Characterisation of the genomic region around the TNF locus within the Human Major Histocompatibility Complex in the Chromosome band 6p21.3

Matt J. Neville
Magdalen College, Oxford

A Thesis submitted for the
degree of Doctor of Philosophy
Hilary 2000

MRC Immunochemistry Unit
Department of Biochemistry
University of Oxford



ABSTRACT

Characterisation of the genomic region around the TNF locus within the Human Major Histocompatibility Complex in the Chromosome band 6p21.3

Matthew J. Neville, Magdalen College, University of Oxford

Thesis submitted for the degree of D.Phil.
Hilary Term 2000

It is becoming increasingly apparent that many of the genes in the class III region of the human Major Histocompatibility Complex encode proteins involved in the immune and inflammatory responses. Furthermore, genetic studies have indicated that genes within the class III region, particularly the telomeric segment containing the TNF gene, could contribute to susceptibility to diseases of immune-related aetiology. To further characterise this region and to identify candidate disease susceptibility genes, two overlapping cosmids, TN62 and TN82, covering an ~82kb segment of DNA around the TNF gene were selected for sequence analysis.

The eight known genes in this region have been precisely positioned with the order: G1/AIF-1, 1C7, LST1 (B144), LTB, TNF, LTA, IKBL, BAT1 (centromere to telomere) and their genomic structures have been defined. Comparison of the G1 genomic region with previously described cDNA and genomic sequences, together with the results of RT-PCR, indicates that three alternative transcripts, G1, Allograft Inflammatory Factor-1 and Interferon-γ Responsive Transcript-1, are all derived from this gene. The completion of the sequence of 1C7 (D6S2570) has revealed that this gene encodes a putative novel member of the immunoglobulin superfamily. A number of alternatively spliced transcripts of 1C7 were identified by RT-PCR, all of which are expressed in immune-related cell lines. Alternative splicing within the immunoglobulin domain-encoding region was seen to result in possible set switching between an IgV domain and an IgC2 domain. In addition to this, a previously unidentified gene, homologous to a number of V-ATPase G-subunits, has been located 1kb telomeric of IKBL. Lastly, the pseudogene UCRH-L and an AIF-1 gene fragment have been identified in the intergenic region between AIF-1 and 1C7.

In order to assess the contribution of loci in this region to disease susceptibility, the genes AIF-1, 1C7, ATP6G and the BAT1 promoter region were subjected to mutation analysis. A total of 28 polymorphisms have been identified, 8 in AIF-1, 10 in 1C7, 7 in ATP6G and 3 in the BAT1 promoter region. Work is at present underway to genotype a number of the identified polymorphisms in control DNAs and in DNA samples from patients with combined variable immunodeficiency (CVID). The information generated from this analysis will bring us closer to explaining the reported linkage of CVID with the telomeric end of the human MHC class III region.

ACKNOWLEDGEMENTS

I owe an enormous debt of gratitude to many people, without whose support and efforts this work would not have been possible, in particular:

- My supervisors Dr Duncan Campbell and Dr Caroline Milner for this opportunity, for guidance over the last four years and for thorough correction of this thesis.

- All members of the F1 lab, past and present, for their advice, encouragement and provision of reagents for this project. In particular Dr Begonã Aguado for reagents and invaluable advice throughout; Dr John Broxholme for help with radioactive work and bioinformatics; Dr Suzanne Jenkins and Dr Ruth March for help with PCR work, mutation analysis and general support and gossip; Dr Gloria Ribas-Despuig for help and support with the sequencing project and mutation detection; and Helene Jones for help with setting up the sequencing work.

- Prof. Ken Reid and members of the MRC Immunochemistry Unit for help, advice and intermittent trips to local ale houses.

- The staff of the Biochemistry Department.

- Dr Ken Welsh, Dr Sara Marshall, Dr Mike Bunce and all other members of the Nuffield Department of Surgery Tissue Typing Laboratory for the use of their DNA samples and all their help and advice whilst setting up the PCR-SSP assays.

- I gratefully acknowledge the receipt of an Medical Research Council (MRC) studentship.

- Finally, I would like to say a much deserved 'Thank You' to big and little Lou and all my friends for putting up with me during my writing up.

ABBREVIATIONS

Nucleotides containing the bases adenine, cytosine, guanine and thymine are abbreviated to A, C, G and T, respectively.

