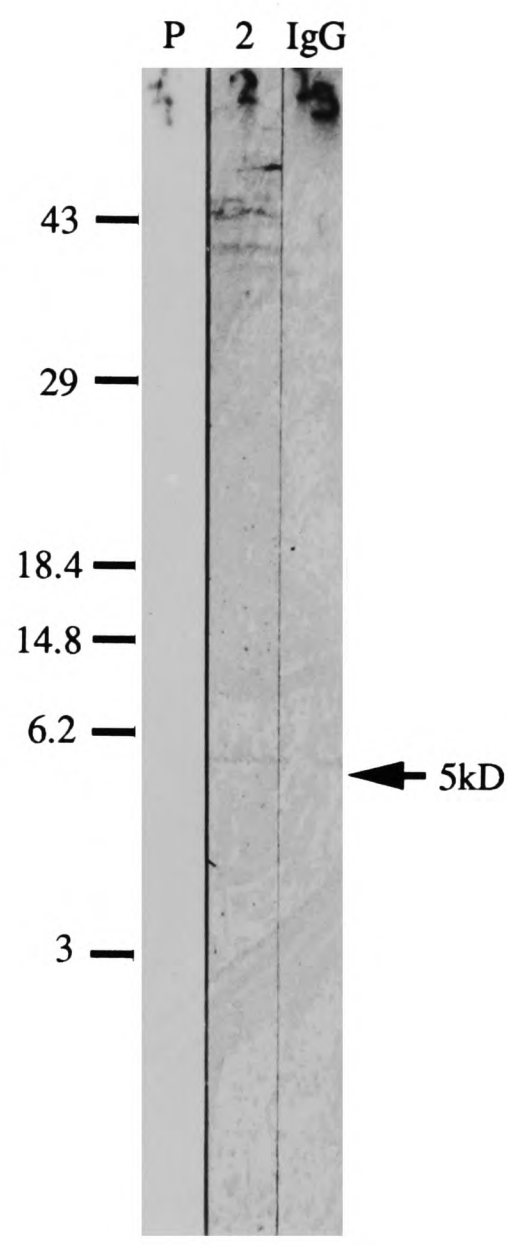


Figure 4.26 Western blot analysis of total U937 protein.
Western blots of U937 cells lysed by sonication were incubated with pre-immune serum (P), second immune serum (2) and IgG purified from the second immune serum (IgG) from rabbit 4848. A specific protein of ~5kDa was detected by both the immune sera. Sizes are in kDa.

SDS-PAGE on a 13% (w/v) polyacrylamide gel. Nitrocellulose strips from this Western blot were incubated with a 1:200 dilution of the pre-immune serum, a 1:200 dilution of the immune serum collected after the third immunisation and a 1:50 dilution of the purified anti-G7b IgG. Following incubation of the blots with anti-rabbit IgG and NBT/BCIP substrate [section 2.14.2], a single protein of ~5kDa was detected with the anti-G7b IgG. A single protein of ~5kDa was also detected by the immune serum, but was not detected by the pre-immune serum (Figure 4.26). In addition, this ~5kDa protein was not detected by antisera raised against the Sj26-G9a fusion protein (provided by CM Milner), indicating that this does not correspond to cross-reaction with a human glutathione S-transferase homologue. This unique product, however, does not correspond to the size of the predicted ~10kDa G7b polypeptide.

4.6 DISCUSSION

4.6.1 The G7b gene

G7b is a newly identified, novel gene in the MHC class III region that lies between the HSP70-Hom and G7a, valyl-tRNA synthetase (VARS2), genes. The G7b gene spans ~10kb and consists of five exons that are clustered into two groups, separated by an ~8kb intron. The reason for the uneven distribution of the exons is unclear.

Northern blot analysis has shown that the G7b gene encodes a ~0.9kb mRNA that is ubiquitously expressed and may be more abundant in monocytes and macrophages. In addition to the G7b gene in the MHC class III region, the G7b cDNA sequence was also shown to cross-hybridise to additional DNA sequences indicating that there are sequences homologous to G7b elsewhere in the human genome, although none of the homologous sequences have so far been characterised.

4.6.2 Western blot analysis using anti-G7b antibodies

The putative G7b polypeptide has been expressed as a fusion protein with glutathione S-transferase in *E.coli* and used to raise antibodies in rabbits. These antibodies detected the G7b polypeptide expressed in *E.coli* by ELISA and Western blot analysis and on Western blots of total U937 cell lysates these antibodies detected

an antigen of molecular weight ~5kDa. However, the predicted G7b protein is ~10kDa in size. Possible explanations for this discrepancy are:

i) Polyclonal antibodies raised against the Sj26-G7b fusion protein may include some with specificity for the peptide encoded by the 5′ untranslated region, between Sj26 and the 95 amino acid G7b polypeptide in the fusion protein. These antibodies may then cross-react with an unknown ~5kDa protein. However, the database search using the complete cDNA translation product [section 4.4.2] did not reveal any similarity between this part of the sequence and any protein in the database. As a result this explanation, although not impossible, is improbable.

ii) The anti-G7b antibodies cross-react with a smaller, homologous ~5kDa protein, that is possibly encoded by one of the regions of the genome which cross-hybridised to the G7b cDNA probe [section 4.2.3]. If this were the case, then it might be expected that multiple, cross-reacting proteins would be detected.

iii) The ~5kDa protein is a product of the G7b polypeptide that has undergone post-translational processing or degradation. Although there are no obvious proteolytic cleavage sites within the G7b polypeptide that might generate a ~5kDa product, multiple processing events may generate a product with an apparent size of ~5kDa.

4.6.3 Functions of the G7b protein and possible roles in disease aetiology

The 95 amino acid putative G7b polypeptide sequence does not contain any regions that are particularly hydrophobic or highly charged. As a result, this protein is probably localised in the cytosol or, because of its small size, it could move freely into the nucleus (von Heijne 1988).

The G7b polypeptide sequence includes two similarities of interest. The first is an 18 amino acid sequence from residues 7-24 that shares similarity with members of the annexin family of proteins. Although the sequence is short, the similarity is high when compared with all the annexin sequences. All the residue changes are conservative, or occur where the annexin family members diverge in sequence from each other. As a result, the G7b sequence probably shares structural similarity with the annexins where their corresponding sequence forms a helix-turn-helix on the outer surface of the protein.

The second similarity is with the small nuclear ribonucleoprotein Sm-D and is high enough to predict, from the ALIGN score, that the two proteins may share a

similar structure, excluding the Gly-Arg repeating structure at the C-terminus of Sm–D. This similarity with Sm-D is of particular interest with respect to the possible association of SLE with the MHC. Patients with SLE contain autoantibodies against many nuclear proteins, including Sm-D, and, assuming that the structural similarity shared between G7b and Sm-D reflects a functional similarity, G7b may also be a component of snRNPs involved in RNA splicing that becomes an autoantigen in SLE patients.

There is also similarity between another SLE autoantigen, HMG-1, and the G9a protein (Milner and Campbell 1993) and, to a lesser extent with the predicted G7 protein product. The association of SLE with the MHC could be due to the fact that genes which encode several autoantigens are closely linked such that they behave as a single locus in genetic analysis due to the strong linkage disequilibrium within the MHC.

CHAPTER 5

EVIDENCE FOR A NEW GENE, G7c

5.1 INTRODUCTION

The central region of the MHC has been shown to encode loci responsible for several genetic traits. The TNP-ficoll response control locus was originally mapped to the region between *H-2S* and *H-2D* in the mouse MHC (Hillstrom-Shapiro *et al.* 1985) and has subsequently been shown to lie between the complement gene cluster and the *Bat5* locus (Snoek *et al.* 1991). The major haemopoietic histocompatibility antigen regulatory locus, *Hh-1*, was mapped telomeric of *H-2S* and centromeric to *H-2D* in the mouse (Rembecki *et al.* 1988), within a region that corresponds to the telomeric end of the human MHC class III region. In man, the *EC1* locus, controlling natural killer cell alloreactivity, was originally mapped to the region between the complement factor B gene and HLA-A (Ciccone *et al.* 1990) Recently this locus has been shown to be strongly associated with HLA-C (Colonna *et al.* 1992b). The locus controlling corticosteroid induced cleft palate susceptibility, *Cps-1*, was initially associated with the mouse *H-2* region (Demant 1985) and has subsequently been mapped, by recombination, to the *Hsp70–Bat5* interval (Snoek *et al.* 1991). The genetic control of susceptibility to induction of autoimmune orchitis in the mouse was originally mapped towards the *H-2D* region by Teuscher *et al.* (1985). By recombination mapping, the locus, *Orch-1*, was mapped, initially between *H2-S* and *Tnfa* (Teuscher *et al.* 1990), then to the same *Hsp70–Bat5* interval as *Cps-1* (Snoek *et al.* 1991), and more recently to the ~60kb *Hsp70.1–G7* region (Snoek *et al.* 1993).

In the human MHC, the region from the HSP70 gene cluster to BAT5 has been shown to encode eight genes; HSP70-2, HSP70-1, HSP70-Hom, G7b, G7a, G7, G6 and BAT5 (Milner and Campbell 1990; Chapter 4; Hsieh and Campbell 1991; Sargent *et al.* 1989a, Spies *et al.* 1989a). Although sequence data is available for the predicted protein products of all these genes (Milner and Campbell 1990; section 4.4; Hsieh and Campbell 1991; section 3.5; Cheng 1991), except BAT5, there is no indication as to whether or not any of these genes are the *Cps-1* or *Orch-1* loci. Thus, the possibility

that there are other genes within the HSP70-Hom to BAT5 interval was investigated.

To identify coding sequences within a relatively short region of genomic DNA, such as the 11kb between G7 and G7a, two types of approach may be used.

In the first approach, the genomic DNA fragment can be used as a hybridisation probe in order to determine the tissues in which the coding sequences are expressed. This can be carried out by Northern blot analysis or by screening for clones in a cDNA library [section 5.2.2]. However, if expression of the coding sequences is restricted to a limited number of tissue types, or a specific stage in development, or the mRNA species has low stability or is of low abundance, then such an approach may be unsuccessful.

The second approach involves analysis of the genomic sequence itself to identify the coding sequences, thereby avoiding the problems associated with detecting expression of the sequences. This can either be carried out by computational analysis of nucleotide sequence data, or by *in vitro* isolation of coding sequences.

If computational analysis is to be carried out, then the nucleotide sequence data must first be collected, either from the whole region or from selected DNA fragments *e.g.* those devoid of repetitive sequences [section 5.2.1]. The determination of genomic DNA sequence will generate a large amount of data from non-coding regions, but selected fragments will not necessarily include all of the coding regions and repetitive sequence elements can occur throughout the sequence.

Even with the complete sequence available, the coding, exonic sequences must still be identified by searching for intron-exon boundaries and the more systematically complex coding regions (Brunak *et al.* 1991). Currently, using this technique, not all exons are identified and false positives are often detected. However, as more sequence is analysed, the base patterns characteristic of coding sequences should, in the future, become easier to identify.

The techniques for *in vitro* isolation of coding sequences are generally referred to as exon trapping. These experimental methods make use of the functional identification of *cis*-acting elements required for RNA splicing. All of the methods described utilise vectors that contain splicing signals that can be used to detect complementary splice signals in genomic DNA fragments inserted into the vector.

The exon trapping technique of Duyk *et al.* (1990) utilises the donor splice site from human β globin exon I to search for downstream 3′ splice acceptor, sites. This

technique involves retroviral packaging of constructs before infection of COS-7 cells and transformation of bacteria before screening can take place using colour selection which detects the removal of the *E.coli* α-complementing factor of β–galactosidase during splicing. In addition, a high frequency of cryptic splice acceptor sites are identified using this technique.

Other exon trapping techniques described are, exon trap cloning (Auch and Reth 1990), exon amplification (Buckler *et al.* 1991) and, more recently, the specific exon trapping system, SETS, (Hamaguchi *et al.* 1992). All of these techniques utilise both donor and acceptor splice sites in the vector to screen for both acceptor and donor sites flanking coding sequences in the genomic DNA inserts. The exon trap cloning system utilises exons II and III from the rat preproinsulin gene in the pL53ln vector but the efficiency of this system has not been reported (Auch and Reth 1990). The exon amplification system utilises the HIV *tat* intron to replace an intron from rabbit β globin which is essential for expression of the rabbit β globin cDNA (Buchman and Berg 1988). The pSPL1 vector may also generate false positive exons [section 4.3.7] (Buckler *et al.* 1991) although the accuracy of splicing is not known. The most recent, specific exon trapping system (SETS), of Hamaguchi *et al.* (1992) utilises the p53 gene intron 10 and its flanking exons, selected for accuracy of splicing and ease of recognition of spliced products, in the pMHC2 vector which appears to offer high accuracy and sensitivity.

Based upon the principles that the coding sequences of genes are non-repetitive and conserved across species, a putative new gene has been identified. Using exon amplification (Buckler *et al.* 1991) to analyse part of the putative gene sequence, two short, non-repetitive, conserved sequences were identified that encode ORFs in two phases. Neither of these sequences shares similarity with any protein sequence in the databases.

5.2 IDENTIFICATION OF POSSIBLE LOCATIONS FOR NEW GENES

5.2.1 Mapping non-repetitive DNA sequences between HSP70-Hom and BAT5

The human HSP70-Hom–BAT5 interval spans ~100kb, of which ~75kb was isolated

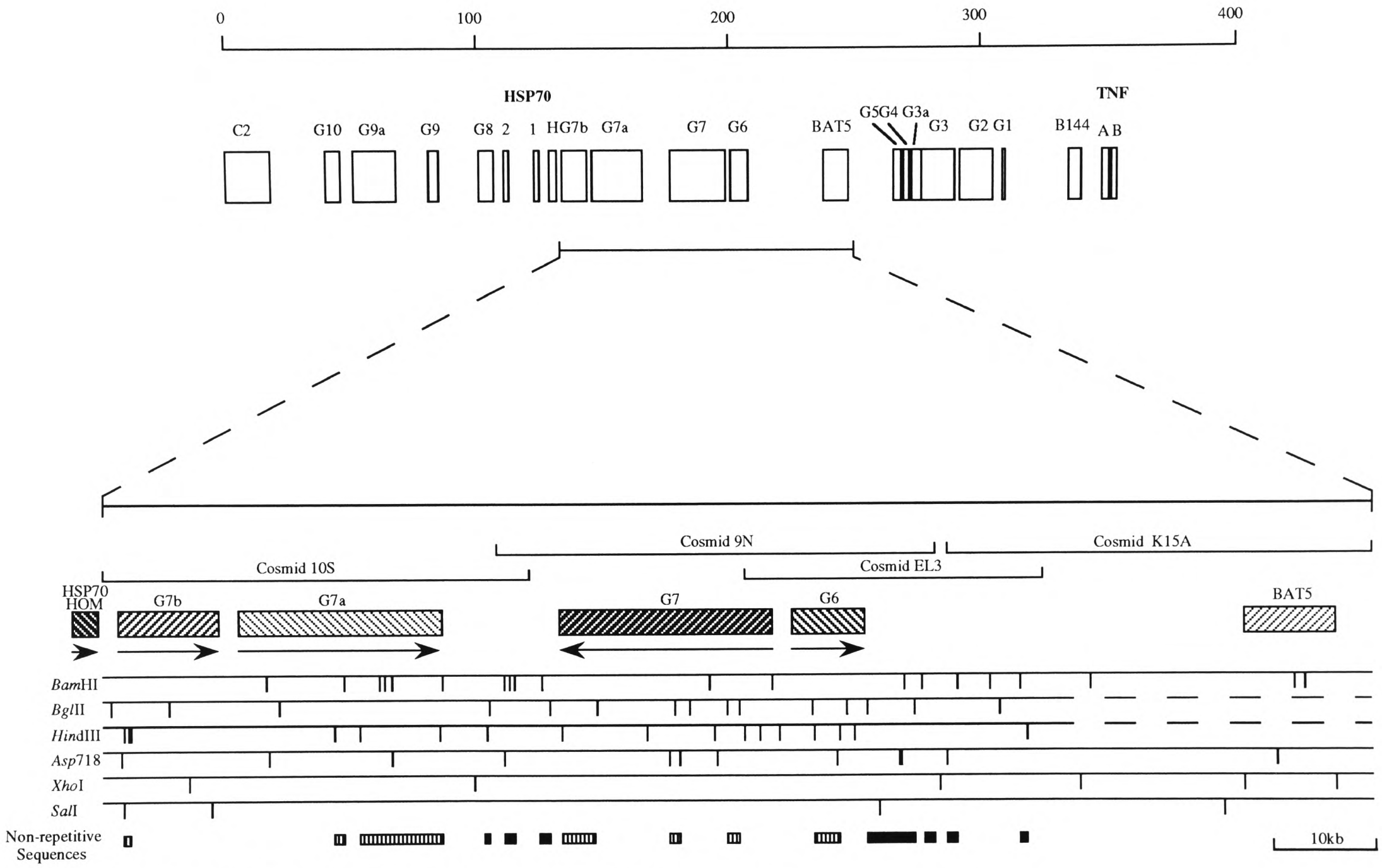


Figure 5.1 Molecular map illustrating the region between HSP70-Hom and BAT5 in MHC class III region.
The upper half of this figure shows the locations of the genes mapped between the C2 and TNF genes. The genes are shown as open boxes and the region corresponding to the genomic inserts of cosmids 10S, 9N, EL3 and K15A is shown as a horizontal bar. The region contained in the cosmid inserts is expanded in the lower half of the figure. The genes within this region are shown as hatched boxes with arrows beneath to indicate their direction of transcription.The positions of the restriction sites mapped in the region and the locations of non-repetitive sequences are shown . Non-repetitive sequences associated with genes already mapped in the region are shown by striped boxes and those within the intergenic regions are shown as filled boxes.

in overlapping cosmid clones by Sargent *et al.* (1989a). Within these clones, the precise limits of the genes HSP70-Hom (Milner and Campbell 1990), G7b [section 4.3], G7a (Hsieh and Campbell 1991), G7 [section 3.3] and G6 (Cheng 1991) have been mapped, and the location of BAT5 was mapped in cosmids telomeric of G6 (Spies *et al.* 1989a) (Figure 5.1).

During the isolation of overlapping cosmid clones by Sargent *et al.* (1989a), walking probes were isolated by identifying non-repetitive sequences that were linked to preexisting probes from the MHC class III region. The non-repetitive sequences were identified by probing Southern blots of digested cosmid DNA with total human genomic DNA, radiolabelled by nick translation. Those fragments not hybridised by total human genomic DNA, devoid of repetitive sequences, were identified by comparing the autoradiographs with photographs of ethidium bromide stained agarose gels (Sargent 1988) and the locations of these non-repetitive sequences were then incorporated into restriction maps of the cosmids (CA Sargent unpublished). The non-repetitive fragments in the cosmids from HSP70-Hom to cosmid EL3, which extends closest to BAT5, are included in the restriction map for this region (Figure 5.1).

Examination of the restriction map (Figure 5.1) revealed a correlation between the locations of the genes and the non-repetitive sequences. All the genes contain at least one DNA fragment devoid of repetitive sequences. However, there are also two intergenic regions of ~11kb between G7 and G7a, and ~17kb between G6 and the end of cosmid EL3, that contain regions of non-repetitive DNA. The other intergenic regions are only ~2kb in length. Although genes of 1kb or less *e.g.* G1 (Olavesen *et al.* 1993) could lie in these regions, none of the non-repetitive sequences occur there.

As a result, it is possible that new genes lie in the G7–G7a interval and between G6 and BAT5. Since the *Orch-1* locus has been mapped to the more restricted, *G7–Hsp70.1* region, it was decided to analyse the G7–G7a gene interval in more detail.

5.2.2 Characterisation of the G7–G7a intergenic region

One strategy for the identification of genes within genomic DNA, has been to demonstrate that sequences in the region of interest are conserved across species (Monaco *et al.* 1986; Rommens *et al.* 1989). To show this, potential coding sequences

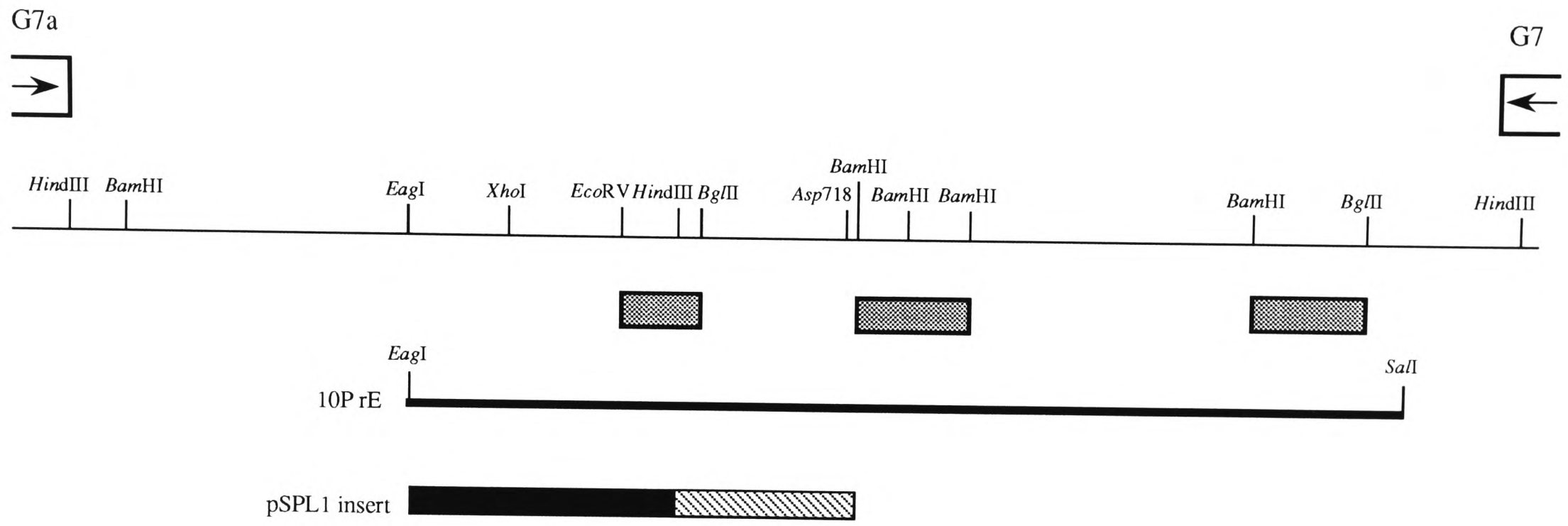


Figure 5.2 Molecular map of the intergenic region between G7 and G7a.
The region of DNA between G7 and G7a is shown with the positions of the restriction sites in the region. The limits of the G7 and G7a genes are shown with arrows to indicate their direction of transcription. The non-repetitive sequences mapped within this region are shown as shaded boxes. The genomic DNA insert of the religated cosmid subclone 10P rE is shown as a horizontal bar with the locations of restriction sites for *Eag*I and *Sal*I which flank the insert. The genomic DNA insert subcloned into pSPL1 is shown as a filled box and the region deleted in plasmid construct pSPL1-ME2.5 is hatched.

are hybridised to Southern blots of digested DNA from different animal species; often referred to as zoo blots. By washing the filters at low stringency, cross-hybridising DNA fragments are detected if homologous sequences are present in the DNA from the different animal species.

In order to isolate an intergenic DNA fragment that does not contain any coding sequence from G7 or G7a, cosmid 10P (Sargent *et al.* 1989a), which lies ~100bp centromeric of the 3' end of G7, was religated at the *Eag*I sites at the telomeric end of the insert and at one end of the cosmid vector. The religated subclone, 10P rE, contained an 8.8kb insert of genomic DNA sequence that does not contain any sequences from G7 or G7a, but includes three regions of non-repetitive DNA (Figure 5.2).

To investigate whether or not the G7-G7a intergenic sequence contained any conserved sequences, the 8.8kb genomic insert was excised by digestion with *Eag*I and *Sal*I and isolated, following separation from the vector sequence, on a LGT agarose gel [section 2.4.3]. The genomic insert was radiolabelled [section 2.6.1] and hybridised, under competitive conditions with 250µg/ml placental DNA [section 2.6.2], to a Southern blot of *Bam*HI digested genomic DNA from man, mouse, cat, sheep, rabbit, chicken and shark. The filter was washed, twice, for 30 minutes each, at 65°C with 2x SSC/0.1% (w/v) SDS. The intergenic DNA fragment cross-hybridised to multiple DNA fragments from man, mouse, cat, sheep, and rabbit, but not chicken or shark (Figure 5.3). This result indicates that there are conserved DNA sequences in the G7–G7a intergenic region that may correspond to the coding regions within exons from a new gene in this region.

In order to investigate whether these conserved DNA sequences are transcribed into a mRNA species, the 8.8kb *Eag*I/*Sal*I DNA probe was hybridised, under competitive conditions with 250µg/ml placental DNA [section 2.6.2], to a Northern blot of total RNA from liver (HepG2), B lymphocyte (Daudi), T lymphocyte (Molt4) and monocyte (HL60 and U937) cultured cell lines and testis tissue (a gift from YH Edwards, MRC Human Biochemical Genetics Unit, London) [section 2.5.2]. Following autoradiography, for three weeks, a mRNA species of 2-3kb was just detectable in all the samples (result not shown).

Subsequently, a cDNA library in CDM8 (a gift from D Simmons, Institute of Molecular Medicine, Oxford), constructed using mRNA derived from human foetal

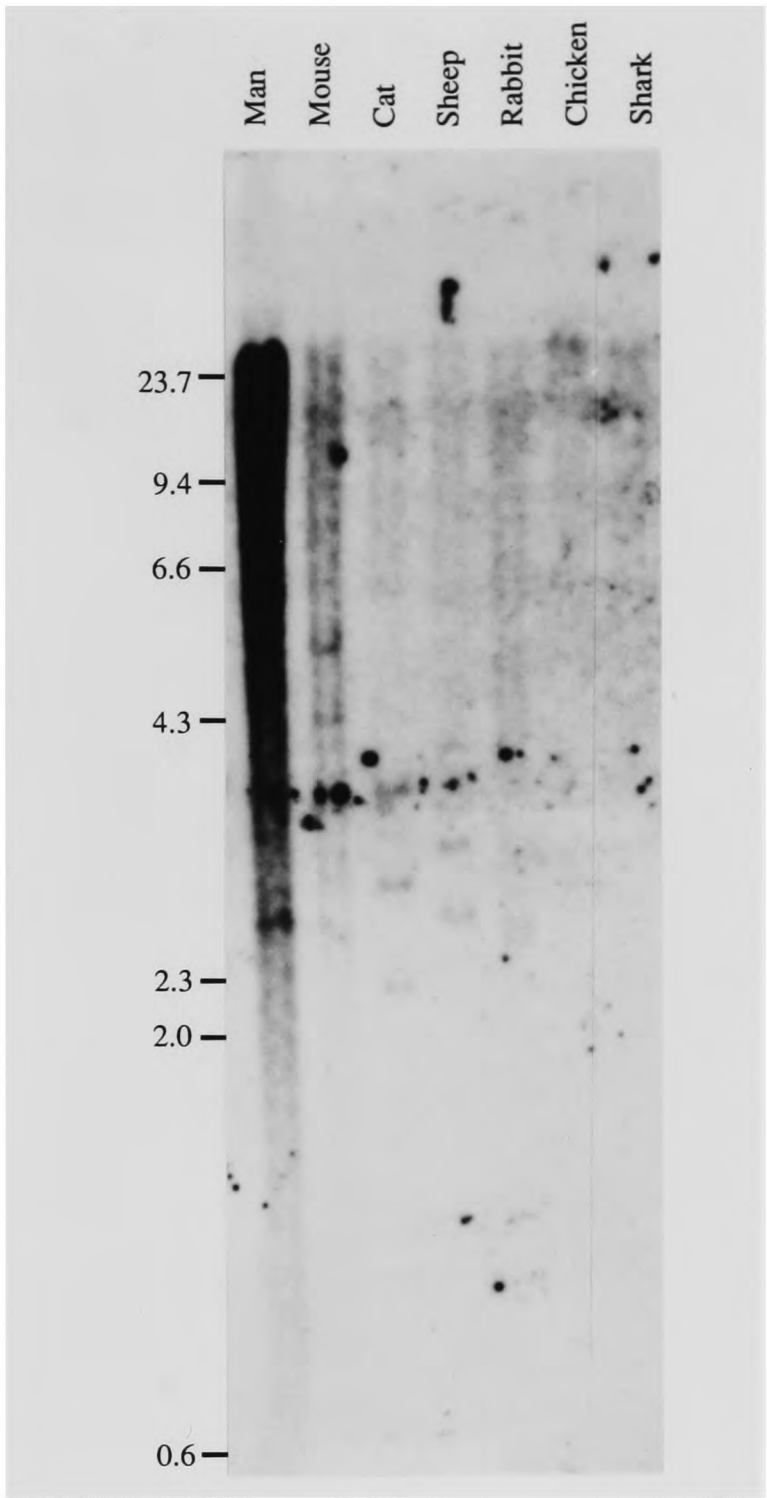


Figure 5.3 Zoo blot analysis of the G7-G7a intergenic region.
Total genomic DNA from the animal species shown digested with *Bam*HI, fractionated on a 0.8% (w/v) agarose gel and transferred onto nitrocellulose. The Southern blot was hybridised with the 8.8kb *Sal*I/*Eag*I genomic DNA insert from cosmid subclone 10P rE competed with 250µg/ml placental DNA. The blot was washed at low stringency at 65°C with 2xSSC/0.1% (w/v) SDS. The probe cross-hybridised to specific DNA fragments from mouse, cat, sheep and rabbit DNA but not chicken or shark DNA. Sizes are in kb.

brain, which expresses a large number of low abundance transcripts (Adams *et al.* 1991), cloned in CDM8, was screened using the 8.8kb *EagI/SalI* DNA probe competed with 250μg/ml placental DNA [section 2.8]. No duplicating positive clones were detected from the $5x10^5$ colonies screened.

5.3 IDENTIFICATION OF THE PUTATIVE GENE, G7c

5.3.1 Use of exon amplification to identify coding sequences in the putative gene, G7c

In order to isolate coding sequences from the G7–G7a interval a specific fragment from the region was subcloned into the pSPL1 vector (Buckler *et al.* 1991) for exon amplification. From the religated cosmid subclone, 10P rE [section 5.2.2], an end-filled [section 2.4.4], 3.8kb *BamHI/EagI* insert (Figure 5.2) was isolated and cloned into *HincII* cut pBC [section 2.7]. The insert was then excised using *EagI*, end-filled [section 2.4.4], and ligated directly into *BamHI* cut, end-filled, pSPL1 [section 2.7]. All the recombinants contained the insert in only one orientation and all contained partial deletions of the 3.8kb insert (Figure 5.2). The subclone with the smallest deletion, of 1.3kb, pSPL1-ME2.5, which retained part of the non-repetitive sequence, was used for electroporation into COS-7 cells [section 2.11.1]. A construct containing the 4.6kb *SalI/BglII* DNA fragment including exon II of the G7b gene (Figure 4.14) was used as a positive control and COS-7 cells electroporated in the absence of DNA were used as a negative control.

After recovery, the cells were harvested and RNA was isolated for RT-PCR [section 2.11.2]. Analysis of the amplified products was carried out by agarose gel electrophoresis on a 1.2% (w/v) gel [section 2.4.2]. The results showed that the first round of amplification produced products from all the RNA samples, including the negative control (Figure 5.4 A). In the second round of PCR reactions, more specific amplification did not amplify any product from the negative control, but the G7c sample produced three major products of ~100, 210 and 320bp and the G7b sample produced two major products of ~160bp and 400bp (Figure 5.4 A). The ~100bp fragment from G7c corresponded to the expected size of product if there was no amplified exon present. The initial, non-specific, ~460bp PCR product amplified from the COS-7 cells indicated that the SD2 and SA2 primers may amplify a β globin

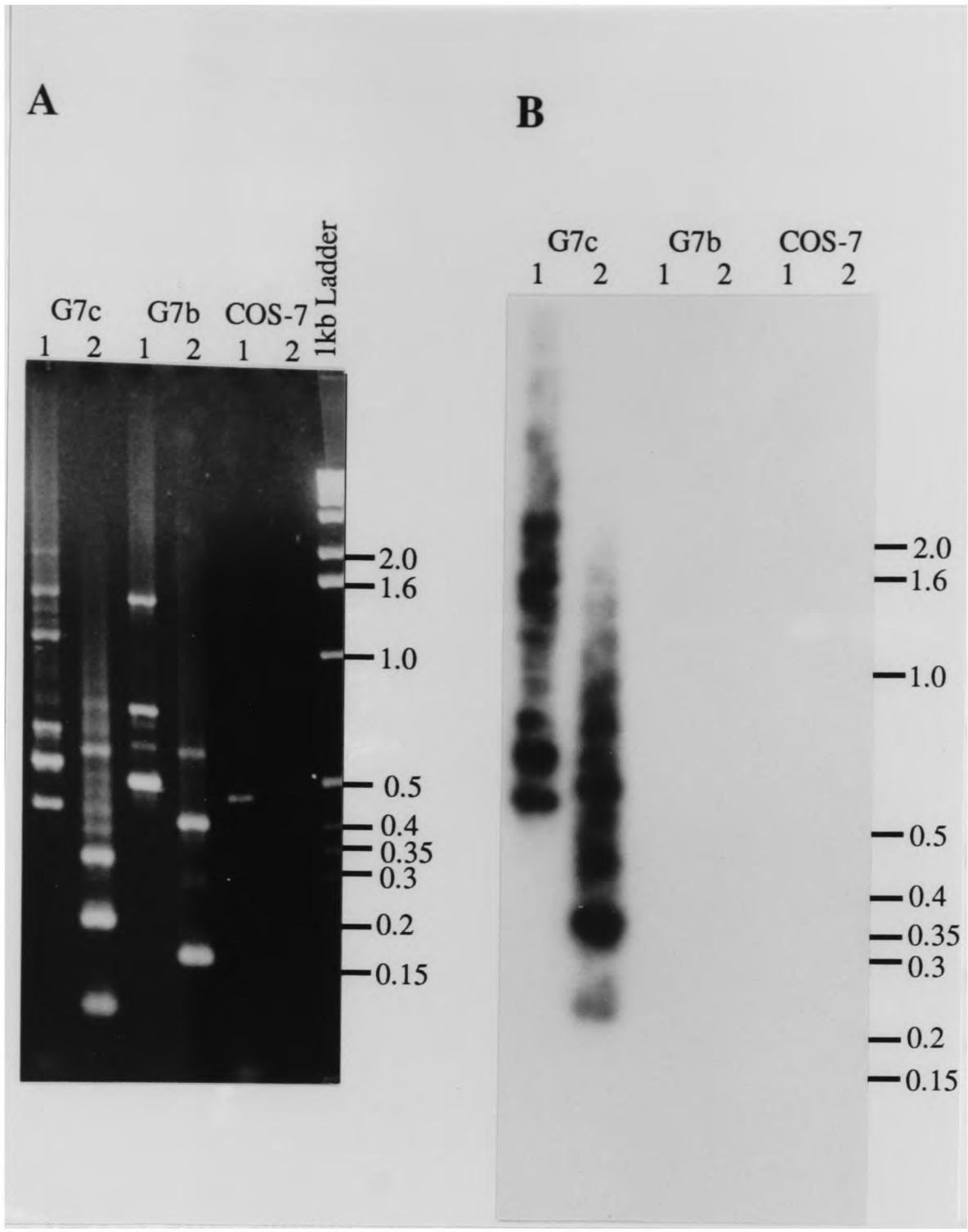


Figure 5.4 Exon amplification analysis of the G7-G7a intergenic region.
RT-PCR was carried out using total RNA isolated from the cells transfected with pSPL1 plasmid constructs containing DNA from G7c and G7b and COS-7 cells electroporated in the absence of DNA. The first round of PCR (1) was carried out using oligonucleotide primers SD2 and SA2 for 35 cycles of 94°C for 1 min, 58°C for 2 min and 72°C for 3 min. The second round of PCR (2) was carried out using oligonucleotide primers SD1 and SA1 for 20 cycles of 94°C for 60 s, 58°C for 60 s and 72°C for 90 s.
A The reaction product DNA was fractionated on a 1.2% agarose gel. Specific products are present in the samples containing RNA from cells transfected with the pSPL1 constructs. A single non-specific product was amplified from the sample containing RNA from electroporated COS-7 cells with the SD2 and SA2 primer pair but not with SD1 and SA1.
B Southern blot of DNA from the PCR reactions hybridised with the genomic DNA insert from the G7c pSPL1 construct competed with 50µg/ml placental DNA. Specific products only from the cells transfected with the G7c pSPL1 construct are detected. following two rounds of PCR, products of ~210bp and 320bp are detected. Sizes are in kb.

related sequence in the first round of PCR amplification; however, this product is not amplified by the SD1 and SA1 primers from the HIV *tat* sequence. Buckler *et al.* (1991) observed an ~300bp product that persisted in the second round of PCR reactions, but the use of an oligo dT primer for reverse transcription, in place of the SA2 primer, may have prevented amplification of this product.

To confirm that the amplified G7c products are derived from the genomic DNA insert of the religated cosmid subclone, 10P rE, the DNA fragments were Southern blotted onto nitrocellulose [section 2.5.1] and hybridised with the ~3.8kb EagI insert isolated from pBC, competed with 50µg/ml placental DNA [section 2.6]. This analysis detected only the G7c-specific products, which included the ~210 and ~320bp products (Figure 5.4 B). The ~100bp product was apparently not hybridised by the G7c genomic DNA insert, but this may have been due to poor transfer of the DNA which occurs when very small DNA fragments are Southern blotted.

Although the genomic probe from G7c (Figure 5.4 B) does not cross-hybridise with the different exon amplification products from the G7b gene, there is a significant amount of background, due to non-specific amplification. This may be caused by the presence of partially processed transcripts or by less efficiently spliced exons that result in the amplification of a heterogeneous population of minor products.

5.3.2 Mapping the G7c exon amplification products in the cosmid subclone, 10P rE

In order to map the major amplified products within the religated cosmid subclone, 10P rE, the three major exon amplification products [section 5.3.2] were isolated from LGT agarose [section 2.4.3] and hybridised to Southern blots of digested, 10P rE DNA. The results showed that the 100bp product does not hybridise to the religated cosmid subclone 10P rE DNA and is, therefore, not derived from G7c, confirming that it is probably derived from mRNA transcripts that did not contain any trapped exons, only the HIV *tat* exonic sequences.

The ~210 and 320bp G7c products both hybridised to the same restriction fragments in the religated cosmid subclone 10P rE, and mapped to an ~1.5kb region in the genomic DNA insert (Figure 5.5 E). However, subsequent cloning and sequencing of the ~210bp product revealed that there were two different amplified DNA inserts, A and B, of approximately the same size, present within this product that were spliced together in the ~320bp product insert [section 5.3.3]. The two

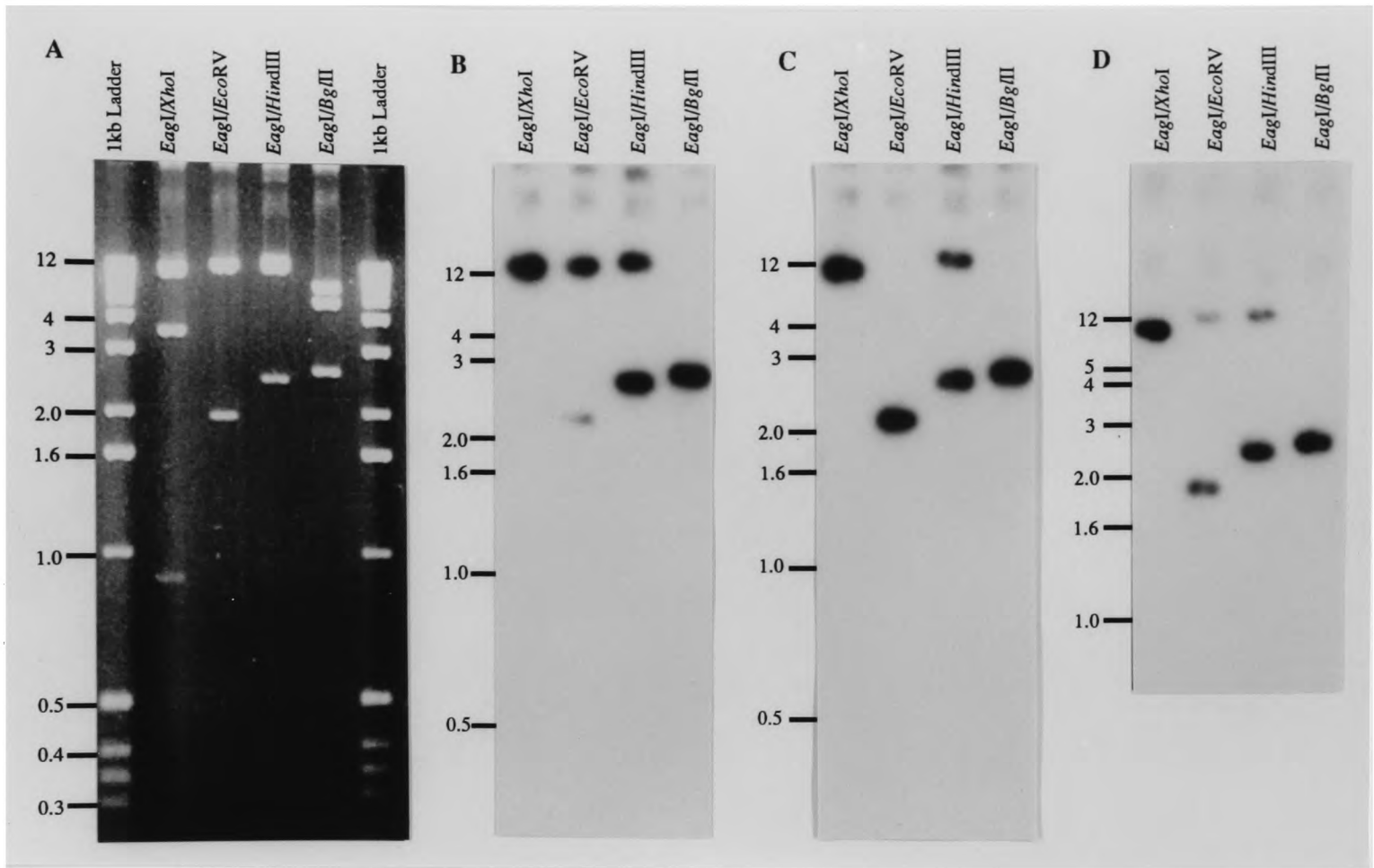


Figure 5.5 Mapping G7c exon amplification products within cosmid subclone 10P rE.
A Digested cosmid subclone 10P rE DNA fractionated on a 0.8% (w/v) agarose gel. Southern blots of cosmid subclone 10P rE were hybridised with the 5' insert from the ~210bp product (**B**) and the 3' insert from the ~210bp product (**C**) and the ~320bp exon amplification product (**D**). Sizes are in kb. **E** The genomic DNA insert of cosmid subclone 10P rE is shown as a horizontal line and restriction sites for *Eag*I (E), *Xho*I (X), *Eco*RV (V), *Hin*dIII (H), *Bgl*II (B), *Asp*718 (A), *Bam*HI (M) and *Sal*I (S) are indicated. The locations of the non-repetitive sequences mapped within this region are shown as open boxes. The G7c pSPL1 insert is shown as a horizontal bar with a striped region to shown the extent of the deletion. The region hybridised by the ~320bp PCR product (ET320) is shown as a hatched box and the regions hybridised by the 3' insert of the ~210bp product (ET210 3') and the 5' insert of the ~210bp product (ET210 5') are shown as shaded boxes. The approximate locations of the two exons A and B are shown as filled boxes and the direction of transcription of the G7c gene is indicated by an arrow.

different DNA inserts were excised from the double-stranded, replicative form of the M13mp10 clones by digestion with *Eco*RI and *Hin*dIII and isolated as probes from LGT agarose [section 2.4.3]. These fragments were hybridised to Southern blots of digested cosmid subclone 10P rE DNA (Figure 5.5 B & C).

These results showed that the two inserts hybridised to different fragments from 10P rE and were mapped to the regions shown in figure 5.5 E. The sequence data from one of the products revealed an *Eco*RV site within the insert and this correlates with the mapping data.

Using oligonucleotide primers that hybridise to the two different ~210bp fragments, and which are oriented towards each other, PCR has been carried out using both cDNA and genomic DNA templates. The results showed that the product from the genomic DNA template is ~200bp larger than that from the spliced product (MR Albertella, personal communication), indicating that the inserts from the two ~210bp sequences are separated by an intron of ~200bp [section 5.3.2]. As a result the two exons can be positioned within the religated cosmid subclone, 10P rE, as shown in figure 5.5 E.

5.3.3 Nucleotide sequence analysis of the exon amplification products

The products of the PCR amplification from the G7c and G7b genes using the SD1 and SA1 primers [section 5.3.1] were digested with *Bam*HI and *Sal*I, restriction sites for which occur in the primers, in preparation for subcloning. The fragments were end-filled [section 2.4.4], fractionated on a 3.5% (w/v) polyacrylamide gel [section 2.4.5] and the ~210bp and ~320bp, major products were isolated [section 2.4.6]. These DNA fragments were subcloned into *Sma*I cut M13mp10 [section 2.7] and the resultant white plaques were used to prepare single-stranded DNA [section 2.9.1]. These templates were sequenced using the M13 universal primer [section 2.9.3].

Examination of the nucleotide sequence confirmed that the inserts contained the HIV *tat* splice junctions and unknown 'trapped' sequences. The ~160bp G7b amplification product contained the G7b exon II sequence (Figure 5.6) and an ~300bp *Bam*HI insert from the ~400bp G7b product contained the G7b exon II sequence spliced to ~150bp of unknown sequence corresponding to the 5′ monomer of an *Alu* repeat. This sequence included an *Eag*I restriction site which corresponds to the site located at the very 3′ end of the G7b genomic DNA insert. In addition the insert

A

*Bam*HI
GGATCCGCGACGAAGACCTC CTCAAGGCAGTCAGACTCAT CAAGTTTCTCTATCAAAGCA

G7b Exon II
CTCTTCTATTCTTTTTTCAA GTCCCTTGTGGGCAAGGATG TGGTCGTGGAACTAAAGAAT

*Sal*I
*GACCTGAG*ACCCACCTCCCA ATCCGAGGGGACCCGACGTC GAC 163

B

*Bam*HI
GGATCCGCGACGAAGACCTC CTCAAGGCAGTCAGACTCAT CAAGTTTCTCTATCAAAGCA

G7b Exon II
CTCTTCTATTCTTTTTTCAA GTCCCTTGTGGGCAAGGATG TGGTCGTGGAACTAAAGAAT

*Eag*I
GACCTGAG/AGACAGAGTCTC GGCCGTGTTGCCCAGGCTGG TCTTGAACTCCTGGACTCAA

GTGATCCTCCCACCTCAGCC TCTCAAAGTGCAGTGACTAC AGGTGTGAGCCAATATGTGT

*Hin*dIII
GTGGCTTTAAAGTGACTTTT CTGTTGGGATGAGAAAGAAG ATCatcaagcttatcg/gat cc 302
*Bgl*II/*Eco*RV *Cla*I/*Bam*HI

Figure 5.6 Nucleotide sequence of G7b exon amplification products.
A Nucleotide sequence of the ~160bp G7b exon amplification product. **B** Nucleotide sequence of the 302bp *Bam*HI fragment from the ~400bp G7b exon amplification product. Flanking sequence from the HIV *tat* intron is underlined. The G7b exon II sequence is shown in italics and the pBC vector and pSPL1 intronic sequences are shown in lower case. Restriction sites are shown above the sequence with the recognition sequence overlined and partial restriction sites remaining from synthesis of the plasmid construct are shown below the sequence.

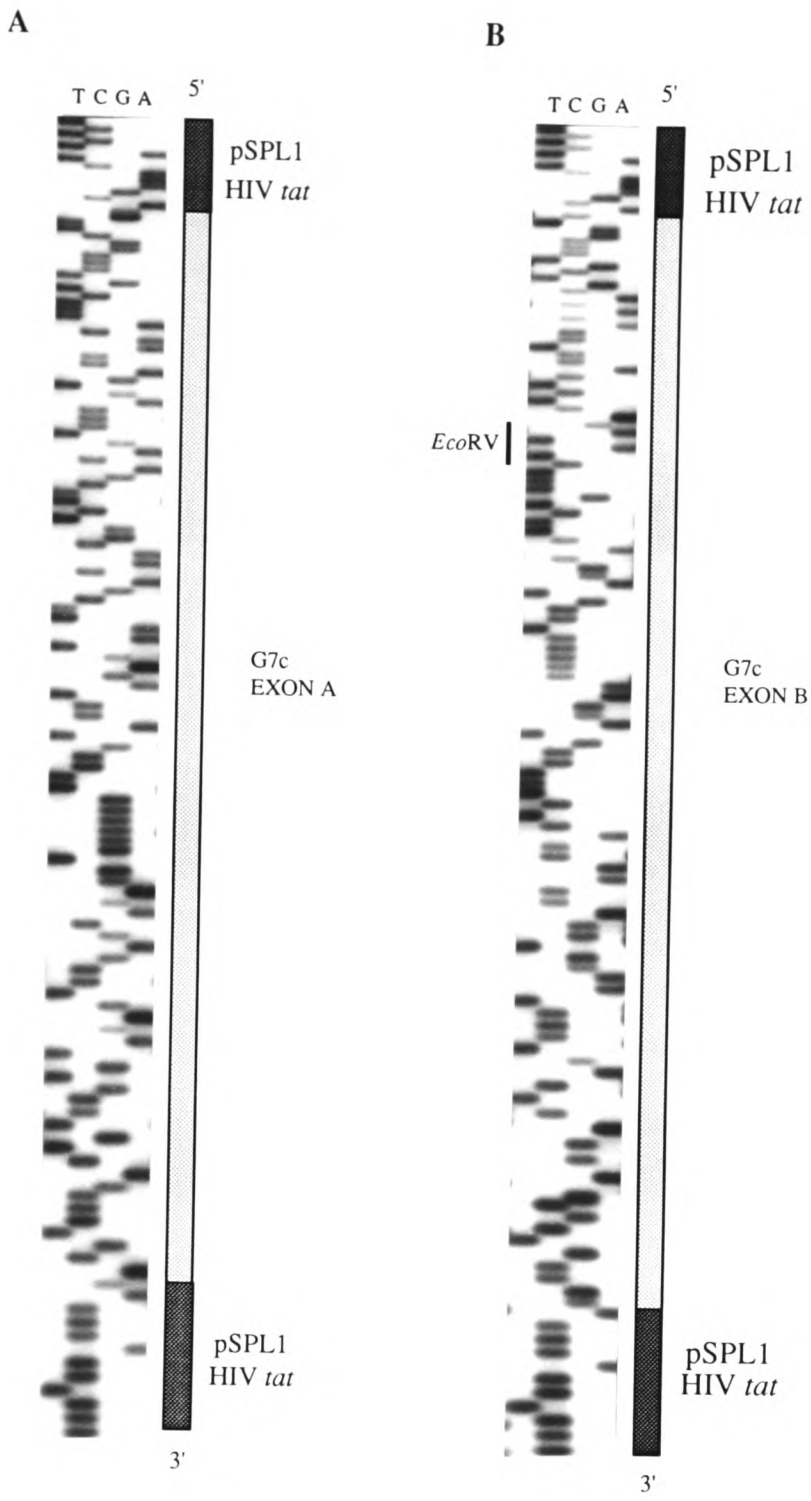


Figure 5.7 Sequence analysis of G7c exon amplification products.
Nucleotide sequence data from the ~210bp G7c exon amplification PCR product shows that there are two different inserts corresponding to the 5' (**A**) and 3' (**B**) sequences of the ~320bp product. The parts of the gels containing HIV *tat* sequences derived from the pSPL1 vector are indicated by dark shading and the parts corresponding to the sequences of the trapped inserts are indicated by light shading.