AH	Ancestral haplotype
AIF-1	Allograft Inflammatory Factor-1
APS	Ammonium persulphate
AS	Ankylosing spondylitis
Asn	Asparagine
ATP6G	Vacuolar-ATPase G-subunit
ATP	Adenosine triphosphate
β2-m	β2-microtubulin
BART-1	Balloon angioplasty responsive transcript-1
Bf	Factor B
bp	Base pair
CAH	Congenital adrenal hyperplasia
CLIP	Class II-associated invariant chain peptide
CVID	Combined variable immunodeficiency
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	Deoxyribonucleotide triphosphate
ddNTP	Dideoxyribonucleotide triphosphate
DTT	Dithiothreitol
EBI	European Bioinformatics Institute
EDTA	Ethylenediaminetetra-acetic acid
ER	Endoplasmic reticulum
EST	Expressed sequence tag
FAM	Free arm monomer
FGFR	Fibroblast growth factor receptor
FISH	Fluorescent *in-situ* hybridisation
FLAM	Free left arm monomer
FRAM	Free right arm monomer
HCG	Haemochromatosis candidate gene
HGMP	Human Genome Mapping Project
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HPLC	High performance liquid chromatography
IDDM	Insulin-dependent diabetes mellitus
IFN-γ	Interferon gamma
Ig	Immunoglobulin
IgSF	Immunoglobulin superfamily
Ii	Invariant chain
IKBL	I kappa B like protein
IPTG	Isopropyl-β-D-thiogalactopyranoside
IRT-1	Interferon gamma responsive transcript
kb	Kilobase pair
LINE	Long interspersed nuclear element
LST-1	Leukocyte specific transcript-1
LTA/B	Lymphotoxin A/B
LTR	Long terminal repeat

MaLR	Mammalian LTR-retroposon
MER	Medium reiterary repeat
MG	Myasthenia gravis
MIR	Mammalian-wide interspersed repeats
MHC	Major histocompatibility complex
mRNA	Messenger RNA
NCBI	National Centre for Biotechnology Institute
ng	Nanogram
nt	Nucleotide
OD_{xxx}	Optical density at xxx nm
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PBL	Peripheral blood lymphocyte
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PCR-SSP	Sequence specific primer polymerase chain reaction
PEG	Polyethylene glycol
PolyA	polyadenylation signal
PPT2	Palmitoyl-protein thioesterase-2
RA	Rheumatoid arthritis
RACE	Rapid amplification of cDNA ends
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RT-PCR	Reverse transcription PCR
SINE	Short interspersed nuclear element
SLE	Systemic lupus erythematosus
SNP	Single nucleotide polymorphism
STS	Sequence tagged site
TAP	Transporter associated with antigen processing
TCR	T cell receptor
TEMED	N, N, N', N' - tetramethylethylenediamine
TIGR	The Institute of Genomic Research
TNF	Tumour necrosis factor
Tris	Tris(hydroxymethyl)-aminoethane
U	Units
UCRHL	Ubiquinol-cytochrome-c reductase hinge protein like
UTR	Untranslated region
VSMC	Vascular smooth muscle cells
YAC	Yeast artificial chromosome

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

1.1 THE MAJOR HISTOCOMPATIBILITY COMPLEX. 1

1.1.1 The MHC class I and II regions. 2

1.1.1.1 The MHC class I region 2

1.1.1.2 The MHC class II region 7

1.1.2 The MHC class III region 11

1.1.2.1 The Complement genes 12

1.1.2.2 The heat shock protein genes 13

1.1.2.3 The Leukocyte Antigen 6 Superfamily gene cluster 13

1.1.2.4 RAGE 14

1.1.2.5 hLPAATα (G15) 14

1.1.2.6 NEU (G9) 15

1.1.2.7 SK12W (SKIV2L/ G11a) 15

1.1.2.8 The G1 – BAT1 interval, a cluster of genes involved in the inflammatory response 16

1.1.2.9 MHC class III region genes with no obvious immune related function. 18

1.1.3 Evolution of the MHC region 20

1.1.4 The MHC and disease 22

1.1.4.1 Disease association within the MHC class III region 24

1.1.4.2 The significance of the G1 to BAT1 interval. 25

1.2 PROJECT RATIONALE AND BACKGROUND 27

1.2.1 Molecular mapping of the MHC class III region 27

1.2.2 Why sequence the MHC class III region ? 28

1.3 AIMS 31

CHAPTER 2: MATERIALS AND METHODS 32

2.1 GENERAL SOLUTIONS, BUFFERS AND MEDIA 32

2.2 BACTERIAL CELL CULTURE 36

2.2.1 Culturing Bacterial Cells 36

2.2.2 Glycerol Stocks (storage of bacterial cultures) 36

2.2.3 Preparation of Competent Bacterial Cells 36

2.2.4 Transformation Procedures 37

2.2.4.1 Chemical transformation of plasmid DNA 37
2.2.4.2 Chemical transformation of M13mp18 37
2.3 NUCLEIC ACID TECHNIQUES 37
2.3.1 Organic Extraction 37
2.3.2 Recovery of DNA from Aqueous Solution 38
2.3.3 Restriction Enzyme Digestion 38
2.3.4 Agarose Gel Electrophoresis 38
2.3.4.1 DNA agarose gel electrophoresis 38
2.3.4.2 Preparative DNA agarose gels 39
2.3.4.3 RNA agarose gel electrophoresis 39
2.3.5 Recovery of DNA from Agarose Gels (Gel Purification) 39
2.3.6 Determination of DNA Concentration 40
2.3.7 Preparation of Plasmid or Cosmid DNA 40
2.3.7.1 Alkaline lysis method 40
2.3.7.2 Wizard miniprep DNA purification (Plasmid only) 41
2.3.8 DNA Cloning Techniques 41
2.3.8.1 Cloning into Plasmid vectors 41
2.3.8.1.1 Preparation of plasmid vectors 41
2.3.8.1.2 Dephosphorylation 41
2.3.8.1.3 Ligation 42
2.3.8.1.4 Transformation 42
2.3.8.1.5 Identification of recombinant clones 42
2.3.8.1.6 pGEM -T Easy vector system (Promega) for the cloning of PCR products 42
2.3.8.2 Cloning into the M13mp18 bacteriophage vector 43
2.3.8.2.1 Preparation of M13mp18 vectors 43
2.3.8.2.2 End Filling 43
2.3.8.2.3 Ligation 44
2.3.8.2.4 Transformation 44
2.3.8.2.5 Identification of recombinant clones 44
2.3.8.2.6 Preparation of single stranded DNA from M13mp18 44
2.3.9 The Polymerase Chain Reaction (PCR) 45
2.3.9.1 Oligonucleotide primer design 45
2.3.9.2 Preparation of oligonucleotide primers 45