A

GGATCCGCGACGAAGACCTC CTCAAGGCAGTCAGACTCAT CAAGTTTCTCTATCAAAGCA

GGTTCGGCCCTGTCTTTACA ACCAGTGACCCTGACAGCTT CTGGCAGCAGCTTAATGAGA

Exon A

TCCATGCCTTGGGGGGTGGA GACGAGCCTGAGATGTGCCT GTCAGCCCTGCAG 113

ACCCACCTCCCAATCCGAGG GGACCCGACGTCGAC

B

GGATCCGCGACGAAGACCTC CTCAAGGCAGTCAGACTCAT CAAGTTTCTCTATCAAAGCA

*Eco*RV

CTGGCCCTGCTGCACACACC TCCACTCTCAGATATCTTTG TCTTCACGGATACCTCCCCC

Exon B

AAGGATGCCTTTCTCACCAA CCAACCAGGTGGAATCCCTG ACTCAGGAGGCGCTGCCGG 119

ACCCACCTCCCAATCCGAGG GGACCCGACGTCGAC

C

GGATCCGCGACGAAGACCTC CTCAAGGCAGTCAGACTCAT CAAGTTTCTCTATCAAAGCA

```
  G  S  A  L  S  L  Q  P  V  T  L  T  A  S  G  S  S  L  M  R  S  M  P  W  G  V  E
   V  R  P  C  L  Y  N  Q  *  P  *  Q  L  L  A  A  A  *  *  D  P  C  L  G  G  W
    F  G  P  V  F  T  T  S  D  P  D  S  F  W  Q  Q  L  N  E  I  H  A  L  G  G  G
GGTTCGGCCCTGTCTTTACAACCAGTGACCCTGACAGCTTCTGGCAGCAGCTTAATGAGATCCATGCCTTGGGGGGTGGA

  T  S  L  R  C  A  C  Q  P  C  S  W  P  C  C  T  H  L  H  S  Q  I  S  L  S  S
R  R  A  *  D  V  P  V  S  P  A  A  G  P  A  A  H  T  S  T  L  R  Y  L  C  L  H
 D  E  P  E  M  C  L  S  A  L  Q  L  A  L  L  H  T  P  P  L  S  D  I  F  V  F  T
GACGAGCCTGAGATGTGCCTGTCAGCCCTGCAGCTGGCCCTGCTGCACACACCTCCACTCTCAGATATCTTTGTCTTCAC

R  I  P  P  P  R  M  P  F  S  P  T  N  Q  V  E  S  L  T  Q  E  A  L  P
 G  Y  L  P  Q  G  C  L  S  H  Q  P  T  R  W  N  P  *  L  R  R  R  C  R
  D  T  S  P  K  D  A  F  L  T  N  Q  P  G  G  I  P  D  S  G  G  A  A  G
GGATACCTCCCCCAAGGATGCCTTTCTCACCAACCAACCAGGTGGAATCCCTGACTCAGGAGGCGCTGCCGG 232
```

ACCCACCTCCCAATCCGAGG GGACCCGACGTCGA

Figure 5.8 Nucleotide sequences of G7c exon amplification products.

The nucleotide sequences of two different ~210bp G7c exon amplification products with inserts corresponding to exon A of 113bp (**A**) and exon B of 119bp (**B**) (figure 5.6) are shown. The nucleotide sequence of the ~320bp G7c exon amplification product (**C**) which consists of exon A at the 5' end spliced to exon B at the 3' end to generate a 232bp insert is shown.The insert sequence was translated in three phases and the derived amino acid sequences are shown above the sequence. Phases +1 and +3 on the top and bottom lines both encode continuous open reading frames and phase +2 encodes several stop codons each of which is denoted by an asterisk (*). The HIV *tat* sequences flanking the trapped inserts are underlined. The *Eco*RV restriction site in exon B is indicated and the recognition sequence is underlined in panels **B** and **C**.

sequence contained 13bp of sequence derived from the pBC vector from which the genomic insert was subcloned.

The fragments cloned from the 210bp G7c amplification product contained two different trapped inserts of 113bp, A, and 119bp, B (Figure 5.7), which were spliced together to form the 232bp trapped insert from the ~320bp G7c product (Figure 5.8). Analysis of the nucleotide sequence revealed an *Eco*RV site in the larger, 3', 119bp exon that corresponds to the *Eco*RV site mapped within G7c [section 5.3.2]. The arrangement of the smaller inserts within the ~320bp fragment revealed that the direction of transcription was from G7 towards G7a, in the same orientation as G7.

In order to investigate whether or not G7c corresponded to a known gene, the 232bp sequence from the ~320bp product was used to screen the GENEMBL database using the GCG package FASTA program [section 2.9.7]. The results revealed that there was no sequence similar to G7c in the database.

To determine the potential polypeptide sequence encoded by the 232bp sequence, the nucleotide sequence was translated in three phases (Figure 5.8 C). In phase +2 there are multiple stop codons, but the other two phases, +1 and +3, contain continuous ORFs. In order to try to predict which of these two putative polypeptide sequences may be encoded by G7c, both predicted partial protein sequences were used to screen the SwissProt protein database using the GCG package FASTA program [section 2.9.7]. Neither of the two sequences contained any significant homology with any protein in the database. As a result it was not possible to determine which polypeptide sequence is the partial G7c protein sequence.

5.3.4 Zoo blot analysis of the amplified exons

In order to demonstrate that the exon sequences isolated from G7c are coding sequences, it was necessary to show that these sequences are conserved across species (Monaco *et al.* 1986; Rommens *et al.* 1989). The ~320bp fragment, containing the 232bp sequence encoding the two exons, subcloned into M13mp10 [section 5.3.3], was excised from the double-stranded replicative form of the vector using *Eco*RI and *Hin*dIII and isolated from LGT agarose [section 2.4.3].

This probe was hybridised to a Southern blot of *Bam*HI-digested total genomic DNA from man, mouse, cat, sheep, rabbit, chicken and shark before washing at low stringency [section 5.2.2]. The results showed that the 232bp sequence cross-

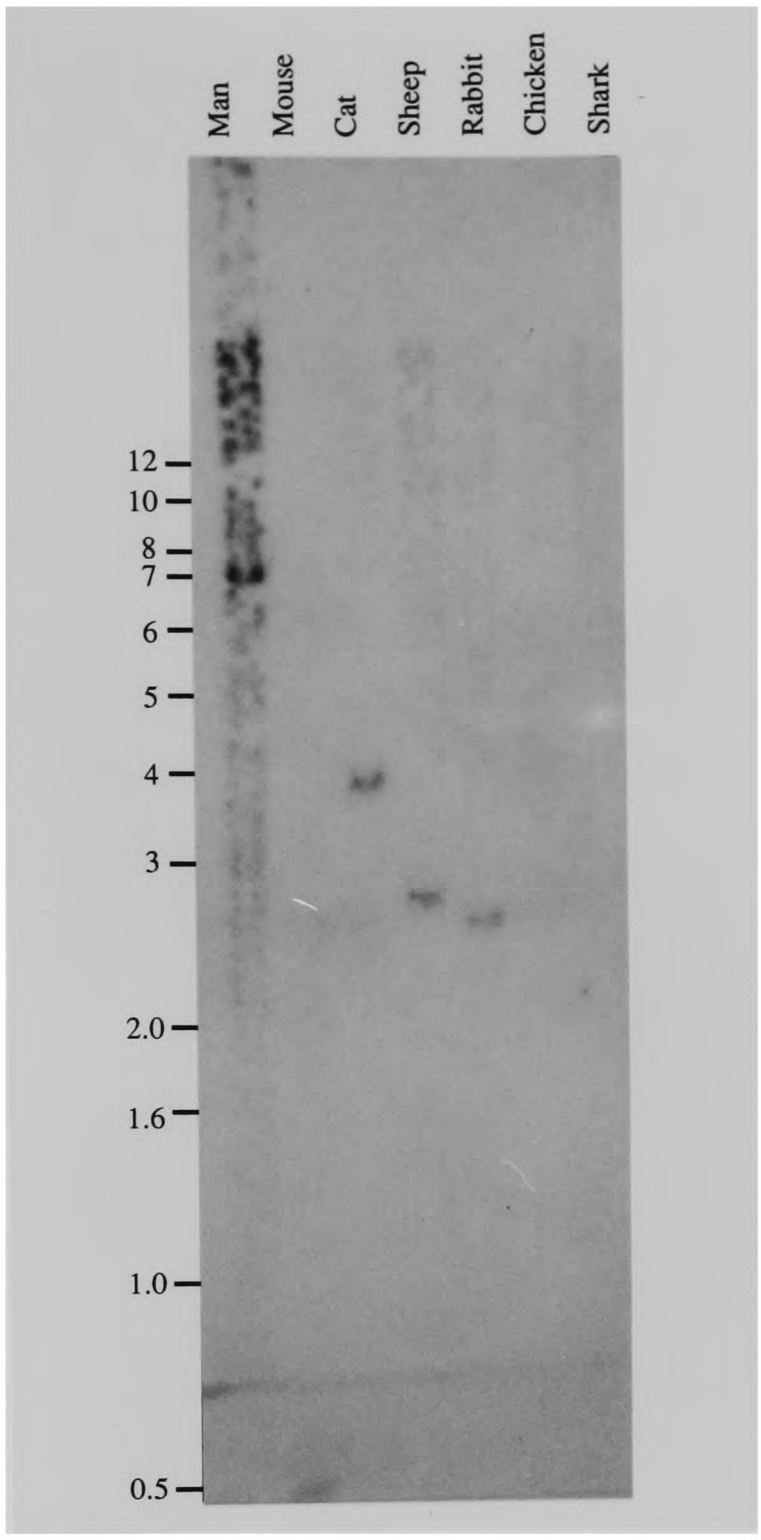


Figure 5.9 Zoo blot analysis of the G7c exon amplification products.
Total genomic DNA from the animal species shown digested with *Bam*HI, was fractionated on a 0.8% (w/v) agarose gel and transferred onto nitrocellulose. The Southern blot was hybridised with the ~320bp G7c exon amplification product insert and washed at low stringency at 65°C with 2xSSC/0.1% (w/v) SDS. The probe cross-hybridised to specific DNA fragments from mouse, cat, sheep and rabbit DNA, but not chicken or shark DNA. Sizes are in kb.

hybridised to DNA fragments in DNA from man, mouse, cat, sheep and rabbit, but not chicken or shark (Figure 5.9). This included an ~7kb DNA fragment from the total human genomic DNA (Figure 5.9) that corresponds to the size of the *Bam*HI fragment within which the sequences were mapped. In addition, the probe detected ~0.5kb, ~4.0kb, ~2.8kb and ~2.7kb fragments from mouse, cat, sheep and rabbit DNA, respectively (Figure 5.9), that correspond to fragments of the same size detected by the 8.8kb *Eag*I/*Sal*I intergenic probe (Figure 5.3). This indicates that the sequence obtained by exon amplification is probably coding sequence, thereby strongly supporting the existence of the G7c gene.

Using the partial coding sequences obtained from the two exons, work is currently being carried out to detect the expression of G7c in order to isolate cDNA clones to obtain the complete coding sequence of G7c.

5.4 DISCUSSION

5.4.1 Strategies for gene detection

The gene density within the human MHC class III region is very high with one gene occurring approximately every 20kb (Milner and Campbell 1992b). These genes have been mapped either in the search for specific genes, *e.g.* C2, C4A, C4B, factor B, TNF–A and TNF-B, or during cosmid walking. This latter approach has utilised the association of genes with HTF-islands as markers for potential genes, or the direct screening of cDNA libraries with cosmid inserts to search for new genes. Although a large number of genes have been detected in this manner their distribution is not even throughout the region. In some places, the intergenic distances are short, 2-5kb, but in other places larger gaps of 10-20kb occur. Is this due to the absence of genes in these regions, or have the genes that lie there not yet been identified?

To map the loci controlling genetic traits at the level of single genes *e.g. Orch-1*, it is important to identify all the possible candidate genes that lie within the region so that they can all be considered, especially if no obvious candidate gene is present. Towards this aim, the correlation of non-repetitive DNA sequences with the locations of genes was examined in order to assess the possibility that more genes lie in the interval between HSP70-Hom and BAT5. Having identified the potential

locations of two genes in this way, one of them was analysed for coding sequences by zoo blot analysis [section 5.2.2] and subsequent exon isolation [section 5.3] to demonstrate the presence of a new gene, G7c.

Although this approach was only applied to a relatively short, ~10kb, region of genomic DNA, it could be applied to larger regions such as whole cosmids, since the mapping of non-repetitive sequences is part of the procedure for isolating probes for cosmid walking. However, a thorough examination of the cosmid, using several restriction enzymes to digest the DNA, may be required in cases where the sites for restriction enzymes lie further apart, *e.g.* in the G7a/G7b region where non-repetitive sequences were not detected.

This approach is more labour intensive than the direct screening of whole cDNA libraries with cosmid DNA inserts, but is not dissimilar to the identification of HTF-island associated genes. However, non-repetitive sequences are associated with all genes, not just those housekeeping, and some tissue-specific, genes that are HTF-island associated. As a result, this approach may be most effective in searching for genes not initially detected by the direct screening methods, such as the hybridisation of cosmid, or even YAC clone inserts onto cDNA libraries (Wallace *et al.* 1990; Elvin *et al.* 1990), or the hybridisation of whole cDNA libraries to immobilised cosmids or YACs (Lovett *et al.* 1991; Parimoo *et al.* 1991), which do not detect all the genes encoded within these clones.

The isolation of more specific, non-repetitive, probes may be more effective in detecting rare mRNA transcripts by Northern blot analysis. Alternatively, 'trapped' exonic sequences can be used to design oligonucleotide primers to screen a wide range of cDNA libraries by PCR, with greater sensitivity, in order to detect gene expression and its tissue distribution

5.4.2 G7c, another novel gene in the human MHC

G7c lies within the ~11kb region between G7 and G7a. The coding sequences obtained by exon amplification and the two derived partial translation products do not share homology with any sequence in the DNA or protein databases, indicating that G7c is a novel, previously unknown, gene. The Northern blot analysis of G7c [section 5.2.2] produced very inconclusive results that do not allow the distribution or size of any G7c transcription product to be determined. However, the partial coding sequences

obtained have allowed the rapid screening of several cDNA libraries by PCR and specific products have been detected in a B lymphocyte derived cDNA library (MR Albertella, personal communication).

CHAPTER 6

POLYMORPHIC ANALYSIS OF *Hsc70t*

6.1 INTRODUCTION

The mapping of the *Orch-1* locus to the *Hsp70–Bat5* interval of the mouse *H-2S* region (Snoek *et al.* 1991) and the subsequent refinement of this position, by recombination mapping, to the ~60kb *Hsp70.1-G7* region (Snoek *et al.* 1993), has reduced the number of candidate genes for the *Orch-1* locus to six. In man this region spans ~70kb centromeric of G7 and encodes G7c, G7a, G7b, HSP70-Hom, HSP70-1 and HSP70-2. In the mouse *H-2* complex, the homologous genes *G7a*, *G7b*, *Hsc70t*, *Hsp70.3* and *Hsp70.1* have been mapped (Snoek *et al.* 1993). Although the human G7c probe has been shown to cross-hybridise with DNA fragments in the mouse genome [section 5.2], the location of the homologous sequence has not yet been mapped. The colinearity of the genes within the human and mouse MHCs has recently been demonstrated (Snoek *et al.* 1993), but the distribution of the genes appears to be different. The *G7b* gene appears to be shorter in the mouse, <4.1kb (Snoek *et al.* 1993) compared with ~10kb in man [section 4.3], although this could be due to regions of sequence sharing low homology and not cross-hybridising between the two species. In addition, the *G7a–G7b* interval in the mouse appears to be significantly larger, ~20kb (Snoek *et al.* 1993), allowing for the possibility that another gene, not present in the human MHC, may be present in the mouse *H-2* complex.

Analysis of the protein products from the human genes in this interval and the pattern of their expression (Milner and Campbell 1990; sections 4.2 & 4.4; Hsieh and Campbell 1991; sections 3.4 & 3.5) does not give any initial indication as to which could potentially be the *Orch-1* locus. However, the mouse HSP70-Hom homologue, *Hsc70t* (Fujimoto *et al.* 1993; Snoek *et al.* 1993), has been shown to be expressed at higher levels in haploid testicular germ cells compared to other cell types (Matsumoto and Fujimoto 1990) and as a result, appears to be the most promising candidate for the *Orch-1* locus.

Hsc70t is a member of the HSP70 protein family, which are highly conserved

proteins that were originally identified as a result of their synthesis in response to heat shock (Welch 1990). However, not all members of the HSP70 protein family are upregulated by stress to the cells and the different proteins can vary in the pattern of their expression and their cellular location. The major HSP70 proteins are expressed constitutively at low levels, but are upregulated when a cell undergoes stress. The human 70kDa heat shock cognate protein (HSC70) is constitutively expressed (Dworniczak and Mirault 1987) and two heat shock proteins have been shown to be specifically expressed at high levels during different stages of germ cell differentiation in mouse testes (Zakeri and Wolgemuth 1987; Zakeri *et al.* 1988). In addition, the heat shock proteins GRP78 (Ting and Lee 1988) and GRP75 (Mizzen *et al.* 1991) are two glucose-regulated human proteins, of which GRP78 is identical to the immunoglobulin heavy chain binding protein BiP (Munro and Pelham 1986).

The heat inducible and constitutively expressed heat shock proteins are normally localised in the cytoplasm and nucleus, whilst GRP78 (Ting and Lee 1988) and GRP75 (Leustek *et al.* 1989; Mizzen *et al.* 1991) are localised in the endoplasmic reticulum and mitochondria, respectively. Under stress, HSP70 proteins move from the cytoplasm to the nucleus and nucleolus (Pelham 1990) where they associate with damaged pre-ribosomes (Welch and Mizzen 1988). HSP70 proteins interact with a diverse range of proteins in the cell and have been shown to bind to synthetic peptides *in vitro* (Flynn *et al.* 1989). It has also been shown that HSC70 can distinguish between folded and unfolded proteins (Palleros *et al.* 1991). In addition, heat shock proteins are believed to play a role as chaperones in the folding, unfolding and translocation of proteins (Ellis and Hemingsen 1989; Langer and Neupert 1991).

Heat shock proteins have ATP-binding and ATPase activities which are associated with a conserved region in the N terminal 382 amino acids (Chappell *et al.* 1987). The ATPase activity is required for the release of bound peptides and proteins and the three dimensional structure of the ATPase domain has been determined (Flaherty *et al.* 1990), which shows structural similarity with the rabbit skeletal muscle actin ATPase domain (Flaherty *et al.* 1991). The role of peptide-binding has been attributed to a 170 amino acid region of HSP70 (Chappell *et al.* 1987; Rippmann *et al.* 1991), which corresponds to residues 383–552. A model structure has been proposed for the HSP70 peptide-binding domain (Rippmann *et al.* 1991) based upon

a consensus secondary structure derived from 30 different HSP70 proteins. The resulting pair of $\beta_4\alpha$ motifs were aligned with the functionally equivalent peptide-binding domain of the class I HLA–A2 molecule, maintaining the positions of residues in the heat shock protein consensus sequence that are structurally important in the HLA molecule. A similar model for the peptide binding domain was proposed by Flajnik *et al.* (1991) following the detection of sequence similarity between the α-1 and α-2 domains from a *Xenopus* MHC class I molecule and the peptide binding domain of human HSC70.

The major HSP70 proteins are involved in immune responses. They are believed to play a role in disease pathogenesis due to molecular mimicry based upon the observation that in bacterial and protozoan infections the major HSP70 proteins are the major antigens that result in a T cell response against heat shock protein epitopes. HSC70 also plays a role in protein translocation across cell membranes and is also believed to be involved in antigen processing and presentation (Pierce *et al.* 1991).

Within the ~70kb region of DNA of the mouse *H2-S* region to which the *Orch-1* locus has been mapped there are three HSP70 genes. Of these three HSP70 genes, one is the *Hsc70t* gene which has been shown to be expressed at high levels in haploid testicular germ cells (Matsumoto and Fujimoto 1990), and therefore appears to be a potential candidate gene for the *Orch-1* locus.

6.2 EXPRESSION OF HSP70-Hom and *Hsc70t*

6.2.1 Northern blot analysis of HSP70-Hom

The expression of HSP70-Hom was investigated by Milner and Campbell (1990) who demonstrated that the gene is expressed constitutively at low levels and is not upregulated following heat shock. However, it has been shown that the mouse homologue, *Hsc70t*, is abundantly expressed in early spermatids (Matsumoto and Fujimoto 1990), although constitutive, low level expression of *Hsc70t* in other cell types was apparently not investigated.

In order to determine whether HSP70-Hom is expressed in the human testis, a

Northern blot of total RNA derived from liver (HepG2), B lymphocyte (Daudi), T lymphocyte (Molt4) and monocyte (HL60 and U937) cultured cell lines and from testis tissue [section 5.2.2], was hybridised with the HSP70-Hom locus specific probe (Milner and Campbell 1990). Following autoradiography for two weeks, very low levels of the ~3kb mRNA transcript were detected in all the tissues examined (result not shown). There was no elevation in the level of HSP70-Hom expression detected in the RNA from testis tissue.

6.2.2 Expression of *Hsc70t* in *H-2* congenic mouse strains

To investigate whether the level of *Hsc70t* expression correlates with susceptibility or resistance to orchitis, the mRNA level in the testes of orchitis susceptible and orchitis resistant mouse strains was studied (Snoek *et al.* 1993). The results showed that the level of *Hsc70t* mRNA in mouse testes is very high, but no difference in expression levels was observed between the mouse strains with different haplotypes (Snoek *et al.* 1993). Therefore, if *Hsc70t* is the *Orch-1* locus, a structural difference in the protein may be responsible for resistance or susceptibility to the disease.

6.3 POLYMORPHIC SEQUENCE ANALYSIS OF *Hsc70t*

6.3.1 Cloning and sequencing of *Hsc70t*

The location of the *Orch-1* locus has been mapped in mouse strains with the $H\text{-}2^d$ and $H\text{-}2^p$ haplotypes, which exhibit susceptibility to autoimmune orchitis accompanied by infertility. This dominant trait has been mapped within the *H-2* region by analysing intra *H-2* recombinant, congenic strains. The B10.A and B10.A(4R) strains represent the susceptible and resistant genotypes, respectively (Teuscher *et al.* 1990). The B10.A(1R), susceptible, and B10.A(2R), resistant, strains have been used to map the *Orch-1* locus, within the DNA segment derived from the $H\text{-}2^d$ haplotype to the *Hsp70.1–G7* interval (Snoek *et al.* 1993). As a result, these mouse strains were selected for polymorphic analysis of the *Hsc70t* gene.

The *Hsc70t* gene was sequenced following isolation of a single, partial cDNA clone and two genomic DNA clones (Matsumoto and Fujimoto 1990). The sequence showed that, like HSP70-Hom (Milner and Campbell 1990), *Hsc70t* is an intronless

```
M A A N K G M A I G I D L G T T Y S C V G V F Q H G K V E I I A D Y Q V N R T T   40
ATGGCTGCTAATAAAGGAATGGCGATCGGCATCGACCTGGGCACCACCTACTCGTGCGTGGGCGTGTTCCAACACGGCAAGGTGGAGATCATCGCCGACTACCAGGTCAACCGCACCACC  120
      HSC70T-5'

P D Y V A F T D T S R L I E R S K N Q V A M N P Q N T V F D A K R L I G R K F N   80
CCTGACTACGTGGCCTTCACCGACACCAGCCGCCTCATCGAGCGCTCCAAGAACCAGGTGGCCATGAATCCCCAGAACACTGTTTTTGATGCCAAACGTCTAATTGGCAGGAAGTTTAAT  240

D P V V Q S N M E A L A I S M I N E A G K P K V M V S Y K G E K K A F Y P E E I  120
GATCCTGTTGTGCAGTCAAATATGGAAGCTTTGGCCATTTCAATGATCAATGAAGCCGGCAAACCCAAGGTGATGGTGTCCTATAAAGGAGAGAAGAAAGCCTTCTACCCAGAGGAGATC  360
                    MHSC1

S S M V L T K M K E T A E V F W A P M S Q R V I T V P A Y F N D S Q R Q A T K D  160
TCATCCATGGTACTGACGAAGATGAAGGAGACTGCAGAGGTTTTTTGGGCACCAATGTCACAACGTGTGATCACGGTGCCAGCCTATTTCAATGACTCTCAGCGGCAAGCCACCAAGGAT  480

A G V I A G L N V L R I I N E P T A V A W A R L D K G S H A E R H V L I F D L G  200
GCAGGTGTCATCGCAGGACTCAATGTGCTGAGAATAATCAATGAGCCCACGGCGGTCGCATGGGCTAGGTTGGATAAAGGAAGTCACGCAGAGCGGCACGTGCTCATCTTCGACCTGGGG  600
                                                    MHSC2

G G T F D L S I L T I D D A S S V K P R G D T H L G G E D F D N R L V S H F V R  240
GGTGGCACGTTCGACTTGTCCATCCTGACGATCGACGACGCATCTTCAGTGAAGCCACGCGGCGACACGCACCTGGGAGGGGAGGACTTCGACAACCGGCTGGTGAGCCACTTCGTGAGA  720

V K R K H K K D N Q A K N K R A V R R L R T A C E R A K R T L S S S T Q A N L E  280
GTCAAGAGGAAGCACAAGAAGGACAATCAAGCCAAGAACAAGCGCGCGGTGCGGCGGCTGCGCACGGCGTGTGAGAGGGCCAAGAGGACGCTGTCGTCCAGCACCCAGGCAAACCTGGAG  840

I D S L Y E G I D F Y T S I T R A R F E E L C A D L F R G T L E P V E K S L R D  320
ATCGACTCTTTATATGAGGGCATCGACTTCTACACGTCCATCACTAGAGCACGGTTTGAAGAGCTGTGTGCAGACCTATTTAGAGGCACACTTGAGCCCGTGGAAAAGTCTCTTCGGGAT  960
                  MHSC3

A K M D K A K I H D I V L V G G S T R I P K V Q K M L Q D Y F N G R D L N K S I  360
GCCAAGATGGATAAGGCTAAAATCCATGACATTGTTCTAGTAGGGGGCTCCACCCGCATCCCAAAGGTGCAAAAAATGCTTCAGGACTACTTTAATGGACGGGATCTCAACAAGAGTATC 1080

N P D E A V A Y G A A V Q G A I L M G D K S E K V Q D L L L L D V A P L S L G L  400
AATCCCGATGAGGCAGTCGCCTACGGAGCTGCAGTCCAGGGAGCTATTTTAATGGGCGACAAATCTGAAAAAGTACAGGATTTGCTTTTGTTGGACGTAGCTCCCCTGTCTCTAGGATTG 1200
                                                    MHSC4

E T A G G V M T V L I K R N S T I P T K Q T K I F T T Y S D N Q P G V L I Q C T  440
GAGACAGCTGGGGGTGTGATGACTGTATTGATCAAGCGGAACTCCACCATCCCCACCAAGCAGACGAAGATCTTCACCACCTACTCGGACAACCAGCCCGGGGTATTGATCCAGTGTACG 1320

R Q T I P D N N L V G P F D L T G I P P A P S V P Q I E V T F D I A R N G I L N  480
AGGCAAACCATACCCGACAACAACCTGGTGGGGCCCTTTGACTTGACTGGAATACCTCCTGCACCTAGCGTGCCACAGATCGAGGTGACCTTCGACATCGCGCGCAACGGTATCCTGAAC 1440

V T A M D K S T G K A N K I T I T N D K G R L S K E E I D S I E S R T A Y K R E  520
GTTACGGCCATGGACAAGAGCACCGGCAAGGCCAACAAGATCACCATCACCAACGACAAGGGTCGCCTGAGCAAGGAGGAGATTGACTCGATCGAGAGCCGGACTGCTTACAAACGTGAG 1560
                  MHSC5

D E G Q R E K I A A K N A L E S Y A F N M K S A C G D E G L K D K I T E S D K K  560
GATGAGGGTCAGAGGGAGAAAATCGCTGCCAAAAATGCCTTAGAATCGTACGCCTTTAATATGAAGAGCGCTTGTGGTGATGAGGGTCTGAAGGACAAGATCACGGAGTCCGATAAAAAG 1680

K I L D K C N E V P F L L S Q T Q L A E K V E F D H K R K E L E N M C N P I I T  600
AAAATACTGGATAAATGCAATGAGGTTCCTTTCCTGCTGAGTCAAACCCAGCTGGCTGAGAAAGTAGAGTTTGATCATAAAAGAAAAGAACTGGAAAATATGTGTAATCCGATCATCACA 1800
                  MHSC6

K L Y Q S G C T G P T V R Q G I L R Q G S T G P T I E E V D *   630
AAACTGTACCAGAGCGGATGCACCGGGCCCACTGTACGCCAGGGTATACTCCGGCAGGGCTCCACAGGCCCTACCATCGAGGAAGTAGATTAGCCTTTTCCAGAATTGCAGGGTGCTAGG 1920
                                                                    BamHI
GTGCCTCTAGGCGAATTTTATTCATCTTCAAACATCAATATGATTCTTGAACTGACTGGACTCAGCCTACGTCACCAGTCCTTTGGATCCCGATGGAGAAGCTCGAACGCACCTTTTCAC 2040
                                                                    HSC70T-3'

ACCCCCACCCGTCGTCAGTCTATGATCCTGAACTGGACCTTTAGGAAAACCAGGCCCCTCTTTAGAGCCATTTGAGAATTTGATGTCTGTATTTATTATCCACACCCCACCTTTCTCCTT 2160

CCTGTGTGTTGGTTATTTGTCTCTCAATAAATTTGTTCCCAAAGGAAACAACTCTGCACTTTTATTTTAGCAGGCCTTACTGATATGAG 2240
```

Figure 6.1 Nucleotide sequence of the mouse *Hsc70t* gene and the derived amino acid sequence.
This shows the nucleotide sequence of *Hsc70t* obtained by Matsumoto and Fujimoto (1990). The derived amino acid sequence is shown above the nucleotide sequence. The sequences from which the oligonucleotide primers HSC70t-5', HSC70T-3' and MHSC1-6 were determined are underlined. The location of the *Bam*HI site in the 3' untranslated region is indicated.

OLIGONUCLEOTIDE	SEQUENCE	POSITION‡	LENGTH	Td	USE
HSC70T-5'	*Nde*I CAGGCATATGGCTGCTAATAAAGGAATGGC	1-23	30 bases*	66°C	PCR (Sense)
HSC70T-3'	*Bam*HI GAGCTTCTCCATCGGGATCC	2005-2024	20 bases	64°C	PCR (Antisense)
MHSC1	GTCAAATATGGAAGCTTTGGCC	255-276	22 bases	64°C	Sequencing
MHSC2	GTCACGCAGAGCGGCACG	563-580	18 bases	62°C	Sequencing
MHSC3	GGCATCGACTTCTACACGTC	859-878	20 bases	62°C	Sequencing
MHSC4	CAGGATTTGCTTTTGTTGGACG	1156-1177	22 bases	64°C	Sequencing
MHSC5	GAGCACCGGCAAGGCCAAC	1458-1476	19 bases	64°C	Sequencing
MHSC6	GCAATGAGGTTCCTTTCCTGC	1697-1717	21 bases	64°C	Sequencing

‡ The positions of the nucleotides are numbered according to the sequence of Matsumoto and Fujimoto (1990) shown in figure 6.1.

* Oligonucleotide HSC70T-5' contains 7 bases of adaptor sequence at the 5' end.

Table 6.1 Oligonucleotides used for PCR amplification and sequence analysis of the *Hsc70t* gene.

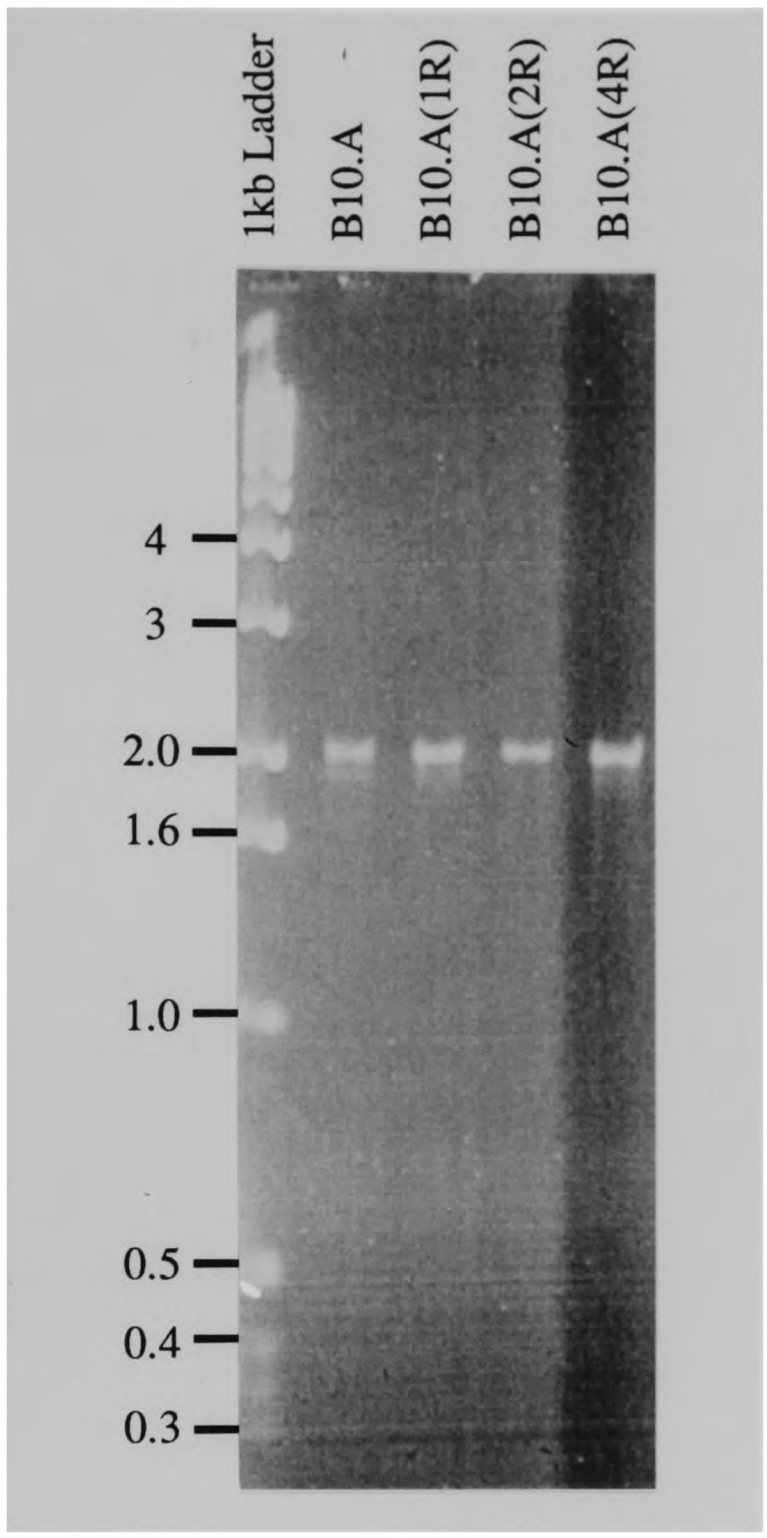


Figure 6.2 Amplification of *Hsc70t* from *H-2* congenic mouse strains. The coding sequence from the *Hsc70t* gene was amplified by PCR from the congenic mouse strains shown using the oligonucleotide primers HSC70T-5' and HSC70T-3'. Thirty five cycles of 1 min at 94°C, 1 min at 62°C and 3 min at 72°C were used to amplify the ~2.1kb product which was fractionated on a 0.8% (w/v) agarose gel.

gene. As a result, *Hsc70t* could be directly amplified by PCR using genomic DNA templates, thereby removing the need to isolate mRNA from testes for reverse transcription before amplification.

Based on the *Hsc70t* nucleotide sequence (Figure 6.1) of Matsumoto and Fujimoto (1990), two specific oligonucleotide primers, unique to the *Hsc70t* gene sequence, HSC70T-5′ and HSC70T-3′ (Table 6.1), were designed for amplification of the gene and six oligonucleotide primers, MHSC1–6 (Table 6.1), were designed as sequencing primers. The 5′, sense, oligonucleotide PCR primer contains the first 23 nucleotides of the published *Hsc70t* sequence (Figure 6.1), preceded by seven bases of an adaptor sequence that generates a unique *Nde*I restriction site at the 5′ end of the coding sequence. The 3′, antisense, oligonucleotide primer sequence is derived from the 3′ untranslated region of *Hsc70t*, including a unique *Bam*HI restriction site (Figure 6.1).

The coding sequence of *Hsc70t* was amplified by PCR from total genomic DNA prepared from the livers of mice from the strains B10.A, B10.A(1R). B10.A(2R) and B10.A(4R) (a gift from M Snoek, Division of Molecular Genetics, Netherlands Cancer Institute, Amsterdam). Using ~200ng of total mouse genomic DNA as the template in 100μl reactions, 35 cycles were carried out using the following conditions: 1 min at 94°C, 1 min at 62°C and 3 min at 72°C. All the samples contained an amplified product of ~2.1kb (Figure 6.2), as predicted from the published nucleotide sequence.

The amplified products were digested with *Nde*I and *Bam*HI, end-filled [section 2.4.4], fractionated on a 3.5% (w/v) polyacrylamide gel [section 2.4.5] and isolated as described in section 2.4.6. The inserts were cloned into *Eco*RV cut pBluescript and transformed into TG1 competent cells [section 2.7.5]. Resulting white colonies were picked, cultured overnight [section 2.2.2] and double-stranded DNA was isolated [section 2.3.2.2]. The cloned inserts were sequenced using the M13 universal primer, the T3 pBluescript primer and the specific primers MHSC1–6, using double-stranded DNA templates [section 2.9.4].