2.3.9.3 General amplification conditions for cosmid and plasmid DNA 46
2.3.9.4 Modifications to the general PCR cycling parameters 47
2.3.9.4.1 Hot start PCR 47
2.3.9.4.2 Touchdown PCR 47
2.3.9.5 Amplification of genomic DNA for direct sequencing 47
2.3.9.6 Amplification of plasmid DNA from E. coli cells (Colony PCR) 48
2.3.10 The Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) 49
2.3.10.1 First strand cDNA synthesis 49
2.3.10.2 Template specific PCR 49
2.3.10.3 Confirmation of identity of PCR product 50
2.3.11 Hybridisation Techniques 50
2.3.11.1 Transfer of DNA from agarose gels onto nitro-cellulose membranes 50
2.3.11.2 Transfer of RNA from agarose gels onto nitro-cellulose 51
2.3.11.3 Transfer of bacterial DNA onto nitro-cellulose membranes 51
2.3.11.4 Radiolabelling of DNA 52
2.3.11.5 Hybridisation of Radiolabelled DNA Probes 52
2.3.11.6 Washing of membranes and autoradiography 53
2.3.12 DNA Sequencing 53
2.3.12.1 Dye-primer sequencing used for M13 clones and M13-tailed PCR sequencing 53
2.3.12.2 Dye-terminator sequencing used for M13 clones, plasmid clones, and the direct sequencing of PCR products 54
2.3.13 Cosmid Genomic Insert Sequencing Strategy 55
2.3.13.1 Preparation of an M13mp18 Shotgun Library 55
2.3.13.2 Automated Sequencing 55
2.3.13.3 Finishing and Troubleshooting 56
2.3.13.3.1 Sequence ambiguities 56
2.3.13.3.2 Areas of single orientation read 57
2.3.13.3.3 Gaps between contigs 57
2.3.14 Genomic Sequencing Approach To Mutation Analysis 58
2.3.15 PCR-SSP genotyping of polymorphisms 59
2.3.15.1 Primer design and preparation 59
2.3.15.1.1 Primer design 59
2.3.15.1.2 Primer preparation 60

2.3.15. 2 PCR-SSP set up. 61
2.3.15.3 PCR machines and thermal cycling parameters. 62
2.3.15.4 Electrophoresis and photography. 63
2.4 BIOINFORMATICS 63

CHAPTER 3: THE SEQUENCING AND INITIAL ANALYSIS OF 82 KILOBASES OF DNA AROUND THE TNF LOCUS WITHIN THE MHC CLASS III REGION 65

3.1 INTRODUCTION. 65
3.1.1 Methodology 65
3.1.1.1 M13 library preparation and sequencing 66
3.1.1.2 Sequence assembly 68
3.1.1.3 Sequence analysis 69
3.2 RESULTS 70
3.2.1 Sequencing and Sequence Assembly 70
3.2.1.1 Finishing 71
3.2.2 Sequence Analysis 72
3.2.2.1 Sequence composition and the positions of previously identified genes ... 72
3.2.2.2 Comparison with other human genomic sequences within the databases .. 72
3.2.2.3 Analysis of known genes 74
3.2.2.3.1 G1 74
3.2.2.3.2 1C7 75
3.2.2.3.3 LST-1(B144) 75
3.2.2.3.4 The LTB/TNF/LTA gene cluster 78
3.2.2.3.5 IKBL 79
3.2.2.3.6 BAT1 79
3.2.2.4 Identification of new coding regions 81
3.2.2.4.1 The AIF-1 to 1C7 genomic region 82
3.2.2.4.2 NB6 genomic region between LTA and IKBL 83
3.2.2.4.3 The IKBL – BAT1 intergenic region 84
3.2.2.5 Repeat sequences 84
3.2.2.6 Alignment of the Human and Mouse genomic sequences 85
3.2.2.6.1 Alignment of the human and murine genomic sequences to aid the identification of potential coding regions in the human sequence. 86