6.3.2 *Hsc70t* nucleotide sequence analysis

The nucleotide sequences of *Hsc70t* from the mouse strains B10.A, B10.A(1R). B10.A(2R) and B10.A(4R) were aligned with the published sequence of Matsumoto and Fujimoto (1990) using the CLUSTAL V program (Higgins and Sharp 1988) to compare the sequences (Figure 6.3). A striking feature of this alignment is the large

```
B10.A       ATGGCTGCTAATAAAGGAAT GGCGATCGGCATCGACTTGG GCACCACCTACTCGTGCGTG GGCGTGTTCCAGCACGGCAA GGTGGAGATCATCGCCAACG ACCAGGGCAACCGCACGACC  120
B10.A(1R)   ATGGCTGCTAATAAAGGAAT GGCGATCGGCATCGACTTGG GCACCACCTACTCGTGCGTG GGCGTGTTCCAGCACGGCAA GGTGGAGATCATCGCCAACG ACCAGGGCAACCGCACGACC  120
B10.A(2R)   ATGGCTGCTAATAAAGGAAT GGCGATCGGCATCGACTTGG GCACCACCTACTCGTGCGTG GGCGTGTTCCAGCACGGCAA GGTGGAGATCATCGCCAACG ACCAGGGCAACCGCACGACC  120
B10.A(4R)   ATGGCTGCTAATAAAGGAAT GGCGATCGGCATCGACTTGG GCACCACCTACTCGTGCGTG GGCGTGTTCCAGCACGGCAA GGTGGAGATCATCGCCAACG ACCAGGGCAACCGCACGACC  120
Hsc70t      ATGGCTGCTAATAAAGGAAT GGCGATCGGCATCGACCTGG GCACCACCTACTCGTGCGTG GGCGTGTTCCAACACGGCAA GGTGGAGATCATCGCCGACT ACCAGGTCAACCGCACCACC  120

B10.A       CCTAGCTACGTGGCCTTCAC CGACACCGAGC GCCTCATCGGAGACGCTGCCAAGAACCAG GTGGCCATGAATCCCCAGAA CACTGTTTTTGATGCCAAAC GTCTAATTGGCAGGAAGTTT  240
B10.A(1R)   CCTAGCTACGTGGCCTTCAC CGACACCGAGC GCCTCATCGGAGACGCTGCCAAGAACCAG GTGGCCATGAATCCCCAGAA CACTGTTTTTGATGCCAAAC GTCTAATTGGCAGGAAGTTT  240
B10.A(2R)   CCTAGCTACGTGGCCTTCAC CGACACCGAGC GCCTCATCGGAGACGCTGCCAAGAACCAG GTGGCCATGAATCCCCAGAA CACTGTTTTTGATGCCAAAC GTCTAATTGGCAGGAAGTTT  240
B10.A(4R)   CCTAGCTACGTGGCCTTCAC CGACACCGAGC GCCTCATCGGAGACGCTGCCAAGAACCAG GTGGCCATGAATCCCCAGAA CACTGTTTTTGATGCCAAAC GTCTAATTGGCAGGAAGTTT  240
Hsc70t      CCTGACTACGTGGCCTTCAC CGACACC-AGCCGCCTCATCG-AG-CGCT-CCAAGAACCAG GTGGCCATGAATCCCCAGAA CACTGTTTTTGATGCCAAAC GTCTAATTGGCAGGAAGTTT  237

B10.A       AATGATCCTGTTGTGCAGTC AGATATG AAGCTTTGGCCATTTCAAGTGATCAATGAAGCC GGCAAACCCAAGGTGATGGT GTCCTATAAAGGAGAGAAGA AAGCCTTCTACCCAGAGGAG  360
B10.A(1R)   AATGATCCTGTTGTGCAGTC AGATATG AAGCTTTGGCCATTTCAAGTGATCAATGAAGCC GGCAAACCCAAGGTGATGGT GTCCTATAAAGGAGAGAAGA AAGCCTTCTACCCAGAGGAG  360
B10.A(2R)   AATGATCCTGTTGTGCAGTC AGATATG AAGCTTTGGCCATTTCAAGTGATCAATGAAGCC GGCAAACCCAAGGTGATGGT GTCCTATAAAGGAGAGAAGA AAGCCTTCTACCCAGAGGAG  360
B10.A(4R)   AATGATCCTGTTGTGCAGTC AGATATG AAGCTTTGGCCATTTCAAGTGATCAATGAAGCC GGCAAACCCAAGGTGATGGT GTCCTATAAAGGAGAGAAGA AAGCCTTCTACCCAGAGGAG  360
Hsc70t      AATGATCCTGTTGTGCAGTC AAATATGGAAGCTTTGGCCATTTCAA-TGATCAATGAAGCC GGCAAACCCAAGGTGATGGT GTCCTATAAAGGAGAGAAGA AAGCCTTCTACCCAGAGGAG  357

B10.A       ATCTCATCCATGGTACTGAC GAAGATGAAGGAGACTGCAG AGGCTTTTTTGGGCCACAAT GTCACCGACGCTGTGATCAC GGTGCCAGCCTATTTCAATG ACTCTCAGCGGCAAGCCACC  480
B10.A(1R)   ATCTCATCCATGGTACTGAC GAAGATGAAGGAGACTGCAG AGGCTTTTTTGGGCCACAAT GTCACCAACGCTGTGATCAC GGTGCCAGCCTATTTCAATG ACTCTCAGCGGCAAGCCACC  480
B10.A(2R)   ATCTCATCCATGGTACTGAC GAAGATGAAGGAGACTGCAG AGGCTTTTTTGGGCCACAAT GTCACCAACGCTGTGATCAC GGTGCCAGCCTATTTCAATG ACTCTCAGCGGCAAGCCACC  480
B10.A(4R)   ATCTCATCCATGGTACTGAC GAAGATGAAGGAGACTGCAG AGGCTTTTTTGGGCCACAAT GTCACCAACGCTGTGATCAC GGTGCCAGCCTATTTCAATG ACTCTCAGCGGCAAGCCACC  480
Hsc70t      ATCTCATCCATGGTACTGAC GAAGATGAAGGAGACTGCAG AGG-TTTTTTGGGCACCAAT GTCAC-AACG-TGTGATCAC GGTGCCAGCCTATTTCAATG ACTCTCAGCGGCAAGCCACC  474

B10.A       AAGGATGCAGGTGTCATCGC AGGACTCAATGTGCTGAGAA TAATCAATGAGCCCACGGCG GCCGCCATCGCCTACGGCTT GGATAAAGGAAGTCACGGAG AGCGGCACGTGCTCATCTTC  600
B10.A(1R)   AAGGATGCAGGTGTCATCGC AGGACTCAATGTGCTGAGAA TAATCAATGAGCCCACGGCG GCCGCCATCGCCTACGGCTT GGATAAAGGAAGTCACGGAG AGCGGCACGTGCTCATCTTC  600
B10.A(2R)   AAGGATGCAGGTGTCATCGC AGGACTCAATGTGCTGAGAA TAATCAATGAGCCCACGGCA GCCGCCATCGCCTACGGCTT GGATAAAGGAAGTCACGGAG AGCGGCACGTGCTCATCTTC  600
B10.A(4R)   AAGGATGCAGGTGTCATCGC AGGACTCAATGTGCTGAGAA TAATCAATGAGCCCACGGCG GCCGCCATCGCCTACGGCTT GGATAAAGGAAGTCACGGAG AGCAGCACGTGCTCATCTTC  600
Hsc70t      AAGGATGCAGGTGTCATCGC AGGACTCAATGTGCTGAGAA TAATCAATGAGCCCACGGCG GTCGC-ATGGGCTA-GG-TT GGATAAAGGAAGTCACGCAG AGCGGCACGTGCTCATCTTC  591

B10.A       GACCTGGGGGGTGGCACGTT CGACGTGTCCATCCTGACGA TCGACGACGGCATCTTCGAG GTGAAGGCCACGGCGGGCGA CACGCACCTGGGAGGGGAGG ACTTCGACAACCGGCTGGTG  720
B10.A(1R)   GACCTGGGGGGTGGCACGTT CGACGTGTCCATCCTGACGA TCGACGACGGCATCTTCGAG GTGAAGGCCACGGCGGGCGA CACGCACCTGGGAGGGGAGG ACTTCGACAACCGGCTGGTG  720
B10.A(2R)   GACCTGGGGGGTGGCACGTT CGACGTGTCCATCCTGACGA TCGACGACGGCATCTTCGAG GTGAAGGCCACGGCGGGCGA CACGCACCTGGGAGGGGAGG ACTTCGACAACCGGCTGGTG  720
B10.A(4R)   GACCTGGGGGGTGGCACGTT CGACGTGTCCATCCTGACGA TCGACGACGGCATCTTCGAG GTGAAGGCCACGGCGGGCGA CACGCACCTGGGAGGGGAGG ACTTCGACAACCGGCTGGTG  720
Hsc70t      GACCTGGGGGGTGGCACGTT CGACTTGTCCATCCTGACGA TCGACGACG-CATCTTC-A- GTGAAG-CCACG-C-GGCGA CACGCACCTGGGAGGGGAGG ACTTCGACAACCGGCTGGTG  716

B10.A       AGCCACTTCGTGGAGGAGTTCAAGAGGAAGCACAAGAAGGACA TCA GCCA GAACAAGCGCGCGGTGCGGCGGCTGCGCAC TGCGTGTGAGAGGGCCAAGA GGACGCTGTCGTCCAGCACC  840
B10.A(1R)   AGCCACTTCGTGGAGGAGTTCAAGAGGAAGCACAAGAAGGACA TCA GCCA GAACAAGCGCGCGGTGCGGCGGCTGCGCAC TGCGTGTGAGAGGGCCAAGA GGACGCTGTCGTCCAGCACC  840
B10.A(2R)   AGCCACTTCGTGGAGGAGTTCAAGAGGAAGCACAAGAAGGACA TCA GCCA GAACAAGCGCGCGGTGCGGCGGCTGCGCAC TGCGTGTGAGAGGGCCAGGA GGACGCTGTCGTCCAGCACC  840
B10.A(4R)   AGCCACTTCGTGGAGGAGTTCAAGAGGAAGCACAAGAAGGACA TCA GCCA GAACAAGCGCGCGGTGCGGCGGCTGCGCAC TGCGTGTGAGAGGGCCAAGA GGACGCTGTCGTCCAGCACC  840
Hsc70t      AGCCACTTCGTG-AG-AGT-CAAGAGGAAGCACAAGAAGGACAATCAAGCCAAGAACAAGCGCGCGGTGCGGCGGCTGCGCAC GGCGTGTGAGAGGGCCAAGA GGACGCTGTCGTCCAGCACC  825

B10.A       CAGGCAAACCTGGAGATCGA CTCTTTATATGAGGGCATCG ACTTCTACACGTCCATCACT AGAGCACGGTTTGAAGAGCT GTGTGCAGACCTATTTAGAG GCACACTTGAGCCCGTGGAA  960
B10.A(1R)   CAGGCAAACCTGGAGATCGA CTCTTTATATGAGGGCATCG ACTTCTACACGTCCATCACT AGAGCACGGTTTGAAGAGCT GTGTGCAGACCTATTTAGAG GCACACTTGAGCCCGTGGAA  960
B10.A(2R)   CAGGCAAACCTGGAGATCGA CTCTTTATATGAGGGCATCG ACTTCTACACGTCCATCACT AGAGCACGGTTTGAAGAGCT GTGTGCAGACCTATTTAGAG GCACACTTGAGCCCGTGGAA  960
B10.A(4R)   CAGGCAAACCTGGAGATCGA CTCTTTATATGAGGGTATCG ACTTCTACACGTCCATCACT AGAGCACGGTTTGAAGAGCT GTGTGCAGACCTATTTAGAG GCACACTTGAGCCCGTGGAA  960
Hsc70t      CAGGCAAACCTGGAGATCGA CTCTTTATATGAGGGCATCG ACTTCTACACGTCCATCACT AGAGCACGGTTTGAAGAGCT GTGTGCAGACCTATTTAGAG GCACACTTGAGCCCGTGGAA  945

B10.A       AAGTCTCTTCGGGATGCCAA GATGGATAAGGCTAAAATCC ATGACATTGTTCTAGTAGGG GGCTCCACCCGCATCCCAAA GGTGCAAAAACTGCTTCAGG ACTACTTTAATGGACGGGAT 1080
B10.A(1R)   AAGTCTCTTCGGGATGCCAA GATGGATAAGGCTAAAATCC ATGACATTGTTCTAGTAGGG GGCTCCACCCGCATCCCAAA GGTGCAAAAACTGCTTCAGG ACTACTTTAATGGACGGGAT 1080
B10.A(2R)   AAGTCTCTTCGGGATGCCAA GATGGATAAGGCTAAAATCC ATGACATTGTTCTAGTAGGG GGCTCCACCCGCATCCCAAA GGTGCAAAAACTGCTTCAGG ACTACTTTAATGGACGGGAT 1080
B10.A(4R)   AAGTCTCTTCGGGATGCCAA GATGGATAAGGCTAAAATCC ATGACATTGTTCTAGTAGGG GGCTCCACCCGCATCCCAAA GGTGCAAAAACTGCTTCAGG ACTACTTTAATGGACGGGAT 1080
Hsc70t      AAGTCTCTTCGGGATGCCAA GATGGATAAGGCTAAAATCC ATGACATTGTTCTAGTAGGG GGCTCCACCCGCATCCCAAA GGTGCAAAAAAATGCTTCAGG ACTACTTTAATGGACGGGAT 1065

B10.A       CTCAACAAGAGTATCAATCC CGATGAGGCAGTCGCCTACG GAGCTGCAGTCCAGGCAGCT ATTTTAATGGGCGACAAATC TGAAAAAGTACAGGATTTGC TTTTGTTGGACGTAGCTCCC 1200
B10.A(1R)   CTCAACAAGAGTATCAATCC CGATGAGGCAGTCGCCTACG GAGCTGCAGTCCAGGCAGCT ATTTTAATGGGCGACAAATC TGAAAAAGTACAGGATTTGC TTTTGTTGGACGTAGCTCCC 1200
B10.A(2R)   CTCAACAAGAGTATCAATCC CGATGAGGCAGTCGCCTACG GAGCTGCAGTCCAGGCAGCT ATTTTAATGGGCGACAAATC TGAAAAAGTACAGGATTTGC TTTTGTTGGACGTAGCTCCC 1200
B10.A(4R)   CTCAACAAGAGTATCAATCC CGATGAGGCAGTCGCCTACG GAGCTGCAGTCCAGGCAGCT ATTTTAATGGGCGACAAATC TGAAAAAGTACAGGATTTGC TTTTGTTGGACGTAGCTCCC 1200
Hsc70t      CTCAACAAGAGTATCAATCC CGATGAGGCAGTCGCCTACG GAGCTGCAGTCCAGGGAGCT ATTTTAATGGGCGACAAATC TGAAAAAGTACAGGATTTGC TTTTGTTGGACGTAGCTCCC 1185
```

```
B10.A       CTGTCTCTAGGATTGGAGAC AGCTGGGGGTGTGATGACTG TACTGATCAAGCGCAACTCC ACCATCCCCACCAAGCAGAC GCAGATCTTCACCACCTACT CGGACAACCAGCCCGGGGTG 1320
B10.A(1R)   CTGTCTCTAGGATTGGAGAC AGCTGGGGGTGTGATGACTG TACTGATCAAGCGCAACTCC ACCATCCCCACCAAGCAGAC GCAGATCTTCACCACCTACT CGGACAACCAGCCCGGGGTG 1320
B10.A(2R)   CTGTCTCTAGGATTGGAGAC AGCTGGGGGTGTGATGACTG TACTGATCAAGCGCAACTCC ACCATCCCCACCAAGCAGAC GCAGATCCTCACCACCTACT CGGACAACCAGCCCGGGGTG 1320
B10.A(4R)   CTGTCTCTAGGATTGGAGAC AGCTGGGGGTGTGATGACTG TACTGATCAAGCGCAACTCC ACCATCCCCACCAAGCAGAC GCAGATCTTCACCACCTACT CGGGCGACCAGCCCGGGGTG 1320
Hsc70t      CTGTCTCTAGGATTGGAGAC AGCTGGGGGTGTGATGACTG TATTGATCAAGCGGAACTCC ACCATCCCCACCAAGCAGAC GAAGATCTTCACCACCTACT CGGACAACCAGCCCGGGGTG 1320

B10.A       CTGATCCAGGTGTGCGAGGG CGAGAGAGCCATGACGCGCG ACAACAACCTGCTGGGGCGC TTTGACTTGACTGGAATACC TCCTGCACCTAGGGGCGTGC CACAGATCGAGGTGACCTTC 1440
B10.A(1R)   CTGATCCAGGTGTACGAGGG CGAGAGAGCCATGACGCGCG ACAACAACCTGCTGGGGCGC TTTGACTTGACTGGAATACC TCCTGCACCTAGGGGCGTGC CACAGATCGAGGTGACCTTC 1440
B10.A(2R)   CTGATCCAGGTGTACGAGGG CGAGAGAGCCATGACGCGCG ACAACAACCTGCTGGGGCGC TTTGACTTGACTGGAATACC TCCTGCACCTAGGGGCGTGC CACAGATCGAGGTGACCTTC 1440
B10.A(4R)   CTGATCCAGGTGTACGAGGG CGAGAGAGCCATGACGCGCG ACAACAACCTGCTGGGGCGC TTTGACTTGACTGGAATACC TCCTGCACCTAGGGGCGTGC CACAGATCGAGGTGACCTTC 1440
Hsc70t      TTGATCCAG-TGTACGAGG- C-A-A-A-CCAT-AC-C-CG ACAACAACCTGGTGGGGCCC TTTGACTTGACTGGAATACC TCCTGCACCTAG---CGTGC CACAGATCGAGGTGACCTTC 1440

B10.A       GACATCGACGC CAACGGTATCCTGAACGTCACGGCCATGG ACAAGAGCACCGGCAAGGCC AACAAGATCACCATCACCAA CGACAAGGGTCGCCTGAGCA AGGAGGAGATTGAGCGCATG 1560
B10.A(1R)   GACATCGACGC CAACGGTATCCTGAACGTCACGGCCATGG ACAAGAGCACCGGCAAGGCC AACAAGATCACCATCACCAA CGACAAGGGTCGCCTGAGCA AGGAGGAGATTGAGCGCATG 1560
B10.A(2R)   GACATCGACGC CAACGGTGTCCTGAACGTCACGGCCATGG ACAAGAGCACCGGCAAGGCC AACAAGATCACCATCACCAA CGACAAGGGTCGCCTGAGCA AGGAGGAGATTGAGCGCATG 1560
B10.A(4R)   GACATCGACGC CAACGGTATCCTGAACGTCACGGCCATGG ACAAGAGCACCGGCAAGGCC AACAAGATCACCATCACCAA CGACAAGGGTCGCCTGAGCA AGGAGGAGATTGAGCGCATG 1560
Hsc70t      GACATCG-CGCGCAACGGTATCCTGAACGTTACGGCCATGG ACAAGAGCACCGGCAAGGCC AACAAGATCACCATCACCAA CGACAAGGGTCGCCTGAGCA AGGAGGAGATTGAGCGCATG 1560

B10.A       GTGCAGGAGGCCG AGCGCT ACAAAGCG GAGGATGAGGGCCAGAGGGAGAAAATCGCTGCCAAAAATGCCTTAGAATCGTA CGCCTTTAATATGAAGAGCG CTGTGGGTGATGAGGGTCTG 1680
B10.A(1R)   GTGCAGGAGGCCG AGCGCT ACAAAGCG GAGGATGAGGGCTAGAGGGAGAAAATCGCTGCCAAAAATGCCTTAGAATCGTA CGCCTTTAATATGAAGAGCG CTGTGGGTGATGAGGGTCTG 1680
B10.A(2R)   GTGCAGGAGGCCG AGCGCT ACAAAGCG GAGGATGAGGGCCGGGGGGGAGAAAATCGCTGCCAAAAATGCCTTAGAATCGTA CGCCTTTAATATGAAGAGCG CTGTGGGTGATGAGGGTCTG 1680
B10.A(4R)   GTGCAGGAGGCCG AGCGCT ACAAAGCG GAGGATGAGGGCCAGAGGGAGAAAATCGCTGCCAAAAATGCCTTAGAATCGTA CGCCTTTAATATGAAGAGCG CTGTGGGTGATGAGGGTCTG 1680
Hsc70t      ATCGAG-AG-CCGGACTGCTTACAAA-CGTGAGGATGAGGGTCAGAGGGAGAAAATCGCTGCCAAAAATGCCTTAGAATCGTA CGCCTTTAATATGAAGAGCG CTGTGGGTGATGAGGGTCTG 1680

B10.A       AAGGACAAGATCAGCGAGTC CGATAAAAAGAAAATACTGGATAAATGCAATGAGGT CCTTTCCTGGCTGGAGGCCAA CC AGCTGGCTGAGAAAGATGAGTTTGATCATAAAAGAAAAGAAC 1800
B10.A(1R)   AAGGACAAGATCAGCGAGTC CGATAAAAAGAAAATACTGGATAAATGCAATGAGGT CCTTTCCTGGCTGGAGGCCAA CC AGCTGGCTGAGAAAGATGAGTTTGATCATAAAAGAAAAGAAC 1800
B10.A(2R)   AAGGACAAGATCAGCGAGTC CGATAAAAAGAAAATACTGGATAAATGCAATGAGGT CCTTTCCTGGCTGGAGGCCAA CC AGCTGGCTGAGAAAGATGAGTTTGATCATAAAAGAAAAGAAC 1800
B10.A(4R)   AAGGACAAGATCAGCGAGTC CGATAAAAAGAAAATACTGGATAAATGCAATGAGGT CCTTTCCTGGCTGGAGGCCAA CC AGCTGGCTGAGAAAGATGAGTTTGATCATAAAAGAAAAGAAC 1800
Hsc70t      AAGGACAAGATCACGGAGTC CGATAAAAAGAAAATACTGGATAAATGCAATGAGGTTCCTTTCCTG-CTG-AGTC-AAACCCAGCTGGCTGAGAAAGTAGAGTTTGATCATAAAAGAAAAGAAC 1771

B10.A       TGGAAAATATGTGTAATCCG ATCATCACAAAACTGTACCA GAGCGGATGCACCGGGCCCA CCTGTACGCCAGGGTATACT CCCGGCAGGGCTGCCACAGG CCCTACCATCGAGGAAGTAG 1920
B10.A(1R)   TGGAAAATATGTGTAATCCG ATCATCACAAAACTGTACCA GAGCGGATGCACCGGGCCCA CCTGTACGCCAGGGTATACT CCCGGCAGGGCTGCCACAGG CCCTACCATCGAGGAAGTAG 1920
B10.A(2R)   TGGAAAATATGTGTAATCCG ATCATCACAAAACTGTACCA GAGCGGATGCACCGGGCCCA CCTGTACGCCAGGGTATACT CCCGGCAGGGCTGCCACAGG CCCTACCATCGAGGAAGTAG 1920
B10.A(4R)   TGGAAAATATGTGTAATCCG ATCATCACAAAACTGTACCA GAGCGGATGCACCGGGCCCA CCTGTACGCCAGGGTATACT CCCGGCAGGGCTGCCACAGG CCCTACCATCGAGGAAGTAG 1920
Hsc70t      TGGAAAATATGTGTAATCCG ATCATCACAAAACTGTACCA GAGCGGATGCACCGGGCCCA C-TGTACGCCAGGGTATACT CC-GGCAGGGCT-CCACAGG CCCTACCATCGAGGAAGTAG 1888

B10.A       ATTAGCCTTTTCCAGAATTG CAGGGTGCTAGGGTGCCTCT AGGCGAATTTTATTCATCTT CAAACATCAATATGATTCTT GAACTGACTGGACTCAAGCC TACGTCACCA TCCTTTGGG ATC 2043
B10.A(1R)   ATTAGCCTTTTCCAGAATTG CAGGGTGCTAGGGTGCCTCT AGGCGAATTTTATTCATCTT CAAACATCAATATGATTCTT GAACTGACTGGACTCAAGCC TACGTCACCA TCCTTTGGG ATC 2043
B10.A(2R)   ATTAGCCTTTTCCAGAATTG CAGGGTGCTAGGGTGCCTCT AGGCGAATTTTATTCATCTT CAAACATCAATATGATTCTT GAATTGACTGGACTCAAGCC TACGTCACCA TCCTTTGGG ATC 2043
B10.A(4R)   ATTAGCCTTTTCCAGAATTG CAGGGTGCTAGGGTGCCTCT AGGCGAATTTTATTCATCTT CAAACATCAATATGATTCTT GAACTGACTGGACTCAAGCC TACGTCACCA TCCTTTGGG ATC 2043
Hsc70t      ATTAGCCTTTTCCAGAATTG CAGGGTGCTAGGGTGCCTCT AGGCGAATTTTATTCATCTT CAAACATCAATATGATTCTT GAACTGACTGGACTCA-GCC TACGTCACCAGTCCTTTGG- ATC 2010
```

Figure 6.3 Alignment of the *Hsc70t* nucleotide sequences.

The nucleotide sequences obtained from the *H-2* congenic mouse strains shown are aligned with the sequence of Matsumoto and Fujimoto (1990). Nucleotide variations between the *H-2* congenic mouse strains are shown in red and variations within the published sequence are shown in blue.

number of differences (>100) between the published sequence and the sequences amplified from the DNA samples from different mouse strains. Although the published *Hsc70t* sequence shares ~95% identity with the amplified sequences, there are a large number of single base changes and small deletions and insertions (Figure 6.3). One of these changes occurs within the sequence to which oligonucleotide MHSC6 hybridises, resulting in a mismatch at the 3' end of the oligonucleotide sequence. This was probably responsible for the weak signal obtained when primer MHSC6 was used for sequencing [section 6.3.1]. Some of these changes may be due to allelic variations, but heat shock proteins are very highly conserved and the translated sequences show even greater variation [section 6.3.3]. This suggests that the differences may be due to PCR errors, cloning artifacts, or most likely sequencing errors in the original sequence.

6.3.3 Analysis of *Hsc70t* protein sequences

In order to determine the effects of the nucleotide sequence variations upon the protein sequence, the nucleotide sequences were translated into protein sequences.

The derived amino acid sequences from the B10.A, B10.A(2R) and B10.A(4R) mouse strains are all 641 residues in length, but the published *Hsc70t* sequence encodes only 630 residues. The protein sequence derived from the human *Hsc70t* homologue, HSP70-Hom, is also 641 residues in length (Milner and Campbell 1990) and agrees with the length of the protein sequences derived from the amplified sequences of *Hsc70t*. The sequence from the B10.A(1R) mouse strain contained a stop codon, TAG, at amino acid position 534 as a result of a nucleotide change at position 1600, possibly caused by an error during PCR amplification of the *Hsc70t* gene. As a result, an 'X' was substituted for the stop codon in order to allow alignment of this sequence with the others.

The derived amino acid sequences of *Hsc70t* from the B10.A, B10.A(1R), B10.A(2R) and B10.A(4R) mouse strains and the published *Hsc70t* sequence were aligned using the CLUSTAL V program (Higgins and Sharp 1988) to compare the sequences (Figure 6.4). The sequences share ~87% identity and ~91% similarity (Table 6.2). Many of the amino acid changes are non-conserved substitutions in addition to several deletions and insertions. This further indicates that many of the changes are a result of sequencing errors in the original gene sequence.

```
B10.A       MAANKGMAIGIDLGTTYSCV GVFQHGKVEIIANDQGNRTT PSYVAFTDTERLIGDAAKNQ VAMNPQNTVFDAKRLIGRKF NDPVVQSDMKLWPFQVINEA GKPKVMVSYKGEKKAFYPEE  120
B10.A(1R)   MAANKGMAIGIDLGTTYSCV GVFQHGKVEIIANDQGNRTT PSYVAFTDTERLIGDAAKNQ VAMNPQNTVFDAKRLIGRKF NDPVVQSDMKLWPFQVINEA GKPKVMVSYKGEKKAFYPEE  120
B10.A(2R)   MAANKGMAIGIDLGTTYSCV GVFQHGKVEIIANDQGNRTT PSYVAFTDTERLIGDAAKNQ VAMNPQNTVFDAKRLIGRKF NDPVVQSDMKLWPFQVINEA GKPKVMVSYKGEKKAFYPEE  120
B10.A(4R)   MAANKGMAIGIDLGTTYSCV GVFQHGKVEIIANDQGNRTT PSYVAFTDTERLIGDAAKNQ VAMNPQNTVFDAKRLIGRKF NDPVVQSDMKLWPFQVINEA GKPKVMVSYKGEKKAFYPEE  120
Hsc70t      MAANKGMAIGIDLGTTYSCV GVFQHGKVEIIADYQVNRTT PDYVAFTDTSRLI-ERSKNQ VAMNPQNTVFDAKRLIGRKF NDPVVQSNMEALAISMINEA GKPKVMVSYKGEKKAFYPEE  199
HSP70HOM   MATAKGTAIGIDLGTTYSCV GVFQHGKVEIIANDQGNRTT PSYVAFTDTERLIGDAAKNQ VAMNPQNTVFDAKRLIGRKF NDPVVQADMKLWPFQVINEG GKPKVLVSYKGENKAFYPEE  200

B10.A       ISSMVLTKMKETAEAFLGHN VTDAVITVPAYFNDSQRQAT KDAGVIAGLNVLRIINEPTA AAIAYGLDKGSHGERHVLIF DLGGGTFDVSILTIDDGIFE VKATAGDTHLGGEDFDNRLV  240
B10.A(1R)   ISSMVLTKMKETAEAFLGHN VTNAVITVPAYFNDSQRQAT KDAGVIAGLNVLRIINEPTA AAIAYGLDKGSHGERHVLIF DLGGGTFDVSILTIDDGIFE VKATAGDTHLGGEDFDNRLV  240
B10.A(2R)   ISSMVLTKMKETAEAFLGHN VTNAVITVPAYFNDSQRQAT KDAGVIAGLNVLRIINEPTA AAIAYGLDKGSHGERHVLIF DLGGGTFDVSILTIDDGIFE VKATAGDTHLGGEDFDNRLV  240
B10.A(4R)   ISSMVLTKMKETAEAFLGHN VTNAVITVPAYFNDSQRQAT KDAGVIAGLNVLRIINEPTA AAIAYGLDKGSHGEQHVLIF DLGGGTFDVSILTIDDGIFE VKATAGDTHLGGEDFDNRLV  240
Hsc70t      ISSMVLTKMKETAEVFWA-P MSQRVITVPAYFNDSQRQAT KDAGVIAGLNVLRIINEPTA VAWA-RLDKGSHAERHVLIF DLGGGTFDLSILTIDDAS-S VKPR-GDTHLGGEDFDNRLV  235
HSP70HOM   ISSMVLTKMKETAEAFLGHP VTNAVITVPAYFNDSQRQAT KDAGVIAGLNVLRIINEPTA AAIAYGLDKGGQGERHVLIF DLGGGTFDVSILTIDDGIFE VKATAGDTHLGGEDFDNRLV  240

B10.A       SHFVEEFKRKHKKDI SQNKRAVRRLRTACERAKRTLSSST QANLEIDSLYEGIDFYTSIT RARFEELCADLFRGTLEPVE KSLRDAKMDKAKIHDIVLVG GSTRIPKVQKLLQDYFNGRD  360
B10.A(1R)   SHFVEEFKRKHKKDI SQNKRAVRRLRTACERAKRTLSSST QANLEIDSLYEGIDFYTSIT RARFEELCADLFRGTLEPVE KSLRDAKMDKAKIHDIVLVG GSTRIPKVQKLLQDYFNGRD  360
B10.A(2R)   SHFVEEFKRKHKKDI SQNKRAVRRLRTACERARRTLSSST QANLEIDSLYEGIDFYTSIT RARFEELCADLFRGTLEPVE KSLRDAKMDKAKIHDIVLVG GSTRIPKVQKLLQDYFNGRD  360
B10.A(4R)   SHFVEEFKRKHKKDI SQNKRAVRRLRTACERAKRTLSSST QANLEIDSLYEGIDFYTSIT RARFEELCADLFRGTLEPVE KSLRDAKMDKAKIHDIVLVG GSTRIPKVQKLLQDYFNGRD  360
Hsc70t      SHFVR-VKRKHKKDNQAKNKRAVRRLRTACERAKRTLSSST QANLEIDSLYEGIDFYTSIT RARFEELCADLFRGTLEPVE KSLRDAKMDKAKIHDIVLVG GSTRIPKVQKMLQDYFNGRD  355
HSP70HOM   SHFVEEFKRKHKKDI SQNKRAVRRLRTACERAKRTLSSST QANLEIDSLYEGIDFYTSIT RARFEELCADLFRGTLEPVE KALRDAKMDKAKIHDIVLVG GSTRIPKVQRLLQDYFNGRD  360

B10.A       LNKSINPDEAVAYGAAVQAA ILMGDKSEKVQDLLLLDVAP LSLGLETAGGVMTVLIKRNS TIPTKQTQIFTTYSDNQPGV LIQVCEGERAMTRDNNLLGR FDLTGIPPAPRGVPQIEVTF  480
B10.A(1R)   LNKSINPDEAVAYGAAVQAA ILMGDKSEKVQDLLLLDVAP LSLGLETAGGVMTVLIKRNS TIPTKQTQILTTYSDNQPGV LIQVYEGERAMTRDNNLLGR FDLTGIPPAPRGVPQIEVTF  480
B10.A(2R)   LNKSINPDEAVAYGAAVQAA ILMGDKSEKVQDLLLLDVAP LSLGLETAGGVMTVLIKRNS TIPTKQTQIFTTYSGDQPGV LIQVYEGERAMTRDNNLLGR FDLTGIPPAPRGVPQIEVTF  480
B10.A(4R)   LNKSINPDEAVAYGAAVQAA ILMGDKSEKVQDLLLLDVAP LSLGLETAGGVMTVLIKRNS TIPTKQTQIFTTYSDNQPGV LIQVYEGERAMTRDNNLLGR FDLTGIPPAPRGVPQIEVTF  480
Hsc70t      LNKSINPDEAVAYGAAVQGA ILMGDKSEKVQDLLLLDVAP LSLGLETAGGVMTVLIKRNS TIPTKQTKIFTTYSDNQPGV LIQC---TRQTIPDNNLVGP FDLTGIPPAP-SVPQIEVTF  471
HSP70HOM   LNKSINPDEAVAYGAAVQAA ILMGDKSEKVQDLLLLDVAP LSLGLETAGGVMTALIKRNS TIPTKQTQIFTTYSDNQPGV LIQVYEGERAMTKDNNLLGR FDLTGIPPAPRGVPQIEVTF  480

B10.A       DIDANGILNVTAMDKSTGKA NKITITNDKGRLSKEEIERM VQEAERYKAEDEGQREKIAA KNALESYAFNMKSAVGDEGL KDKISESDKKKILDKCNEVL SWLEANQLAEKDEFDHKRKE  600
B10.A(1R)   DIDANGILNVTAMDKSTGKA NKITITNDKGRLSKEEIERM VQEAERYKAEDEG*REKIAA KNALESYAFNMKSAVGDEGL KDKISESDKKKILDKCNEVL SWLEANQLAEKDEFDHKRKE  600
B10.A(2R)   DIDANGVLNVTAMDKSTGKA NKITITNDKGRLSKEEIERM VQEAERYKAEDEGRGEKIAA KNALESYAFNMKSAVGDEGL KDKISESDKKKILDKCNEVL SWLEANQLAEKDEFDHKRKE  600
B10.A(4R)   DIDANGILNVTAMDKSTGKA NKITITNDKGRLSKEEIERM VQEAERYKAEDEGQREKIAA KNALESYAFNMKSAVGDEGL KDKISESDKKKILDKCNEVL SWLEANQLAEKDEFDHKRKE  600
Hsc70t      DIARNGILNVTAMDKSTGKA NKITITNDKGRLSKEEIDS- IESRTAYKREDEGQREKIAA KNALESYAFNMKSACGDEGL KDKITESDKKKILDKCNEVP FLLSQTQLAEKVEFDHKRKE  590
HSP70HOM   DIDANGILNVTATDKSTGKV NKITITNDKGRLSKEEIERM VLDAEKYKAEDEVQREKIAA KNALESYAFNMKSVVSDEGL KGKISESDKNKILDKCNELL SWLEVNQLAEKDEFDHKRKE  600

B10.A       LENMCNPIITKLYQSGCTGP TCTPGYTPGRAATGPTIEEVD  641
B10.A(1R)   LENMCNPIITKLYQSGCTGP TCTPGYTPGRAATGPTIEEVD  641
B10.A(2R)   LENMCNPIITKLYQSGCTGP TCTPGYTPGRAATGPTIEEVD  641
B10.A(4R)   LENMCNPIITKLYQSGCTGP TCTPGYTPGRAATGPTIEEVD  641
Hsc70t      LENMCNPIITKLYQSGCTGP TVRQGILR-QGSTGPTIEEVD  630
HSP70HOM   LEQMCNPIITKLYQGGCTGP ACGTGYVPGRPATGPTIEEVD  641
```

Figure 6.4 Alignment of the *Hsc70t* and HSP70-Hom amino acid sequences.

The amino acid sequences of *Hsc70t* are aligned with the human HSP70-Hom sequence. Amino acid variations between the *H-2* congenic mouse strains are shown in red, variations within the published sequence are shown in blue and amino acid substitutions within the HSP70-Hom sequence are shown in green.

A

		HSP70-Hom
	Consensus	84.9%
Hsc70t	95.0%	81.1%

B

	Hsc70t	Consensus	HSP70-Hom	
Hsc70t	-	87.0%	83.6%	
Consensus	90.8%	-	94.4%	% Identity
HSP70-Hom	88.9%	96.9%	-	
		% Similarity		

Table 6.2 Heat shock protein sequence similarities.

A Percentage nucleotide sequence identities between the *Hsc70t* sequence of Matsumoto and Fujimoto (*Hsc70t*), the consensus *Hsc70t* sequence determined from the nucleotide sequences obtained from mouse strains B10.A, B10.A(1R), B10.A(2R) and B10.A(4R) and the sequence of the human *Hsc70t* homologue HSP70-Hom.

B Percentage amino acid sequence identities and similarities between the derived amino acid sequences from the *Hsc70t* sequence of Matsumoto and Fujimoto (*Hsc70t*), the consensus *Hsc70t* sequence from mouse strains B10.A, B10.A(1R), B10.A(2R) and B10.A(4R) and the sequence of the human *Hsc70t* homologue HSP70-Hom. The percentage amino acid sequence similarity includes the substitution of chemically equivalent amino acids.

Alignment of the amino acid sequences derived from *Hsc70t* with the human homologue, HSP70-Hom, sequence of Milner and Campbell (1990) (Figure 6.4), shows that there is greater similarity between the HSP70-Hom and the amplified *Hsc70t* amino acid sequences (~94% sequence identity) than to the published *Hsc70t* amino acid sequence (~84% sequence identity). If conservative amino acid substitutions are included these values rise to ~97% and ~89%, respectively. This indicates that there is greater sequence similarity between the human sequence and the amplified mouse sequences than between the published mouse sequence and the amplified mouse sequences. The high level of sequence conservation that exists within the heat shock protein family indicates that the amplified *Hsc70t* sequences, which are more similar to HSP70-Hom, are more likely to be correct. The variations in sequence between *Hsc70t* and HSP70-Hom occur throughout the sequence (Figure 6.4) and do not appear to be clustered, except for 6 amino acid changes over 17 residues at the C-terminus which is the most divergent region in all the heat shock proteins (Milner and Campbell 1990).

6.3.4 Polymorphic analysis of *Hsc70t* in *H-2* congenic mouse strains

By comparing the *Hsc70t* nucleotide sequences amplified from DNA isolated from the mouse strains B10.A, B10.A(1R). B10.A(2R) and B10.A(4R), a total of 15 nucleotide differences between all four sequences were observed (Table 6.3), compared with over 100 differences between the sequence derived from B10.A mice (Figure 6.3) and the published *Hsc70t* sequence (Matsumoto and Fujimoto 1990). All the differences occur in only one mouse strain of the four at any one position and none of the differences correlate with susceptibility or resistance to orchitis in both the parental and progeny strains.

Comparison of the derived *Hsc70t* protein sequences from the four mouse strains revealed 11 amino acid changes (Table 6.3). Four of the nucleotide changes, at positions 540, 876, 2005 and 2034, are silent mutations. Of the amino acid changes, two are conserved changes, residues 273 and 487, one causes a stop codon at position 534 and the other 8 are non-conserved substitutions, including one at position 534. Of these non-conserved substitutions, six occur within the peptide binding domain between residues 383-552 [section 6.1]. The polymorphic analysis of the three human MHC-linked HSP70 genes by Milner and Campbell (1992a) revealed only one

		DNA					PROTEIN		
POSITION	B10.A	B10.A(1R)	B10.A(2R)	B10.A(4R)	POSITION	B10.A	B10.A(1R)	B10.A(2R)	B10.A(4R)
427	G	A	A	A	143	D	N	N	N
540	G	G	A	G	180	A	A	A	A
584	G	G	G	A	195	R	R	R	Q
818	A	A	G	A	273	K	K	R	K
876	C	C	C	T	292	G	G	G	G
1288	T	C	T	T	430	F	L	F	F
1304	A	A	G	A	435	D	D	G	D
1306	A	A	G	A	436	N	N	D	N
1334	G	A	A	A	445	C	Y	Y	Y
1459	A	A	G	A	487	I	I	V	I
1600	C	T	C	C					
1601	A	A	G	A	534	Q	*	R	Q
1603	A	A	G	A	535	R	R	G	R
2005	C	T	C	C	3' UT				
2034	C	C	C	T	3' UT				

Table 6.3 Nucleotide and amino acid sequence variations of *Hsc70t* between the *H-2* congenic mouse strains. The positions of nucleotide and amino acid changes correspond to those shown in figures 6.3 and 6.4.

polymorphic residue, at position 493 in HSP70-Hom, but no polymorphism was detected in *Hsc70t* at this position. The differences that occur within the peptide binding domain all occur at exposed positions in model '1' for the predicted structure for this domain (Rippmann *et al.* 1991) where they can interact with bound peptide or other proteins. In model '2' (Rippmann *et al.* 1991) all the differences also occur at exposed positions except for residue 534 which is buried within the structure.

6.4 DISCUSSION

6.4.1 *Hsc70t* – a candidate for the *Orch-1* locus?

Hsc70t is one of the heat shock cognate proteins which, although their expression is not upregulated following heat shock, share significant homology with the major HSP70 proteins. The localisation of the genes for *Hsc70t* in the mouse *H-2S* region (Fujimoto *et al.* 1993; Snoek *et al.* 1993) and its human homologue HSP70-Hom in the human MHC class III region (Milner and Campbell 1990), between the HSP70 genes and G7, and the abundant expression of *Hsc70t* in mouse early spermatids (Matsumoto and Fujimoto 1990), makes it a potential candidate gene for the *Orch-1* locus, responsible for susceptibility to EAO (Snoek *et al.* 1993). The absence of any variation in mRNA levels of *Hsc70t* in spermatids between mouse strains that are resistant or susceptible to orchitis (Snoek *et al.* 1993) indicated that, if *Hsc70t* is the *Orch-1* locus, then variation in the protein sequence may be responsible for susceptibility or resistance to EAO.

6.4.2 Polymorphic analysis of *Hsc70t*

The *Hsc70t* gene sequence was amplified by PCR from four mouse strains, two of which show susceptibility to EAO, B10.A and B10.A(1R), and two that show resistance to EAO, B10.A(2R) and B10.A(4R). The nucleotide sequences obtained showed few variations, but examination of the sequence obtained by Matsumoto and Fujimoto (1990) revealed a large amount of sequence variation from the *Hsc70t* sequences amplified from the *H-2* congenic mouse strains. Translation of these nucleotide sequences and comparison of the derived amino acid sequences with the human *Hsc70t* homologue, HSP70-Hom, protein sequence (Milner and Campbell

1990) revealed that the HSP70-Hom protein sequence is more similar to the sequence derived from the amplified *Hsc70t* sequences than the *Hsc70t* sequence of Matsumoto and Fujimoto (1990).

Analysis of the derived protein sequences from the *H-2* congenic mouse strains B10.A, B10.A(1R), B10.A(2R) and B10.A(4R) showed no correlation of the sequence variation with resistance or susceptibility to EAO. The variations observed are probably due to PCR errors during amplification and the number of changes, $5x10^{-5}$ per nucleotide per cycle, is comparable with the value of 10^{-5} per nucleotide per cycle observed by Erlich *et al.* (1991). Work is currently under way to sequence the *Hsc70t* gene directly from the amplification products to confirm whether or not the sequence variations observed are PCR artifacts or whether they correspond to polymorphisms that may be responsible for susceptibility or resistance to EAO.

6.4.3 Heat shock proteins and the immune system

The functions of heat shock cognate proteins is of interest in relation to the immune system and their possible roles in autoimmune disease aetiology. HSC70 participates in the translocation of peptides across intracellular membranes (Craig *et al.* 1991; Pelham 1990), plays a role in the disassembly of clathrin-coated vesicles (Brodsky *et al.* 1991) and, in common with the major HSP70 proteins, migrates to the nucleus and nucleolus in response to heat shock, where it is thought to aid the recovery of nucleolar function (Lewis and Pelham 1985). There is little evidence for the direct involvement of HSP70 proteins in the processing and presentation of antigens, although the proposed similarity of the heat shock protein peptide binding site structure with that of MHC class I molecules and their role as molecular chaperones have important implications.

MHC class I molecules usually bind and present intracellular antigenic peptides synthesised within infected cells (Townsend and Bodmer 1989) and the binding of peptide is essential for the stability of this complex and its expression at the cell surface (Townsend *et al.* 1989). BiP (GRP78) binds to the Ig heavy chain during antibody assembly (Monro and Pelham 1986) and it may be possible that another member of the heat shock protein family acts as a chaperone during MHC class I molecule assembly, by stabilising the complex before peptide binding. Alternatively, a member of the heat shock protein family may play a role in MHC class II molecule

assembly by stabilising αβ dimers in the endoplasmic reticulum, possibly after removal of the invariant chain (VanBuskirk *et al.* 1989, 1991).

A more direct role for heat shock proteins in antigen processing and presentation could be by direct interaction with the antigenic peptides. This could be in the role of peptide transport (Gething and Sambrook 1992) or in binding peptide within the proposed cleft of the peptide binding domain protecting them from complete degradation.

Evidence for direct involvement of HSC70 in the immune system has recently come from the observations of Nadler *et al.* (1992). They have shown that the potent immunosuppressant deoxyspergualin (DSG) binds to a 70kDa protein with high affinity and the hydrolysis of ATP by the protein increases the efficiency of DSG release. Partial amino acid sequences from the 70kDa protein revealed homologies with heat shock proteins, but the only protein identical with all the peptides was HSC70.

6.4.4 Heat shock proteins as autoantigens

The major antigens recognised by the immune system in response to a variety of bacterial and parasitic pathogens are heat shock proteins (Young *et al.* 1990). There is a high level of sequence conservation between heat shock proteins of the pathogens *Plasmodium falciparum*, *Mycobacterium leprae* and *Mycobacterium tuberculosis* and their human homologues. *P. falciparum* shares 60% identity with a 72kDa human heat shock protein (Bianco *et al.* 1986). This high level of similarity has resulted in the concept that 'molecular mimicry' (Oldstone 1987), due to the cross reaction of antibodies against heat shock proteins from pathogens with their human homologues, may cause autoimmune diseases. Serum from patients with SLE contain anti-HSC70 autoantibodies (Minota *et al.* 1988) and patients with rheumatoid arthritis, ankylosing spondylitis and infectious mononucleosis contain anti-HSP70 autoantibodies in their serum (Tsoulfa *et al.* 1989). However, there is no conclusive evidence for the involvement of heat shock proteins in disease pathology.

This wide variety of possible roles for *Hsc70t*, however, does not provide any indication as to its potential contribution to resistance or susceptibility to EAO.

CHAPTER 7

SUMMARY & CONCLUSIONS

The human MHC has been shown to influence susceptibility to over 40 diseases, most of which are autoimmune in nature (Tiwari and Terasaki 1985; Batchelor and McMichael 1987; Thomson 1988; Nepom and Erlich 1991). Some diseases are more strongly associated with combinations of particular class I and class II alleles (extended haplotypes) than with any single locus *e.g.* SLE with B8 and DR3, Coeliac disease with B8 and DR3 or DR7. However, more detailed molecular and immunogenetic analysis has shown that some of the strongest associations are with alleles at the HLA-B and HLA-DR loci which flank the class III region *e.g.* Ankylosing spondylitis with B27, IDDM with DQB encoding Asp at position 57. Consequently, characterisation of the protein products encoded by genes in the MHC class III region may provide some indication as to whether or not any of the disease loci do in fact lie in this region.

The class III region of the human MHC contains at least 38 genes (Figure 7.1) (Sargent *et al.* 1989a,b, Spies *et al.* 1989a,b, 1990; Kendall *et al.* 1990; Cheng 1991; Matsumoto *et al.* 1992b; Chapters 4 & 5). These genes encode proteins that are distinct from the highly polymorphic HLA class I and class II antigens and most are also unrelated to each other. Some of these proteins are involved in the immune system and are of obvious interest in the study of autoimmune diseases, but there may be other proteins, for which the function has not been determined, which may also play a role in autoimmune disease pathogenesis. In addition, some proteins encoded in the MHC class III region may not be involved in the immune system, but could still play a role in other diseases.

7.1 CHARACTERISATION OF NOVEL GENES

7.1.1 The G7 gene

The G7 gene has been located ~60kb telomeric of the HSP70 gene cluster, has been

shown to contain 27 exons that span ~22kb. The usage of two of these exons is mutually exclusive and this results in two different mRNA transcripts that encode two different novel, putative intracellular protein products of 749 and 788 amino acids. The mRNA encoding the smaller protein is more abundant in tissue culture cell lines. Expression of the G7 gene appears to be ubiquitous, although increased levels of G7 mRNA were detected in monocytic cell lines. The function of these gene products is unknown, but the putative G7 protein sequences show very significant homology with several bacterial DNA mismatch binding proteins and homologous proteins in man, mouse and yeast, as well as the *S. pombe* Swi4 protein which plays a role in mating type switching. The putative G7 protein products are smaller than the DNA mismatch repair proteins which have previously been described and do not contain a nuclear localisation signal of the type located near the N-terminus of the eukaryotic DNA mismatch repair protein homologues. In addition, the putative G7 protein products do not contain the consensus sequence of a DNA binding site such as those located near the C-terminus of all the DNA mismatch repair proteins and their homologues. The putative G7 protein products do, however, contain the type A nucleotide binding site for ATP (Saraste *et al.* 1990) and the three unique, conserved motifs which are present in all the DNA mismatch repair proteins discovered to date and their homologues. This similarity suggests that the putative G7 protein products may have a role in DNA repair even if they do not contain the DNA binding site motif present in proteins that play a role in the recognition of mismatched DNA following replication or recombination.

A *Pvu*II RFLP has been detected within G7 that was present in ~25% of the haplotypes analysed. This RFLP is associated with a potential Thr→Ala substitution at the C-terminus of the putative G7 protein products, beyond the conserved motifs identified within the DNA mismatch repair proteins and their homologues.

7.1.2 The G7b gene

The G7b gene lies adjacent to the 3′ end of the HSP70–Hom gene, and contains 5 exons that span ~10kb of DNA. There appear to be several genes or pseudogenes homologous to G7b elsewhere in the human genome.

G7b mRNA was found to be expressed in all the cell lines examined which encodes a putative polypeptide apparently 95 amino acids in length. The function of

the putative G7b polypeptide is unknown, but the absence of a hydrophobic leader peptide indicates that it could be an intracellular protein. The only significant homology detected for the putative G7b polypeptide sequence is with the human small ribonucleoprotein, Sm-D, autoantigen and this homology is over the whole of the G7b sequence. The small nuclear ribonucleoprotein complexes are associated with the small nuclear RNAs involved in RNA splicing (Rokeach *et al.* 1988) and G7b may also be involved in this process. The presence of anti-Sm-D antibodies in patients suffering from SLE is of particular interest as this is one of the many autoimmune diseases which has been shown to be associated with the MHC (Tiwari and Terasaki 1985).

7.1.3 The new novel gene G7c

Recombination mapping of the *Orch-1* locus in the mouse *H-2S* region to the *Hsp70–Bat5* interval (Snoek *et al.* 1991) led to the search for additional genes in this region. Segments of non-repetitive DNA sequence mapped within this region were found to correlate with the locations of four genes already mapped to this region. The presence of two other regions containing non-repetitive DNA sequences, located between *G7* and *G7a* and between *G6* and *Bat5*, indicated that there may be other genes in this interval. Refinement of the position of the *Orch-1* locus to the region between *Hsp70.1* and *G7* (Snoek *et al.* 1993) resulted in further analysis of the ~11kb of DNA separating the *G7* and *G7a* genes. This region was found to contain DNA sequences that were conserved across species in zoo blot analysis, indicative of coding sequences (Monaco *et al.* 1986; Rommens *et al.* 1989). Exon trapping analysis (Buckler *et al.* 1991) of part of this region resulted in the isolation of two discrete exons that were shown to correspond to part of the conserved sequences identified by zoo blot analysis. The sequences of these exons, and the translated protein sequences, do not show homology with any other known sequences.

Screening cDNA libraries by PCR and RT-PCR analysis of RNA samples derived from a variety of cell lines, using oligonucleotide primers that hybridise to the trapped exons, has subsequently indicated that G7c is expressed in cultured B lymphocyte cell lines.

7.2 NON-REPETITIVE SEQUENCES AS MARKERS FOR MAPPING NEW GENES

Current approaches for the identification of new genes within stretches of genomic DNA are largely based upon the hybridisation of cDNA libraries with whole YAC or cosmid DNA inserts (Wallace *et al.* 1990; Elvin *et al.* 1990; Lovett *et al.* 1991; Parimoo *et al.* 1991). Recombination based approaches (Vollrath *et al.* 1988; Kurnit and Seed 1990) can also be applied across large cloned regions, but neither of these strategies provide any indication of the total number of genes present. These approaches are 'one hit' analyses and can vary widely in the efficiency with which they identify new genes, often as a result of the level or tissue distribution of gene expression. By preparing uniform abundance, or normalised, cDNA libraries (Patanjali *et al.* 1991)) to increase the frequency of rare cDNA species and by combining cDNA libraries, these problems can be partly overcome.

CpG-rich HTF-islands (Bird 1986, 1987; Gardiner-Garden and Frommer 1987) can be used to identify potential genes within large regions of genomic DNA, but not all genes are HTF-island associated. Consequently, any approach based upon mapping HTF-islands to identify genes will not detect genes that are not HTF-island-associated, unless they happen to lie close enough to a HTF-island to be detected by chance *e.g.* G1 (Olavesen *et al.* 1993).

All genes contain unique or low copy number sequences that can potentially be identified as non-repetitive sequences within cloned genomic DNA and these can be used as markers to identify the genes associated with them. This approach, however, may not be applicable in the initial search for genes over a large region because if the gene density is high then the regions of non-repetitive DNA associated with each gene cannot be distinguished from one another. If, however, following the identification of genes by other approaches, regions of non-repetitive sequence are detected that do not correlate with them, attention can be focused upon these regions to search for additional genes.

Zoo blot analysis can be carried out to confirm whether sequences are conserved across species (Monaco *et al.* 1986; Rommens *et al.* 1989) and the non-repetitive sequences can either be subjected to exon trapping analysis or subcloned and sequenced. The identification of exon sequences allows primers to be synthesised to screen multiple cDNA libraries by PCR for expression of the gene.

Within the MHC class III region there is sufficient space between the genes already mapped for at least an additional ten genes, and non-repetitive sequences within these gaps are currently being analysed to search for even more genes within this region where the gene density is already ~1 every 20kb (Milner and Campbell 1992b).

The human MHC is located in the chromosome band 6p21.3 which is a light band upon Giemsa (G) staining. Most (~75%) of the genes localised by *in situ* hybridisation are also located in G-light bands (Bickmore and Sumner 1989), which have been shown to be associated with unmethylated CpG dinucleotides (Adolph and Hameister 1990), characteristic of HTF-islands which are frequently associated with the 5′ ends of genes. If the gene density of ~1 every 20kb is consistent within all the G-light bands, which constitute ~42% of chromosomal DNA, and these contain ~75% of the genes, then the human genome may contain ~80,000 genes. This is significantly more than the 50,000-60,000 genes estimated upon the frequency of CpG islands in the human genome (Larsen *et al.* 1992).

7.3 COLINEARITY OF GENES IN THE HUMAN AND MOUSE CLASS III REGIONS

The human MHC on the short arm of chromosome 6 and the mouse *H-2* complex on chromosome 17 span ~3.5Mb (Milner and Campbell 1992b) and at least 2.3Mb (Muller *et al.* 1987a), respectively, and are arranged in a similar fashion, with the exception of the mouse *H-2K* class I genes which lie centromeric of the *H-2I* class II region (Weiss *et al.* 1984). The colinearity of the MHC class II regions in man and mouse was studied by Hanson and Trowsdale (1991), who observed striking similarity between the two species. The exceptions were the absence of *DPA1*, *DPB1*, *DPA2* and *DQB2*, *DQA2*, *DQB2* cluster homologues in the mouse and the insertion of the *H2-K* and *H2-K2* class I genes between the *RING1* homologue and *KE3*. The class I regions in man and mouse are less well characterised, but both lie telomeric of their respective class III regions, with the exception of the mouse *H-2K* region, described above.

The organisation of the class III region of the human and mouse MHCs also appears to be highly conserved (Figure 7.1). At the telomeric end, the novel human gene BAT1 (Spies *et al.* 1989a), D6S81E, has been shown to have a mouse homologue

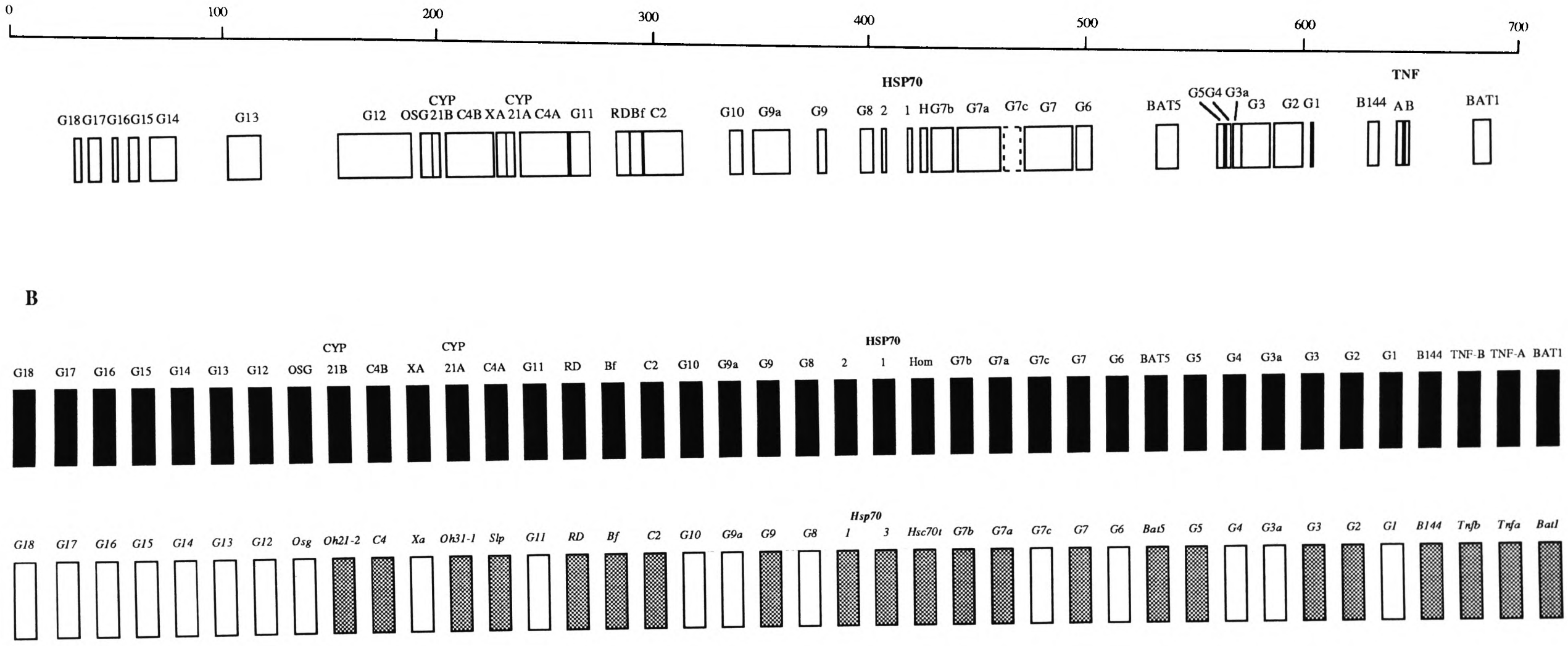


Figure 7.1 The organisation of the human MHC class III region and the mouse *H-2S* region.
A Molecular map of the human MHC class III region. The positions of genes identified in this region are indicated by open boxes. G7c is shown as a dashed box as the expression of this gene has not been confirmed.
B Schematic map of the human MHC class III region (top) and the mouse *H-2S* region (below). The relative positions of the human genes are shown by filled boxes and the corresponding positions of mouse homologues of the human genes are shown as shaded boxes. The open boxes indicate the relative positions for the other human genes not yet mapped in the mouse *H-2S* region.

at the same position (Wroblewski *et al.* 1990). However, the distance from BAT1 to HLA-B in man is much larger, 170kb, than the *Bat1* to *H-2D* interval of 10kb in mouse (Wroblewski *et al.* 1990). The human TNF-A and -B loci (Carroll *et al.* 1987) and the mouse *Tnfa* and *Tnfb* loci (Muller *et al.* 1987b) both lie ~35kb from the corresponding D6S81E loci (Wroblewski *et al.* 1990). Centromeric of the TNF genes, the mouse B144 gene and its human homologue have been identified (Tsuge *et al.* 1987; Sargent *et al.* 1989a). The locations of G2/BAT2, G3/BAT3, G5/BAT4 and BAT5 were mapped in cosmids by Sargent *et al.* (1989a) and Spies *et al.* (1989a) and the corresponding positions of the mouse homologues for all these genes except G6 have been demonstrated by PFGE and RFLP analysis by Snoek *et al.* (1991) and Lafuse *et al.* (1992).