3.3 DISCUSSION 87
3.3.1 Sequencing strategy 87
3.3.2 Genomic structure, other genomic database sequences, and repeat sequences . 89
3.3.2.1 Genomic structure 89
3.3.2.1.1 The putative pseudogene UCRHL and a fragment of the AIF-1 gene 89
3.3.2.2 Variation between the different genomic sequences within the databases . 90
3.3.2.3 Conservation between Human and Mouse 91
3.3.2.4 Repeat sequences 92
3.3.3 Alternative splicing of the Leukocyte specific transcript-1(LST-1) gene 95
3.3.4 A novel splice variant of the BAT1 gene 104
3.3.5 The hunt for potential new coding regions 107
3.3.6 Tools and databases: Gene hunting strategies summarised. 108
3.3.6.1 The use of EST databases. 108
3.3.6.2 Exon prediction analysis. 109

CHAPTER 4: ANALYSIS AND CHARACTERISATION OF THE 1C7, AIF-1 AND ATP6G GENOMIC REGIONS. 111

4.1 INTRODUCTION 111
4.2 RESULTS 112
4.2.1 The G1/AIF-1 genomic region 112
4.2.1.1 Alignments with human database entries 112
4.2.1.2 Alignment of database entries from non-human organisms 113
4.2.1.3 Analysis of the encoded amino acid sequences of the AIF-1 gene 115
4.2.1.4 Expression profiling of the AIF-1 gene encoded transcripts by RT-PCR 115
4.2.2 The 1C7 gene 116
4.2.2.1 Database interrogation, DNA sequence analysis and cDNA clone characterisation 116
4.2.2.2 Expression profiling of the 1C7 gene transcripts and the identification of further splice variants by RT-PCR 117
4.2.2.3 Amino acid sequence analysis of the 1C7 gene encoded transcripts 119
4.2.3 An ATPase G subunit Homolog 120
4.2.3.1 Database interrogation, DNA sequence analysis and cDNA clone characterisation 120

4.2.3.2 Expression profiling of the ATP6G gene encoded transcripts by RT-PCR 122
4.2.3.3 Amino acid sequence analysis of the ATP6G gene encoded transcripts.. 123
4.3 DISCUSSION 123
4.3.1 AIF-1, G1 and IRT-1 123
4.3.2 1C7, A novel member of the immunoglobulin super-family 130
4.3.2.1 Speculations on the evolution of the IgSF from the 1C7 Ig domain splicing 135
4.3.3 A Putative Human Homolog of the Tobacco Hornworm V-ATPase G Subunit 138

CHAPTER 5: INVESTIGATION OF POLYMORPHISMS WITHIN THE AIF-1, 1C7 AND ATP6G GENOMIC REGIONS AND WITHIN THE BAT1 PROMOTER REGION. 142
5.1 INTRODUCTION 142
5.2 MUTATION DETECTION WITHIN THE AIF-1, 1C7 AND ATP6G GENOMIC REGIONS AND CHARACTERISATION OF THE BAT1 5' DELETION. 144
5.2.1 Methodology 144
5.2.2 Results 146
5.2.2.1 Polymorphisms within the AIF-1 genomic region 147
5.2.2.2 Polymorphisms within the 1C7 genomic region 148
5.2.2.3 Polymorphisms within the ATP6G genomic region 149
5.2.2.4 Polymorphisms within the BAT1 promoter region 149
5.3 THE SCREENING OF POLYMORPHISMS IDENTIFIED IN THIS STUDY FOR ASSOCIATIONS WITH COMMON VARIABLE IMMUNODEFICIENCY AND IgA DEFICIENCY. 150
5.3.1 Resources and Methodology 152
5.3.1.1 DNA resources 152
5.3.1.2 Polymorphisms screened 152
5.3.1.3 Methodology 153
5.3.2 Results 154
5.4 DISCUSSION 155
5.4.1 The significance of sequence variations in 1C7, AIF-1 and BAT1 and the ongoing genotyping effort 155

5.4.2 Variation within the AIF-1, 1C7, ATP6G and BAT1 genes 156
5.4.2.1 The AIF-1 Gene..... 156
5.4.2.2 The 1C7 gene 157
5.4.2.3 The ATP6G gene..... 159
5.4.2.3 The BAT1 promoter region..... 160
5.4.3 DNA variation and the Human Genome Project 161

CHAPTER 6: CONCLUSIONS 163
6.1 THE COMPLEXITY OF TRANSCRIBED REGIONS IDENTIFIED..... 163
6.2 THE FUNCTION OF ENCODED GENE PRODUCTS 164
6.3 FUTURE WORK..... 165
6.4 FINAL COMMENTS 166

APPENDICIES
APPENDIX I..... 167
APPENDIX II..... 186
APPENDIX IIIA..... 190
APPENDIX IIIB..... 191

REFERENCES 192

TABLE OF FIGURES

FIGURE 1.1 Gene map of the Human Major Histocompatibility complex 1
FIGURE 1.2 Map of the human MHC class III region 11