The interval from BAT5 to the HSP70 gene cluster contains G6, G7, G7a/BAT6 and G7b in man (Sargent *et al.* 1989a; Spies *et al.* 1989b; Hsieh and Campbell 1991; Chapter 4) and homologues have been mapped, in cosmids linked to the *Hsp70* gene cluster, to the corresponding positions in mouse, although the *G7a-G7b* interval is larger in mouse than in man (Snoek *et al.* 1993). The organisation of the HSP70 gene cluster, consisting of HSP70-2, HSP70-1 and HSP70-Hom, determined by Sargent *et al.* (1989b) and Milner and Campbell (1990), is the same in the mouse, where the corresponding genes are *Hsp70.1*, *Hsp70.3* and *Hsc70t* (Snoek *et al.* 1993). A mouse homologue to the human G9 gene, mapped centromeric of the HSP70 gene cluster (Sargent *et al.* 1989a), has been mapped to a corresponding position in the cosmids linked to the *Hsp70* gene cluster (Snoek *et al.* 1993).

In the mouse, most of the genes in the *H-2* region have been mapped by RFLP analysis of intra *H-2* recombinant mouse strains using human cDNA probes. Most of the recombinations have taken place telomeric of the *Hsp70* gene cluster so that the relative positions of the majority of the novel genes centromeric of the *Hsp70* gene cluster have not yet been mapped.

Within the complement cluster, the organisation and orientation of the C2, C4 factor B and steroid 21-hydroxylase genes (Carroll *et al.* 1984, 1987; Dunham *et al.* 1987) have been shown to be the same in mouse and man (Muller *et al.* 1987a). An additional, novel gene, RD, has been mapped within this cluster between the factor B and C4A/*Slp* genes in both mouse and man (Levi-Strauss *et al.* 1988).

Thus, all the genes that have been shown to have homologues in both man and

mouse are colinear. The only major variations that have been identified are in the intergenic distances separating the D6S81E locus and the most proximal class I locus, and the G7a-G7b interval. Both of these differences are large enough to accommodate further genes. In addition, the colinearity of these and the novel genes not yet mapped in the mouse remains to be demonstrated. However, the total conservation of colinearity so far observed may be expected to be maintained for the other genes. This high level of conservation allows direct comparison of disease loci in man and mouse by extrapolation of results obtained by genetic mapping and physical mapping from one species to another.

7.4 THE MHC CLASS III REGION AND DISEASE ASSOCIATION

In the mouse, intra *H-2* recombinant congenic mouse strains, associated with different genetic traits, are being used to map disease loci by mapping the recombination break points using RFLPs and polymorphic $(CA)_n$ repeats. These loci include: *Hh-1*, the haemopoietic histocompatibility locus controlling natural killer cell-mediated resistance to haemopoietic allografts (Rembecki *et al.* 1988), which has been mapped to the *Bat5–H-2D* interval (Snoek *et al.* 1991), *Neu-1*, the locus controlling the degree of sialylation of lysosomal enzymes, mapped to the C2–Bat5 interval (Lafuse *et al.* 1992), the TNP-Ficoll response control locus (Hillstrom-Shapiro *et al.* 1985), mapped between C4 and Bat5 (Snoek *et al.* 1991), *Cps-1*, one of the loci that influences susceptibility to cortisone-induced cleft palate (Gasser *et al.* 1988), mapped between the *Hsp70* genes and *Bat5* (Snoek *et al.* 1991) and *Orch-1*, the EAO susceptibility locus, which has been mapped to the ~50kb region between, but not including, *Hsp70.3* and *G7* (Snoek *et al.* 1993; Snoek unpublished data) *i.e. Hsc70t*, *G7b*, *G7a* and probably *G7c*. Of these genes *Hsc70t* is the most promising candidate gene for the *Orch-1* locus as it is expressed at high levels in the testis, but no polymorphism associated with the susceptible or resistant mouse strains has been detected within the coding region of the gene.

In man, fewer disease loci have been mapped with certainty within the MHC class III region. However, the human gene G9 encodes a polypeptide that shares 22% sequence identity, over its whole length, with the sialidase of *Clostridium*

perfringens. and it is possible that G9 is the human homologue of the *Neu-1* locus. In addition, CYP21B encodes the enzyme steroid 21-hydroxylase and deficiency or dysfunction of this protein results in congenital adrenal hyperplasia (New *et al.* 1982).

The *EC1* locus, regulating susceptibility or resistance to lysis by natural killer cell clones was mapped between the factor B and HLA-A genes (Ciccone *et al.* 1990) and has subsequently shown strong association with the HLA-C locus (Colonna *et al.* 1992b). It has also been suggested that the human *EC1* locus may have a mouse counterpart in the *Hh-1* locus (Ciccone *et al.* 1990) and it is interesting to note that both *EC1* and *Hh-1* have been mapped towards the telomeric end of the mouse and human class III regions.

IgA deficiency (IgA-D) and common variable immunodeficiency (CVID) are believed to be caused by a single underlying genetic defect (Schaffer *et al.* 1989; Volonakis *et al.* 1992). Analysis of individuals with these immunodeficiencies revealed an association with rare C2 and C4A alleles and deletions of C4A and CYP21, indicating that IgA-D and CVID are related disorders for which susceptibility is determined by a gene or genes within or near the MHC class III region (Schaffer *et al.* 1989). Subsequent analysis of families including individuals with IgA-D and CVID using polymorphic markers across the MHC class III region suggested that the common gene or genes may lie between the C4B and C2 genes (Volonakis *et al.* 1992). Detailed characterisation of the novel genes in this C4B–C2 interval is currently being carried out to try to identify the possible susceptibility gene or genes.

The linkage disequilibrium observed between the HLA class I and class II antigens and the presence of so many closely linked genes in the class III region has made the task of identifying potential disease susceptibility loci very difficult. In addition, the derived amino acid sequences from all of the novel genes identified has not yet assisted in the identification of any of the disease susceptibility loci within the class III region. However, the identification of highly polymorphic dinucleotide repeat markers within the class III region (Hsieh 1992) and RFLPs (Cross *et al.* 1991) that subdivide the extended haplotypes will help in identifying the disease susceptibility loci.

Some of the loci currently mapped to the mouse class III region appear to be regulatory loci, whose products may have a secondary effect upon other proteins,

rather than encoding extracellular proteins which may be easier to identify. As a result the identification of these loci and determination of their function may be difficult. However, ~70kb of DNA within the human MHC class III region, encoding G7, G7b, G7c and HSP70-Hom, has been characterised in detail to facilitate identification of human homologues of the loci currently mapped to the corresponding region of the mouse *H-2* complex. In addition, human cDNA probes for these genes have been used to refine the location of the *Orch-1* locus within the mouse *H-2S* region which in man may control susceptibility to idiopathic infertility. A new potential candidate gene for the *Orch-1* locus, G7c, has also been identified in man using non-repetitve sequences mapped in cloned genomic DNA to identify the location of the new gene. This approach is currently being applied to other intergenic intervals within the MHC class III region in order to identify other new genes, possibly expressed in a tissue specific manner, that may contribute to disease susceptibility.

REFERENCES

REFERENCES

Abraham LJ, Chin Du D, Zahedi K, Dawkins RL, Whitehead AS (1991) *Immunogenet.* **33** 50-53

Accolla RS, Auffray C, Singer DS, Guardiola J (1991) *Immunol. Today* **12** 97-99

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) *Science* **252** 1651-1656

Adolph S, Hameister H, (1990) *Cytogenet. Cell Genet.* **54** 132-136

Ahn NG, Teller DC, Bienkowski MJ, McMullen BA, Lipkin EW, de Haen C (1988) *J. Biol. Chem.* **263** 18657-18663

Alper CA (1976) *J. Exp. Med.* **144** 111-115

Arnold D, Driscoll, Androlewicz M, Hughes E, Cresswell, Spies T (1992) *Nature* **360** 171-174

Attaya M, Jameson S, Martinez CK, Hermel E, Aldrich C, Forman JF, Fischer Lindahl K, Bevan MJ, Monaco JJ (1992) *Nature* **355** 647-649

Auch D, Reth M (1989) *Nucleic Acids Res.* **18** 6743-6744

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (1989) *Short Protocols in Molecular Biology.* Greene Publishing Associates with Wiley-Interscience

Babbitt BP, Allen PM, Matsueda M, Haber E, Unanue ER (1985) *Nature* **317** 359-360

Bach FH, van Rood JJ (1976) *New Eng. J. Med.* **295** 806-813, 872-878, 927-936

Bahram S, Arnold D, Bresnahan M, Strominger JL, Spies T (1991) *Proc. Natl. Acad. Sci. USA* **88** 10094-10098

Banjeri J, Sands J, Strominger JL, Spies T (1990) *Proc. Natl. Acad. Sci. USA* **87** 2374-2378

Barbosa MS, Lowy DR, Schiller JT (1989) *J. Virol.* **63** 1404-1407

Barton GJ, Sternberg MJE (1990) *J. Mol. Biol.* **212** 389-402

Barton GJ (1990) *Methods Enzymol.* **183** 403-428

Barton GJ (1993) *Protein Engineering* **6** 37-40

Batchelor JR, McMichael AJ (1987) *Brit. Med. Bull.* **43** 156-183

Beck S, Hanson IM, Kelly A, Pappin DJ, Trowsdale J (1992) *DNA Seq.* **2** 203-210

Beckmann JS, Weber JL (1992) *Genomics* **12** 627-631

Bell GI, Karam JH, Rutter WJ (1981) *Proc. Natl. Acad. Sci. USA* **78** 5759-5763

Bell JI, Denney D, MacMurray A, Foster L, Watling D, McDevitt HO (1987) *J. Immunol.* **139** 562-573

Belt KT Yu CY, Carroll MC, Porter RR (1985) *Immunogenet.* **21** 173-180

Benacerraf B (1981) *Science* **212** 1229-1238

Benoist C, Mathis D (1990) *Ann. Rev. Immunol.* **8** 681-715

Bergman RK, Munoz JJ, Portis JL (1978) *Infect. Immun.* **21** 627-637

Beutler B, Cerami A (1989) *Ann. Rev. Immunol.* **7** 625-655

Bianco AE, Favaloro JM, Burkot TR, Culvenor JG, Crewther PE, Brown GV, Anders RF, Coppel RL, Kemp DJ (1986) *Proc. Natl. Acad. Sci. USA* **83** 8713-8717

Bickmore WA, Sumner AT (1989) *Trends Genet.* **5** 144-148

Bigazzi PE, Kosuda LL, Hsu KC, Andres GA (1976) *J. Exp. Med.* **143** 382-404

Birboim HC , Doly J (1979) *Nucleic Acids Res.* **7** 1513-1523

Bird AP (1986) *Nature* **321** 209-213

Bird AP (1987) *Trends Genet.* **3** 342-347

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC (1987a) *Nature* **239** 506-512

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC (1987b) *Nature* **239** 512-518

Bjorkman PJ, Parham P (1990) *Ann. Rev. Biochem.* **59** 253-288

Britten RJ, Baron WF, Stout DB, Davidson EH (1988) *Proc. Natl. Acad. Sci. USA* **85** 4770-4774

Brodsky FM, Hill BL, Acton SL, Nathke I, Wong DH, Ponnambalam S, Parham P (1991) *Trends Biochem. Sci.* **16** 208-213

Brown JH, Jardetzky TS, Saper MA, Samraoui B, Bjorkman PJ, Wiley DC (1988) *Nature* **332** 845-850

Brown MG, Driscoll J, Monaco JJ (1991) *Nature* **353** 355-357

Brunak S, Engelbrecht J, Knudsen S (1991) *J. Mol. Biol.* **220** 49-65

Buchman AR, Berg P (1988) *Mol. Cell. Biol.* **8** 4395-4405

Buckler AJ, Chang DD, Graw SL, Brook JD, Haber DA, Sharp PA, Housman DE (1991) *Proc. Natl. Acad. Sci. USA* **88** 4005-4009

Burns AL, Magendzo K, Shirvan A, Srivastava M, Rojas E, Alijani MR, Pollard HB (1989) *Proc. Natl. Acad. Sci. USA* **86** 3798-3802

Campbell RD (1987) *Brit. Med. Bull.* **43** 37-49

Campbell RD Law SKA, Reid KBM, Sim RB (1988) *Ann. Rev. Immunol.* **6** 161-195

Campbell RD, Dunham I, Kendall E, Sargent CA (1990) *Expl. Clin. Immunogenet.* **7** 69-84

Carroll MC, Campbell RD, Bentley DR, Porter RR (1984) *Nature* **307** 237-241

Carroll MC, Campbell RD, Porter RR (1985) *Proc. Natl. Acad. Sci. USA* **82** 521-525

Carroll MC, Katzman P, Alicot EM, Koller BH, Geraghty D, Orr HT, Strominger JL, Spies T (1987) *Proc. Natl. Acad. Sci. USA* **84** 8535-8539

Cavener DR, Ray SC (1991) *Nucleic Acids Res.* **19** 3185-3192

Cerundolo V, Alexander J, Anderson K, Lamb C, Cresswell P, McMichael A, Gotch F, Townsend ARM (1990) *Nature* **345** 449-452

Chappell TG, Konforti BB, Schmid SL, Rothman JE (1987) *J. Biol. Chem.* **262** 746-751

Cheng J (1991) *D. Phil Thesis*, Oxford

Chimini G, Boretto J, Marguet D, Lanau F, Lanquin G, Pontarotti P (1990) *Immunogenet.* 32 419-426

Chirgwin JM, Przybyla, AE, MacDonald RJ, Rutter WJ (1979) *Biochem.* **18** 5294-5299

Chorney MJ, Sawada I, Gillespie GA, Srivastava R, Pan J, Weissman SM (1990) *Mol. Cell Biol.* **10** 243-253

Churchill MEA, Travers AA (1991) *Trends Biochem. Sci.* **16** 92-97

Ciccone E, Colonna M, Viale O, Pende D, Di Donato C, Reinharz D, Amoroso A, Jeannet M, Guardiola J, Moretta A, Spies T, Strominger J, Moretta L (1990) *Proc. Natl. Acad. Sci. USA* **87** 9794-9797 (Correction *Proc. Natl. Acad. Sci. USA* **88** 5477)

Collier S, Sinnott PJ, Dyer PA, Price DA, Harris R, Strachan T (1989) *EMBO J.* **8** 1393-1402

Colonna M, Bresnahan M, Bahram S, Strominger JL, Spies T (1992a) *Proc. Natl. Acad. Sci. USA* **89** 3932-3936

Colonna M, Spies T, Strominger JL, Ciccone E, Moretta A, Moretta L, Pende D, Viale O (1992b) *Proc. Natl. Acad. Sci. USA* **89** 7983-7985

Colonna M, Ferrara GB, Strominger JL, Spies T (1993) In *Proceedings of the 11th International Histocompatibility Workshop* Oxford University Press, Oxford

Corden J, Wasylyk B, Buchwalder A, Sassone-Corsi P, Kedinger C, Chambon P (1980) *Science* **209** 1406-1414

Craig EA, Gross CA (1991) *Trends Biochem. Sci.* **16** 135-140

Cresswell P, Turner MJ, Strominger JL (1973) *Proc. Natl. Acad. Sci. USA* **70** 1603-1607

Crompton MR, Owens RJ, Totty NF, Moss SE, Waterfield MD Crumpton MJ (1988) *EMBO J.* **7** 21-27 (Erratum *EMBO J.* **7** 1914)

Cross SJ, Edwards JH, Bentley DR, Campbell RD (1985) *Immunogenet.* **21** 39-49

Cross SJ, Tonks S, Trowsdale J, Campbell RD (1991) *Immunogenet.* **34** 376-384

Dausset J (1981) *Science* **213** 1469-1474

David-Watine B, Israel A, Kourilsky P (1990) *Immunol. Today* **11** 286-292

Dawkins RL,Christiansen FT, Kay PH, Garlepp MJ, McCluskey J, Hollingsworth PN, Zilko PJ (1983) *Immunol. Rev.* **70** 1-22

Dayhoff MO, Scwartz RM, Orcutt BC (1978) In *Atlas of Protein Sequence and Structure* **5** 345 National Biomedical Research Foundation

Dayhoff MO, Barker WC, Hunt LT (1983) *Methods Enzymol.* **91** 524-545

Demant P, Capkova J, Hinzova E, Voracova B (1973) *Proc. Natl. Acad Sci. USA* **70** 863-864

Demant P (1985) *Immunogenet.* **22** 183-188

Denizot F, Mattei MG, Vernet C, Pontarotti P, Chimini G (1992) *Genomics* **14** 857-862

Deshaies RJ, Koch BD, Werner-Washburne M, Craig EA, Schekman R (1988) *Nature* **332** 800-805

Devereux J, Haeberli P, Smithies O (1984) *Nucleic Acids Res.* **12** 387-395

Deverson EV, Gow IR, Coadwell J, Monaco JJ, Butcher GW, Howard JC (1990) *Nature* **348** 738-741

Dingwall C, Laskey RA (1991) *Trends Biochem. Sci.* **16** 478-481

Dodds AW, Law SKA, Porter RR (1985) *EMBO J.* **4** 2239-2244

Dooher GB, Artzt K, Bennet D Hurtenbach U (1981) *J. Reprod. Fertil.* **62** 505-511

Dunham I, Sargent CA, Trowsdale J, Campbell RD (1987) *Proc. Natl. Acad Sci. USA* **84** 7237-7241

Dunham I, Sargent CA, Dawkins RL, Campbell RD (1989) *Genomics* **5** 787-796

Dunham I, Sargent CA, Kendall E, Campbell RD (1990) *Immunogenet.* **32** 175-182

Duyk GM, Kim S, Myers R, Cox DR (1990) *Proc. Natl. Acad Sci. USA* **87** 8995-8999

Dworniczak B, Mirault M-E (1987) *Nucleic Acids Res.* **15** 5181-5197

Dykes DP, De Furio M, Polesky HF (1983) *Am. J. Hum. Genet.* **35** 652-655

Earnshaw WC, Sullivan KF, Machlin PS, Cooke CA, Kaiser DA, Pollard TD, Rothfield NF, Cleveland DW (1987) *J. Cell. Biol.* **104** 817-829

Earnshaw WC (1987) *J. Cell. Biol.* **105** 1479-1489

Economou EP, Bergen AW, Warren AC, Antonarakis SE (1990) *Proc. Natl. Acad Sci. USA* **87** 2951-2954

Ellis RJ, Hemingsen SM (1989) *Trends Biochem. Sci.* **14** 339-342

Elvin P, Slynn G, Black D, Graham A, Butler R, Riley J, Anand R, Markham AF (1990) *Nucleic Acids Res.* **18** 3913-3917

Erlich HA, Gelfand D, Sninsky JJ (1991) *Science* **252** 1643-1651

Feinberg AP, Vogelstein B (1984) *Anal. Biochem.***137** 266-267

Figueroa P, Klein D, Tewarson S, Klein J (1982) *J. Immunol.* **129** 2089-2093

Figueroa F, Klein J (1986) *Immunol. Today* **7** 78-81

Flaherty KM, DeLuca-Flaherty C, McKay DB (1990) *Nature* **346** 623-628

Flaherty KM, McKay DB, Kabsch W, Holmes KC (1991) *Proc. Natl. Acad Sci. USA* **88** 5041-5045

Flajnik MF, Canel C, Kramer J, Kasahara M (1991) *Immunogenet.* **33** 295-300

Fleck O, Michael H, Heim L (1992) *Nucleic Acids Res.* **20** 2271-2278

Flynn GC, Chappell TG, Rothman JE (1989) *Science* **245** 385-390

Fourney RM, Miyakoshi J, Day RS, Paterson MC (1988) *Focus* (Bethesda Research Laboratories) **10** 5-7

Freemont PS, Hanson IM, Trowsdale J (1991) *Cell* **64** 483-484

Fremont DH, Matsumura M, Stura EA, Peterson PA, Wilson IA (1992) *Science* **257** 917-927

Fritz TE, Lombard LS, Tyler SA, Norris WP (1976) *Exp. Mol. Pathol.* **24** 142-158

Fujii H, Shimada T (1989) *J. Biol. Chem.* **264** 10057-10064

Fujimoto H, Ito Y, Ando A, Matsumoto M, Fujio K, Miura K, Shirai M, Inoko H (1993) In *Proceedings of the 11th International Histocompatibility Workshop* Oxford University Press, Oxford

Gardiner-Garden M, Frommer M (1987) *J. Mol. Biol.* **196** 261-282

Garrett TPJ, Saper MA, Bjorkman PJ, Strominger JL, Wiley DC (1989) *Nature* **342** 692-696

Gaskins HR, Monaco JJ, Leiter EH (1992) *Science* **256** 1826-1828

Gasser DL, Yadvish KN, Trammell MA, Goldman AS (1988) *Teratology* **38** 571-577

Geraghty DE, Wei X, Orr HT, Koller BH, (1990) *J. Exp. Med.* **171** 1-18

Geraghty DE, Pei J, Lipsky B, Hansen JA, Taillon-Miller P, Bronson SK, Chaplin DD (1992a) *Proc. Natl. Acad. Sci. USA* **89** 2669-2673

Geraghty DE, Koller BH, Hansen JA, Orr HT (1992b) *J. Immunol.* **149** 1934-1946

Gething M-J, Sambrook J (1992) *Nature* **355** 33-45

Gitelman SE, Bristow J, Miller WL (1992) *Mol. Cell. Biol.* **12** 2124-2134

Glynne R, Powis SH, Beck S, Kelly A, Kerr L-A, Trowsdale J (1991) *Nature* **353** 357-360

Goodfellow PN, Jones EA, van Heyningen V, Solomon E, Bobrow M, Miggiano V, Bodmer WF (1975) *Nature* **254** 267-269

Gorer PA, Lyman S, Snell GD (1948) *Proc. Royal Soc. London* **135** 499-505

Grilley M, Welsh KM, Su SS, Modrich P (1989) *J. Biol. Chem.* **264** 1000-1004

Grundmann U, Abel K-J, Bohn H, Loebermann H, Lottspeich F, Kuepper H (1988) *Proc. Natl. Acad. Sci. USA* **85** 3708-3712

Guo H-C, Jardetzsky TS, Garrett TPJ, Lane WS, Strominger JL, Wiley DC (1992) *Nature* **360** 364-366

Gussow D, Rein RS, Meijer I, de Hoog W, Seemann GHA, Hochstenbach FM, Pleogh HL (1987) *Immunogenet.* **25** 313-322

Haber LT, Pang PP, Sobell DI, Mankovich JA, Walker GC (1988) *J. Bacteriol.* **170** 197-202

Haber LT, Walker GC (1991) *EMBO J.* **10** 2707-2715

Hamaguchi M, Sakamoto H, Tsuruta H, Sasaki H, Muto T, Sugimura T, Terada M (1992) *Proc. Natl. Acad. Sci. USA* **89** 9779-9783

Hamman HC, Gaffey LC, Lynch KR, Creutz CE (1988) *Biochem. Biophys. Res. Commun.* **156** 660-667

Hanson IM, Gorman P, Lui VCH, Cheah KSE, Solomon E, Trowsdale J (1989) *Genomics* **5** 925-931

Hanson IM, Trowsdale J (1991) *Immunogenet.* **34** 5-11

Hanson IM, Poustka A, Trowsdale J (1991) *Genomics* **10** 417-424

Harding CV, Collins DS, Slot JW, Geuze HJ, Unanue ER (1991) *Cell* **64** 393-401

Hardy DA, Bell JI, Long EO, Lindsten T, McDevitt HO (1986) *Nature* **323** 453-455

Harlow E, Lane D (1988) *Antibodies: A Laboratory Manual*. Cold Spring Harbour Laboratory Press New York

Hauptmann R, Maurer-Fogy I, Krystek E, Bodo G, Andree H, Reutelingsperger CPM (1989) *Eur. J. Biochem* **185** 63-71

Higgins CF, Gallagher MP, Hyde SC, Mimmack ML, Pearce SR (1990) *Phil. Trans. R. Soc.* **326** 353-365

Higgins DG, Sharp PM, (1988) *Gene* **73** 237-244

Hillstrom-Shapiro L, Dugan ES, Neiderhuber JE (1985) *J. Exp. Med.* **162** 1477-1493

Hoeijmakers JHJ, Bootsma D (1992) *Nature Genetics* **1** 313-314

Hood L, Steinmetz M, Malissen B (1983) *Ann. Rev. Immunol.* **1** 529-568

Houlihan JM, Biro PA, Fergar-Payne A, Simpson KL, Holmes CH (1992) *J. Immunol.* **149** 668-675

Hsieh S-L, Campbell RD (1991) *Biochem. J.* **278** 809-816

Hsieh SL (1992) *D. Phil Thesis*, Oxford

Huber R, Romisch J, Paques EP (1990) *EMBO J.* **9** 3867-3874

Huber R, Berendes R, Burger A, Schneider M, KarshikovA, Luecke H, Romisch J, Paques EP (1992) *J. Mol. Biol.* **223** 683-704

Janatipour M, Naumov Y, Ando A, Sugimura K, Okamoto N, Tsuji K, Abe K, Inoko H (1992) *Immunogenet.* **35** 272-278

Jardetzky TS, Lane WS, Robinson RA, Madden DR, Wiley DC (1991) *Nature* **353** 326-329

Johnston PA, Perin MS, Reynolds GA, Wasserman SA, Suedhof TC (1990) *J. Biol. Chem.* **265** 11382-11388

Jones NC, Rigby PWJ, Ziff EB (1988) *Genes Dev.* **2** 267-281

Jurka J, Smith T (1988) *Proc. Natl. Acad. Sci. USA* **85** 4775-4778

Kappes D, Strominger JL (1988) *Ann. Rev. Biochem.* **57** 991-1028

Kaufman JF, Auffray C, Korman AJ, Shackleford DA, Strominger J (1984) *Cell* **36** 1-13

Kelly AP, Monaco JJ, Cho S, Trowsdale J (1991a) *Nature* **353** 571-573

Kelly AP, Powis SH, Glynne R, Radley E, Beck S, Trowsdale J (1991b) *Nature* **353** 667-668

Kelly AP, Powis SH, Kerr L-A, Mockridge I, Elliott T, Bastin J, Uchanska-Zeigler B, Zeigler A, Trowsdale J, Townsend ARM (1992) *Nature* **355** 641-644

Kendall E, Sargent CA, Campbell RD (1990) *Nucleic Acids Res.* **18** 7251-7257

Kendall E, Todd JA, Campbell (1991) *Immunogenet.* **34** 349-357

Kendall E (1992) *D. Phil. Thesis*, Oxford

Khanna A, Campbell RD (1993) In *Proceedings of the 11th International Histocompatibility Workshop* Oxford University Press, Oxford

Khono S, Munoz JA, Williams TM, Teuscher C, Bernard CCA, Tung KSK (1983) *J. Immunol.* **130** 2675-2682

Kleijmeer MJ, Kelly A, Geuze HJ, Slot JW, Townsend ARM, Trowsdale J (1992) *Nature* **357** 342-344

Klein J (1975) *Biology of the Mouse H-2 Complex.* Springer-Verlag, New York

Klein J, Figueroa F, David CS (1983) *Immunogenet.* **17** 553-596

Klein J, Figueroa F (1986) *Immunol. Today* **7** 41-44

Knott V, Rees DJG, Cheng Z, Brownlee GG (1988) *Nucleic Acids Res.* **16** 2601-2612

Koller BH, Geraghty DE, Shimuzu Y, DeMars R, Orr HT, (1988) *J. Immunol.* **141** 897-904

Koller BH, Geraghty DE, DeMars R, Duvick L, Rich SS, Orr HT (1989) *J. Exp. Med.* **169** 469-480

Kozak M (1991) *J. Cell. Biol.* **115** 887-903

Krainer AR, Maniatis T (1988) In *Transcription and Splicing* Ed Hames BD, Glover DM. IRL Press, Oxford 131-206

Krangle MS, Orr HT, Strominger JL (1979) *Cell* **18** 979-991

Kurnit DM, Seed B (1990) *Proc. Natl. Acad. Sci. USA* **87** 3166-3169

Kvist S, Wiman K, Claesson L, Peterson PA, Dobberstein B (1982) *Cell* **28** 61-69

Laemmli EK (1970) *Nature* **227** 680-685

Lafuse WP, Lanning D, Spies T, David CS (1992) *Immunogenet.* **36** 110-116

Langer T, Neupert W (1991) *Current Topics Microbiol. Immunol.* **167** 3-30

Larsen F, Gundersen G, Lopez R, Prydz H (1992) *Genomics* **13** 1095-1107

Laskey RA, Mills AD (1977) *FEBS Letts.* **82** 314-316

Law SKA, Dodds AW, Porter RR (1984) *EMBO J.* **3** 1819-1823

Law SKA, Reid KBM (1988) *Complement.* IRL Press, Oxford

Lawlor DA, Zemmour J, Ennis PD, Parham P (1990) *Ann. Rev. Immunol.* **8** 23-63

Lawrance SK, Smith CL, Srivastava R, Cantor CR, Weissman SM (1987) *Science* **235** 1387-1390

Leustek T, Dalie B, Amir-Shapira D, Brot N, Weissbach H (1989) *Proc. Natl. Acad. Sci. USA* **86** 7805-7808

Levi-Strauss M, Carroll MC, Steinmetz M, Meo T (1988) *Science* **240** 201-204

Levy F, Gabathuler R, Larsson R, Kvist S (1991) *Cell* **67** 265-274

Lewis MJ, Pelham HRB (1985) *EMBO J.* **4** 3137-3143

Lim HM, Pene JJ (1988) *Gene Anal. Techniques* **5** 32-39

Linton JP, Yen J-YJ, Selby E, Chen Z, Chinsky JM, Liu K, Kellems RE, Crouse GF (1989) *Mol. Cell. Biol.* **9** 3058-3072

Locker J, Buzard G (1990) *J. DNA Sequencing & Mapping* **1** 3-11

Lotteau V, Teyton L, Peleraux A, Nilsson T, Karlsson L, Schmid SL, Quaranta V, Peterson PA (1990) *Nature* **348** 600-605

Lovett M, Kere J, Hinton LM (1991) *Proc. Natl. Acad. Sci. USA* **88** 9628-9632

Madden DR, Gorga JC, Strominger JL, Wiley DC (1991) *Nature* **353** 321-325

Mahi-Brown CA, Yule TD, Tung KSK (1987) *Cell. Immunol.* **106** 408-419

Martinez CK, Monaco JJ (1991) *Nature* **353** 664-667

Matsumoto M, Fujimoto H (1990) *Biochem. Biophys. Res. Commun.* **166** 43-49

Matsumoto K, Ishihara N, Ando A, Inoko H, Ikemura T (1992a) *Immunogenet.* **36** 400-403

Matsumoto K, Arai M, Ishihara N, Ando A, Inoko H, Ikemura T (1992b) *Genomics* **12** 485-491

Matsumura M,Fremont DH, Peterson PA, Wilson IA (1992) *Science* **257** 927-934

Mauff G, Alper CA, Awdeh Z, Batchelor JR, Bertrams T, Braun-Petersen G, Dawkins RL, Demant P, Edwards J, Grosse-Wilde H, Hauptmann G, Klouda P, Lamm L, Mollenhauer E, Nerl C, Olaisen B, O'Neill GO, Ritner C, Roos MH, Skanes V, Teisberg P, Wells L (1983) *Immunobiol.* **164** 72-79

McDevitt HO, Chinitz A (1969) *Science* **163** 1207-1208

McDevitt HO, Deak BD, Shreffler DC, Klein J, Stimpfling JH, Snell GD (1972) *J. Exp. Med.* **135** 1259-1278

McKnight SL, Kingsbury R (1982) *Science* **217** 316-324

Messer G, Spengler U, Jung MC, Holand G, Blomer K, Pape GR, Riethmuller G, Weiss EH (1991) *J. Exp. Med.* **173** 209-219

Messer G, Zemmour J, Orr HT, Parham P, Weiss EH, Girdlestone J (1992) *J. Immunol.* **148** 4043-4053

Messing J, Vieira J (1982) *Gene* **19** 269-276

Miller WL (1988) *Endocr. Rev.* **9** 295-318

Milner CM, Campbell RD (1990) *Immunogenet.* **32** 242-251

Milner CM, Campbell RD (1992a) *Immunogenet.* **36** 357-362

Milner CM, Campbell RD (1992b) *BioEssays* **14** 565-571

Milner CM, Campbell RD (1993) *Biochem. J.* In press

Minota S, Cameron B, Welch WJ, Winfield JB (1988) *J. Exp. Med.* **168** 1475-1480

Mizzen LA, Kabiling AN, Welch WJ (1991) *Cell Regulation* **2** 165-179

Modrich P (1991) *Ann. Rev. Genet.* **25** 229-253

Momburg F, Ortiz-Navarrette V, Neefjes J, Goulmy E, van de Wal Y, Spits H, Powis SJ, Butcher GW, Howard JC, Walden P, Hammerling GJ (1992) *Nature* **360** 174-177

Monaco AP, Neve RL, Colletti-Feener C, Bertelson CJ, Kurnit DM, Kunkel LM (1986) *Nature* **323** 646-650

Monaco JJ, Cho S, Attaya M (1990) *Science* **250** 1723-1726

Morel Y, Bristow J, Gitelman SE, Miller WL (1989) *Proc. Natl. Acad. Sci. USA* **86** 6582-6586

Moss SE, Crompton MR, Crumpton MJ (1988) *Eur. J. Biochem.* **177** 21-27

Mount SM (1982) *Nucleic Acids Res.* **10** 459-472

Muller U, Stephan D, Philipsen P, Steinmetz M (1987a) *EMBO J.* **6** 369-373

Muller U, Jongeneel CV, Nedospasov SA, Fischer Lindahl K, Steinmetz M (1987b) *Nature* **325** 265-267

Munro S, Pelham HRB (1986) *Cell* **46** 291-300

Murray V (1989) *Nucleic Acids Res.* **17** 8889

Nadler SG, Tepper MA, Schacter B, Mazzucco CE (1992) *Science* **258** 484-486

Nepom GT, Erlich H (1991) *Ann. Rev. Immunol.* **9** 493-525

Nerbert DN, Gonzalez FJ (1987) *Ann. Rev. Biochem.* **56** 945-993

New MI, Dupont B, Grumback K, Levine LS (1982) In *Metabolic Basis of Inherited Disease* McGraw-Hill, New York 973-1000

Niederhuber JE, Frelinger JA (1976) *Transplant. Rev.* **30** 101-121

Olaisen B, Sakaguchi AY, Naylor SL (1987) *Cytogenet. Cell Genet.* **46** 147-169

Olavesen MG, Thomson, Cheng J, Campbell RD (1993) In *Proceedings of the 11th International Histocompatibility Workshop* Oxford University Press, Oxford

Oldstone MBA (1987) *Cell* **50** 819-820

Oohira T, Nagata N, Akoboshi I, Matsuda I, Naito S (1985) *Hum. Genet.* **70** 341-343

Ortiz-Navarrete V, Seelig A, Gernold M, Frentzel S, Kloetzel PM, Hammerling GJ (1991) *Nature* **353** 662-664

Pabo CO, Sauer RT (1984) *Ann. Rev. Biochem.* **53** 293-321

Palleros DR, Welch WJ, Fink AL (1991) *Proc. Natl. Acad. Sci. USA* **88** 5719-5723

Parham P, Lomen CE, Lawlor DA, Ways JP, Holmes N, Coppin HL, Salter RD, Wan AM, Ennis PD (1988) *Proc. Natl. Acad. Sci. USA* **85** 4005-4009

Parimoo S, Patanjali SR, Shukla H, Chaplin DD, Weissman SM (1991) *Proc. Natl. Acad. Sci. USA* **88** 9623-9627

Partanen J, Campbell RD (1993) In *Proceedings of the 11th International Histocompatibility Workshop* Oxford University Press, Oxford

Patanjali SR, Parimoo S, Weissman SM (1991) *Proc. Natl. Acad. Sci. USA* **88** 1943-1947

Pearson WR, Lipman DJ (1988) *Proc. Natl. Acad. Sci. USA* **85** 2444-2448

Pelham HRB (1990) In *Stress Proteins in Biology and Medicine*. Morimoto RI, Tissieres A, Georgopoulos C (Eds) Cold Spring Harbour Laboratory Press New York 287-299

Pepinsky RB, Tizard R, MattalianoRJ, Sinclair LK, Miller GT, Browning JL, Chow EP, Burne C, Huang KS, Pratt D, Wachter L, Hession C, Frey AZ, Wallner BP (1988) *J. Biol. Chem*. **263** 10799-10811

Person PL, Snoek M, Demant P, Teuscher C (1993) *Reg. Immunol.* In press

Pfaeffle M, Ruggerio F, Hofmann H, Fernandez MP, Selmin O, Yamada Y, Garrone R, von der Mark K (1990) EMBO J. 9 1336-1336

Pierce SK, DeNagel DC, VanBuskirk AM (1991) *Current Topics Microbiol. Immunol.* **167** 83-92

Powis SJ, Townsend ARM, Deverson EV, Bastin J, Butcher GW, Howard JC (1991) *Nature* **354** 528-531

Powis SH, Mockridge I, Kelly A, Kerr L-A, Glynne R, Gileadi U, Beck S, Trowsdale J (1992a) *Proc. Natl. Acad. Sci. USA* **89** 1463-1467

Powis SJ, Deverson EV, Coadwell WJ, Ciruela A, Huskisson NS, Smith H, Butcher G W, Howard JC (1992b) *Nature* **357** 211-215

Priebe SD, Hadi SM, Greenberg B,Lacks SA (1988) *J. Bacteriol.* **170** 190-196

Proudfoot NJ, Brownlee GG (1976) *Nature* **263** 211-214

Prudhomme M, Mejean V, Martin B, Claverys J-P (1991) *J. Bacteriol.* **173** 7196-7203

Ragoussis J, Bloemer K, Pohla H, Messer G, Weiss EH, Ziegler A (1989) *Genomics* **4** 301-308

Ragoussis J, Monaco A, Mockridge I, Kendall E, Campbell RD, Trowsdale J (1991) *Proc. Natl. Acad. Sci. USA* **88** 3753-3757

Raleigh E, Wilson G, Latimer E, Card C (1988) *Nucleic Acids Res*. **16** 1563-1575

Ratanachaiyavong S, Campbell RD, McGregor AM (1989) *Hum. Immunol*. **26** 310-320

Rembecki RM, Kumar V, David CS, Bennett M (1988) *J. Immunol*. **141** 2253-2260

Rippmann F, Taylor WR, Rothbard JB, Green NM (1991) *EMBO J*. **10** 1053-1059

Rodrigues NR, Dunham I, Yu CY, Carroll MC, Porter RR, Campbell RD (1987) *EMBO J*. **6** 1653-1661

Rokeach LA, Haselby JA, Hoch SO (1988) *Proc. Natl. Acad. Sci. USA* **85** 4832-4836

Rommens JM, Iannuzzi MC, Kerem B-S, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N, Zsiga M, Buchwald M, Riordan JR, Tsui L-C, Collins FS (1989) *Science* **245** 1059-1065

Rudensky AY, Preston-Hurlburt P, Hong S-C, Barlow A, Janeway CA (1991) *Nature* **353** 622-627

Sadegh-Nasseri S, Germain RN (1992) *Immunol. Today* **13** 43-46

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA (1988) *Science* **239** 487-491

Salter RD, Benjamin RJ, Wesley PK, Buxton SE, Garrett TPJ, Clayberger C, Krensky AM, Norment AM, Littman DR, Parham P (1990) *Nature* **345** 41-46

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbour Laboratory Press New York

Sanger F, Coulson AR, Barrell BG, Smith AJH, Roe BA (1980) *J. Mol. Biol.* **143** 161-178

Saper MA, Bjorkman PJ, Wiley DC (1991) *J. Mol. Biol.* **219** 277-319

Saraste M, Sibbald PR, Wittinghofer A (1990) *Trends Biochem. Sci.* **15** 430-434

Sargent CA (1988) *D. Phil. Thesis*, Oxford

Sargent CA, Dunham I, Campbell RD (1989a) *EMBO J.* **8** 2305-2312

Sargent CA, Dunham I, Trowsdale J, Campbell RD (1989b) *Proc. Natl. Acad. Sci. USA* **86** 1968-1972

Schaffer FM, Palermos J, Zhu Z-B, Barger BO, Cooper MD, Volanakis JE (1989) *Proc. Natl. Acad. Sci. USA* **86** 8015-8019

Schmid CW, Jelinek WR (1982) *Science* **216** 1065-1070

Sealey PG, Whittaker PA, Southern EM (1985) *Nucleic Acids Res.* **13** 1905

Seed B (1987) *Nature* **329** 840-842

Shanahan F, Weinstein WM (1988) *New Eng. J. Med.* **319** 782-783

Shaw S, Duquesnoy RJ, Smith PL (1981) *Immunogenet.* **14** 153-162

Shirakawa H, Tsuda K, Yoshida M (1990) *Biochem.* **29** 4419-4423

Schlensog V, Block A (1991) *J. Bacteriol.* **173** 7414-7415

Short JM, Fernandez JM, Sorge JA, Huse WD (1988) *Nucleic Acids Res.* **16** 7583-7600

Shreffler DC (1976) *Transplant. Rev.* **32** 140-167

Shukla H, Gillespie GA, Srivastava R, Collins F, Chorney MJ (1991) *Genomics* **10** 905-914

Silver ML, Guo H-C, Strominger JL, Wiley DC, (1992) *Nature* **360** 367-369

Simmons D, Seed B (1988) *J. Immunol.* **141** 2797-2800

Siri A, Carnemolla B, Saginati M, Leprini A, Casari G, Baralle F, Zardi L (1991) *Nucleic Acids Res.* **19** 525-531

Smith DB, Johnson KS (1988) *Gene* **67** 31-44

Snoek M, Groot PC, Spies T, Campbell RD, Demant P (1991) *Immunogenet.* **34** 409-412

Snoek M, Jansen M, Olavesen MG, Campbell RD, Teuscher C, van Vugt H (1993) *Genomics* In press

Southern EM (1975) *J. Mol. Biol.* **98** 503-517

Speiser PW, White PC (1989) *DNA* **8** 745-751

Spence MA, Spurr NK, Field LL (1989) *Cytogenet. Cell Genet.* **51** 149-165

Spies T, Bresnahan M, Strominger JL (1989a) *Proc. Natl. Acad. Sci. USA* **86** 8955-8958

Spies T, Blanck G, Bresnahan M, Sands J, Strominger JL (1989b) *Science* **243** 214-217

Spies T, Bresnahan M, Bahram S, Arnold D, Blanck G, Mellins E, Pious D, DeMars R (1990) *Nature* **348** 744-747

Spies T, DeMars R (1991) *Nature* **351** 323-324

Spies T, Cerundolo V, Colonna M, Cresswell P, Townsend ARM, DeMars R (1992) *Nature* **355** 644-646

Srivastava R, Duceman BW, Biro PA, Ashwani KS, Weissman SM (1985) *Immunol. Rev.* **84** 93-119

Srivastava M, McBride OW, Fleming PJ, Pollard HB, Burns AL (1990) *J. Biol. Chem.* **265** 14922-14931

Su S-S, Modrich P (1986) *Proc. Natl. Acad. Sci. USA* **83** 5057-5061

Tabor S, Richardson CC (1987) *Proc. Natl. Acad. Sci. USA* **84** 4767-4771

Tait JF, Sakata M, McMullen BA, Miao CH, Funakoshi T, Hendrickson LE, Fujikawa K (1988) *Biochem.* **27** 6268-6276

Teuscher C, Smith SM, Goldberg EH, Shearer GM, Tung KSK (1985) *Immunogenet.* **22** 323-333

Teuscher C, Gasser DL, Woodward SR, Hickey WF (1990) *Immunogenet.* **32** 337-344

Teyton L, Peterson PA (1992) *Trends Cell Biol.* **2** 52-56

Thomson G (1988) *Ann. Rev. Genet.* **22** 31-50

Ting J, Lee AS (1988) *DNA* **7** 275-286

Tiwari JL, Terasaki PI (1985) *HLA & Disease Associations.* Springer-Verlag, New York

Todd JA (1990) *Immunol. Today* **11** 122-129

Todd JA, Bell JI, McDevitt HO (1987) *Nature* **329** 599-604

Townsend ARM, Bodmer H (1989) *Ann. Rev. Immunol.* **7** 601-624

Townsend ARM, Ohlen C, Bastin J, Ljunggren H-G, Foster L, Karre K (1989) *Nature* **340** 443-448

Trowsdale J, Young JAT, Kelly AP, Austin PJ, Carson S, Meunier H, So A, Erlich HA, Spielman RS, Bodmer J, Bodmer WF (1985) *Immunol. Rev.* **85** 5-43

Trowsdale J, Hanson I, Mockridge I, Beck S, Townsened A, Kelly A (1990) *Nature* **348** 741-744

Trowsdale J, Ragoussis J, Campbell RD (1991) *Immunol. Today* **12** 443-446

Tsoulfa G, Rook GA, van Embden JD, Young DB, Mehlert A, Isenberg DA, Hay FC, Lydyard PM (1989) *Ann. Rheum. Dis.* **48** 118-123

Tsuge I, Shen F-W, Steinmetz M, Boyse EA (1987) *Immunogenet.* **26** 378-386

Tung KSK (1978) *Science* **201** 833-835

Tung KSK, Alexander NJ (1980) *Am. J. Path.* **101** 17-30

Tung KSK, Ellis L, Teuscher C, Meng A, Blaustein JC, Kohno S, Howell R (1981) *J. Exp. Med.* **154** 1016-1032

Valle G, Bergantino E, Lanfranchi G, Carignani G (1991) *Yeast* **7** 981-988

VanBuskirk A, Crump BL, Margoliash E, Pierce SK (1989) *J. Exp. Med.* **170** 1779-1089

VanBuskirk AM, DeNagel DC, Guagliardi LE, Brodsky FM, Pierce SK (1991) *J. Immunol.* **146** 500-506

Venditti CP, Chorney MJ (1992) *Genomics* **14** 1003-1009

Vollrath D, Davis RW, Connelly C, Hieter P (1988) *Proc. Natl. Acad. Sci. USA* **85** 6027-6031

Volonakis JE, Zhu Z-B, Schaffer FM, Macon KJ, Palermos J, Barger BO, Go R, Campbell RD, Schroeder HW, Cooper MD (1992) *J. Clin. Invest.* **89** 1914-1922

von Heijne G (1988) *Biochim. Biophys. Acta* **947** 307-333

Wallace MR, Marchuk DA, Anderson LB, Letcher R, Odeh HM, Saulino AM, Fountain JW, Brereton A, Nicholson J, Mitchell AL, Brownstein BH, Collins FS (1990) *Science* **249** 181-186

Weber K, Johnsson N, Plessmann U, Van PN, Soling H-D, Ampe C, Vandekerckhove J (1987) *EMBO J.* **6** 1599-1604

Weber JL, May PE (1989) *Am. J. Hum. Genet.* **44** 388-396

Weber JL (1990) *Genomics* **7** 524-530

Weiss EH, Golden L, Fahrner K, Mellor AL, Devlin JJ, Bullman H, Tiddens H, Bud H, Flavell RA (1984) *Nature* **310** 650-655

Welch WJ, Mizzen LA (1988) *J. Cell. Biol.* **106** 1117-1130

Welch WJ (1990) In *Stress Proteins in Biology and Medicine.* Morimoto RI, Tissieres A, Georgopoulos C (Eds) Cold Spring Harbour Laboratory Press New York 223-278

Wen L, Huang JK, Johnson BH, Reeck GR (1989) *Nucleic Acids Res.* **17** 1197-1214

White PC, Grossberger D, Onufer BJ, Chaplin DD, New MI, Dupont B, Strominger JL (1985) *Proc. Natl. Acad. Sci. USA* **82** 1089-1093

Williams AF, Barclay AN (1988) *Ann. Rev. Immunol.* **6** 381-405

Wordsworth BP, Lanchbury JSS, Sakkas LI, Welsh KI, Panayi GS, Bell JI (1989) *Proc. Natl. Acad. Sci. USA* **86** 10049-10053

World Health Organisation (1968) *WHO Bull.* **39** 935-938

World Health Organisation (1981) *WHO Bull.* **59** 489-491

WHO Nomenclature Committee (1992) *Immunogenet.* **36** 135-148

Wroblewski JM, Kaminsky SG, Milisauskas VK, Pittman AM, Chaplin DD, Spies T, Nakamura I (1990) *Immunogenet.* **32** 200-204

Wu LC, Morley BJ, Campbell RD (1987b) *Cell* **48** 331-342

Yang Y, Fruh K, Chambers J, Waters JB, Wu L, Spies T, Peterson PA (1992) *J. Biol. Chem.* **267** 11669-11672

Young DB, Mehlert A, Smith DF (1990) in; *Stress Proteins in Biology and Medicine.* Morimoto RI, Tissieres A, Georgopoulos C (Eds) Cold Spring Harbour Laboratory Press New York 131-165

Yu CY, Belt KT, Giles CM, Campbell RD, Porter RR (1986) *EMBO J.* **5** 2873-2881

Zakeri ZF, Wolgemuth DJ (1987) *Mol. Cell. Biol.* **7** 1791-1796

Zakeri ZF, Wolgemuth DJ, Hunt CR (1988) *Mol. Cell. Biol.* **8** 2925-2932

Zhang W, Young ACM, Imarai M, Nathenson SG, Sacchettini JC (1992) *Proc. Natl. Acad. Sci. USA* **89** 8403-8407

Zhu Z-B, Hsieh SL, Bentley DR, Campbell RD, Volonakis JE (1992) *J. Exp. Med.* **175** 1783-1787

Zinkernagel RM, Doherty PC (1979) *Adv. Immunol.* **27** 51-177