FIGURE 3.1 The telomeric half of the MHC class III region 65
FIGURE 3.2 Lengths of high quality and vector clipped sequence generated from 70
FIGURE 3.3 The relative positions of genomic sequences found within the EMBL database that overlap with the G1-BAT1 genomic region 73
FIGURE 3.4 The genomic structure and splice variants of LST-1 75
FIGURE 3.5 the LST-1 genomic region from the genomic sequence generated in this study, co-ordinates 33450-30601 (reverse orientation) 75
FIGURE 3.6 Polypeptides encoded by open reading frames (ORFs) within the different splice variants of the LST-1 gene 77
FIGURE 3.7 The BAT1 genomic region contained within the TN62/TN82 genomic sequence 80
FIGURE 3.8 A map of transcripts and other features within the human genomic region covered by overlapping cosmids TN62 and TN82 84
FIGURE 3.9 Comparison of the Human TN62/TN82 genomic region with the equivalent Murine genomic region 85
FIGURE 3.10 Multiple alignment of the B144 gene expressed sequences held within the databases 98
FIGURE 3.11 Multiple alignment of the amino acid sequence of the two identified variants of the murine B144 with the first open reading frame (ORF1) of human LST1-5 splice variant, as presented in Figure 3.6 99
FIGURE 3.12 The NIX graphical output for the TN62/TN82 genomic sequence 110

FIGURE 4.1 The AIF-1 genomic region corresponding to co-ordinates 4500-2501 (reverse orientation) of the TN62/TN82 genomic sequence 113
FIGURE 4.2 AIF-1 splice variants 113
FIGURE 4.3 A multiple alignment of the AIF-1 amino acid sequences from different species and alternatively spliced forms found in man 114
FIGURE 4.4 A multiple alignment of the AIF-1 amino acid sequence, as in Figure 4.3, with addition of the BART-1 deduced amino acid sequence from Rat 114

FIGURE 4.5 The 1C7 genomic region corresponding to co-ordinates 26401-30700 of the TN62/TN82 genomic sequence 117
FIGURE 4.6 1C7 splice variants detected by RT-PCR and EST database searches 117
FIGURE 4.7 The amino acid sequence and putative structure of 1C7 119
FIGURE 4.8 Structure based amino acid sequence alignments of the putative immunoglobulin domains of 1C7 119
FIGURE 4.9 The ATP6G genomic region from the genomic sequence generated in this study (co-ordinates 72500-75399) 120
FIGURE 4.10 The genomic structure and expression pattern of the ATP6G gene 120
FIGURE 4.11 Multiple alignment of human ATP6G with orthologous V-ATPase G-subunits 121
FIGURE 4.12 The PROSITE consensus sequence for the E-F hand loop 126
FIGURE 4.13 The putative EF-hand calcium binding domain of AIF-1 126
FIGURE 4.14 The dot-matrix alignment of the rat BART-1 transcript amino acid sequence against the human AIF-1 genomic region DNA sequence 129
FIGURE 4.15 The putative structure of 1C7 with the structures of the cell surface molecules CD152 (CTLA-4) and the T cell receptor (TcR) for comparison 133

FIGURE 5.1 The location of PCR products used for mutation analysis across the genomic sequences of AIF-1, 1C7 and ATP6G 145
FIGURE 5.2 Examples of polymorphisms identified by sequence analysis. 146
FIGURE 5.3 Polymorphisms, and database differences, identified within a 2.3kb segment of genomic DNA covering the AIF-1 region 147
FIGURE 5.4 Polymorphisms and database differences identified within a 2.3kb segment of genomic DNA covering the 1C7 region 148
FIGURE 5.5 Polymorphisms and database differences identified within a 2.2kb segment of genomic DNA covering the ATP6G region 149
FIGURE 5.6 Polymorphisms and database differences identified within a 0.55kb segment of genomic DNA covering the BAT1promoter region 149
FIGURE 5.7 PCR-SSP results for 1C7 and AIF-1 154

TABLE OF TABLES

TABLE 2.1 PCR-SSP thermal cycling parameters 62

TABLE 3.1 Summary of the sequencing of cosmids TN62 and TN82 70
TABLE 3.2 Summary of gaps filled by PCR amplification to complete the sequencing of the overlapping cosmids TN62 and TN82 71
TABLE 3.3 The co-ordinates for the genomic sequence database entries in relation to the deduced sequence of overlapping cosmids TN62 and TN82 (Figure 3.3) 72
TABLE 3.4 The database sequences found to encode the different splice variants of LST-1 76
TABLE 3.5 The co-ordinates of additional potential coding regions of no assigned status, comprising alignments with ESTs, known genes, and predictions by the GRAIL exon prediction program 81
TABLE 3.6 The RepeatMasker server output summary for the genomic sequence contained within the overlapping Cosmids TN62 and TN82 84
TABLE 3.7 The exon co-ordinates for known genes, putative genes, pseudogenes and gene fragments found within the combined 81800 nucleotides of DNA sequence of overlapping Cosmids TN62 and TN82 84
TABLE 3.8 Details of the LST-1 gene open reading frames 102

TABLE 4.1 Transcript specific RT-PCR primers for the AIF-1 gene 116
TABLE 4.2 Transcript specific RT-PCR primers for the 1C7 gene 118
TABLE 4.3 Transcript specific RT-PCR primers for the ATP6G gene 123

TABLE 5.1 PCR templates generated for mutation analysis of the AIF-1, 1C7 and ATP6G genomic regions and the BAT1 promoter region 145
TABLE 5.2 The HLA haplotypes for the 12 heterozygous control DNAs used in the mutation analysis and the homozygous database entries that overlap with the genomic regions analysed 145
TABLE 5.3 PCR-SSP genotyping assay primer mixes 154

Chapter 1

CHAPTER 1: INTRODUCTION

1.1 THE MAJOR HISTOCOMPATIBILITY COMPLEX.

The human Major Histocompatibility Complex (MHC) is among the most extensively and diversely studied regions of the human genome. Within the scientific community it has stimulated the interest of many disciplines, including: immunology, molecular biology, evolutionary biology, classical genetics and disease association. Techniques developed in the study of the MHC are now being applied to the whole genome. Orthologous regions have been studied in detail in mouse, chicken, domestic cattle, pig, marsupials and fish. At the time of writing this thesis the genomic DNA sequence of the MHC had been completed in human, mouse and chicken. Projects were also underway to achieve the same in puffer fish. Furthermore, regions within the MHC, specifically the MHC class III region, are considered to be among the most gene-dense regions of the human genome and an increasing proportion of the genes encoded within the MHC as a whole are being found to be involved in the immune or inflammatory responses (Aguado *et al.* 1996; Dawkins *et al.* 1999; Price *et al.* 1999, and references therein).

The MHC spans approximately 4 megabases (Mb) of DNA in the chromosome band 6p21.3 and is divided into three regions. The class I and class II regions are located at either end of the complex, with the class I region at the telomeric end and the class II region at the centromeric end. Separating these is approximately 0.9 Mbp of DNA that has been termed the class III region (Campbell & Trowsdale 1997). A detailed map of the human MHC is shown in Figure 1.1.

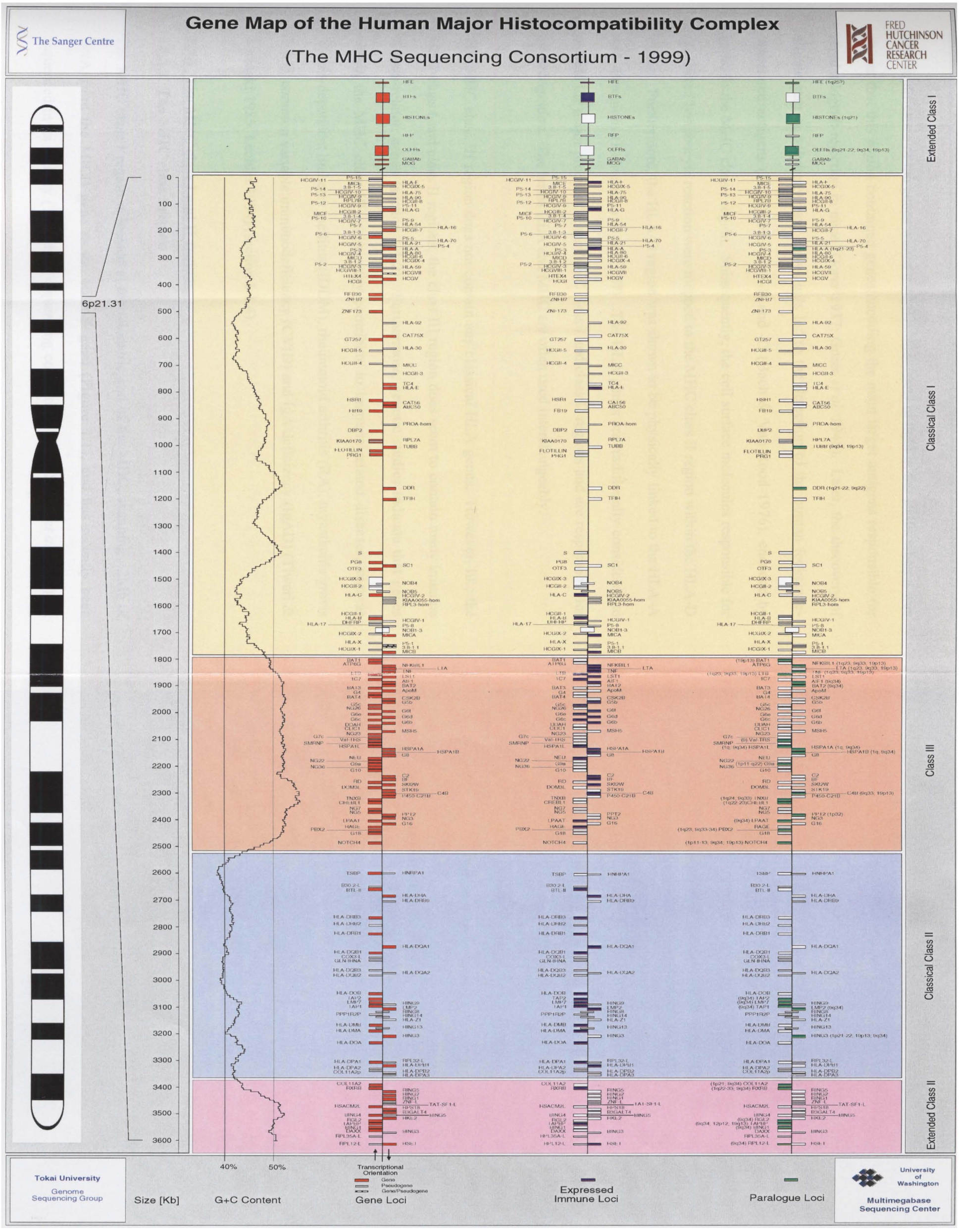


Figure 1.1 Gene map of the Human Major Histocompatibility Complex.
This figure has been taken from: The MHC sequencing consortium, *Nature* **401,** 921-923 (1999). The scale starts at the classical class I gene HLA-F and ends at the HSET gene in the extended class II region. The estimated 4Mb of the extended class I region at the telomeric end is not shown to scale. The G+C content is shown next to the scale. The 'gene loci' column shows the approximate positions and transcriptional orientations of all identified genes and pseudogenes. The 'expressed immune loci' column highlights genes that encode proteins that are thought to be involved in some way in the immune system. The 'paralogue loci' column highlights the genes for which paralogues have been identified on other chromosomes. HLA-A is marked as representative of all class I-like paralogues. The cytogenic location(s) of the paralogue(s) is/are indicated next to the gene name.

The significance of the genes within the MHC first came to light in the 1940s when it was shown that this region contained genes encoding the main target antigens responsible for rapid allograft rejection in mice (Gorer 1937; Gorer *et al.* 1948). Later, the discovery of alloantisera in polytransfused individuals led to the detection of the human leukocyte-associated antigens (HLA), HLA-A, -B and -C (Bach & Van-Rood 1976), encoded by genes in the MHC class I region. Subsequently, the cell surface molecules responsible for the mixed lymphocyte reaction were mapped to the MHC class II region within the HLA-D region. Furthermore, the HLA-D region was shown to be physically linked to the HLA-A and -B regions (Reinsmoen *et al.* 1977). The HLA genes are traditionally considered to encode the major products of the MHC class I and class II regions and are responsible for allorecognition. The terms HLA and MHC are often used interchangeably.

The MHC class III region does not contain any classical HLA genes. However, like the MHC class I and class II regions, the class III region does seem to contain many immune system-related genes (section 1.1.2). Furthermore, genetic studies indicate that genes within the MHC class III region may contribute to immune-related diseases, such as insulin-dependent diabetes mellitus (IDDM), rheumatoid arthritis (RA), myasthenia gravis (MG), common variable immunodeficiency (CVID) and IgA deficiency (IgAD) (Tiwari & Terasaki 1985; Lechler 1994) (section 1.1.4).

1.1.1 The MHC class I and II regions.

1.1.1.1 The MHC class I region

The MHC class I region is located at the telomeric end of the MHC, spanning ~2Mb of DNA from the MICB gene to the HLA-F gene (Figure 1.1), and has recently been sequenced in its entirety (The MHC sequencing consortium 1999; Shiina *et al.* 1999). A

total of twenty five overlapping cosmid clones subcloned from three YAC clones, two BAC and eight PAC clones, covering 1,796,938bp of DNA were sequenced to achieve this. So far a total of 118 genes have been identified in the class I region. These include, 30 expressed genes and another 27 potential coding sequences. In addition, 61 pseudogenes have been identified.

As mentioned above, the MHC is best known for the three classical HLA class I genes HLA-A, HLA-B and HLA-C. These encode the α-chains (or heavy chains) of the class I molecules, which associate with β2-microglobulin (β2-m) to form functional MHC class I molecules. They are expressed on most nucleated cells, but at varying levels in different tissues and cell types (Israel & Kourilsky 1996). HLA-A, with 83 reported alleles, and HLA-B, with 186 reported alleles, are highly polymorphic while HLA-C is considerably less polymorphic having only 42 alleles (Bodmer *et al.* 1997). The classical class I molecules are involved in the binding and presentation of peptides derived from the degradation of intracellular proteins. They were originally identified as the molecules responsible for tumour and allograft rejection between unrelated individuals and were subsequently shown to be required for T cell-mediated cytotoxicity towards virally infected cells. This cytotoxic T cell response depends on the presentation of foreign antigen in the context of self MHC class I molecules at the cell surface (reviewed in Germain 1994). Endogenous or viral proteins are degraded to generate short peptides, which are transported to the endoplasmic reticulum (ER), where they associate with MHC class I molecules. The peptide-MHC class I complex is subsequently transported to the cell surface, where the foreign peptides are recognised by the T cell receptors of $CD8^+$ cytotoxic T lymphocytes, resulting in destruction of the target cell (Germain 1994).

Also within the class I region are the non-classical class I genes: HLA-E, HLA-F and HLA-G (reviewed in Geraghty 1993). They are highly related to the classical class I genes (Bjorkman & Parham 1990) but show only limited polymorphism. The protein products of HLA-E, -F and -G are known to be expressed and to associate with β2-m (Geraghty 1993). However, they have shortened cytoplasmic tails compared to classical class I α chains and their cell surface expression is limited. Cell surface expression of HLA-E is known to depend upon its association with the leader peptide of other class I molecules (Braud *et al.* 1998). Receptors for HLA-E have been identified on natural killer (NK) cells and have been shown to be involved in both the positive and negative regulation of NK-cell-mediated cytotoxicity (Braud *et al.* 1998). HLA-G expression has been detected on two types of foetal cytotrophoblast cells: the villous cytotrophoblasts from the first trimester of gestation; and the chorionic-membrane cytotrophoblasts from the third trimester (reviewed in Carosella *et al.* 1996). This has led to the suggestion that HLA-G has an important role in materno-fetal immune tolerance. Other class I-related sequences, such as HLA-X (Shiina *et al.* 1998) and HLA-J (Messer *et al.* 1992), encode pseudogenes or gene fragments (Geraghty *et al.* 1992; Geraghty 1993).

Members of yet another lineage of class I-related genes and pseudogenes, called MIC genes (MHC class I chain related genes), are located on the border between the class I and the class III regions (Bahram *et al.* 1994). Only two of these are known to be expressed: MICA and MICB. However, they are highly divergent from the classical and non-classical class I genes and differ in their patterns of expression (Groh *et al.* 1996), MICA being expressed in fibroblasts and epithelial cells (Bahram *et al.* 1994). Also, unlike the classical class I molecules, MICA and MICB are not associated with β2-microtubulin and are conformationally stable in the absence of peptide ligands (Groh *et al.* 1996). MICA is

highly polymorphic with more than 20 alleles identified to date (Ando *et al.* 1997) whilst, MICB is somewhat less polymorphic with 8 polymorphic sites identified (Visser *et al.* 1998). Interestingly, the presence of diagnostic residues in the MICA amino acid sequence suggests that MICA at least, despite its divergence from the classical HLA class I antigens, folds in a similar manner to the classical class I chains and, therefore, may have the capacity to bind peptide (Bahram *et al.* 1994). Also, the predominant expression of MICA in gastrointestinal epithelium and the evidence that both MICA and MICB are recognised by intestinal epithelial T cells expressing diverse V delta 1 gamma delta TCRs (Griffith *et al.* 1998) has prompted the suggestion that these two molecules function as stress-induced antigens and that they may play a central role in the immune surveillance of damaged, infected or otherwise stressed intestinal epithelial cells (Groh *et al.* 1996; Griffith *et al.* 1998; Steinle *et al.* 1998).

In addition to the HLA genes there are more than 50 non-HLA related genes within the class I region. Among these are a family of olfactory receptor genes (OLFR) (Fan *et al.* 1995); the OCT3 oncogene (Geraghty *et al.* 1992); a β-tubulin-like gene (TUBB) (Geraghty *et al.* 1992; Volz *et al.* 1994); HSR1, which encodes a GTP-binding protein (Vernet *et al.* 1994); the *trans*-activating factor SC1 (TCF19) with similarity to the *tre* oncogene (Goldsworthy *et al.* 1997); and the NOB genes 1-5 (Mikzuki *et al.* 1997) which were identified by Northern blot analysis, but show no similarity to known sequences within the databases.

A number of multigene families have been identified within the class I region including the P5 and HCG families. The P5 multigene family (Vernet *et al.* 1993) contains at least eight members, of unknown function. P5-1 has, however, been found to be specifically

expressed in lymphoid cells and tissues. A paper by Avoustin *et al* (1994) indicates that members of the P5 family represent chimeric transcripts, where the first exon of an HLA class I gene, encoding the signal peptide, has been spliced to an unrelated sequence. The P5 family also shows some similarity to the HERV-L and HERV-16 human endogenous retroviruses (Kulski & Dawkins 1999) which has led to the suggestion that P5-1, at least, expresses an antisense transript to retroviral pol mRNA and may have a role in immunity to retroviral infection. A number of genes termed HCG genes (haemochromatosis candidate genes) (Giffon *et al.* 1996) are located within the class I region. Of these two belong to multigene families: HCG II and HCG IV, with so far a total of 3 and 11 members identified, respectively (Pichon *et al.* 1996). No information is at present available on the structure or function of these two gene families. Two further multigene families have also been described in the literature: 3.8-1 and pMC6.7 (Venditti *et al.* 1994; Pichon *et al.* 1997). However, little is known about them.

The class I region also contains 15 HLA class I pseudogenes, 14 P5 pseudogenes, 15 HCG pseudogenes, four 3.8-1 pseudogenes and 13 others (Geraghty 1993; Campbell & Trowsdale 1997; Shiina *et al.* 1999). The large numbers of both pseudogenes and multi-gene families within this region suggests that genomic duplication contributed considerably to the evolution of the class I region (Shiina *et al.* 1999). In fact, analysis of the entire MHC class I genomic sequence has revealed a number of duplicated segments (Shiina *et al.* 1999). Interestingly, there is also a class I-related gene, HLA-HFE, more than 4 Mb telomeric of HLA-A, suggesting a further duplication event. This has given rise to the idea of the 'extended' MHC class I region, which includes this region (Figure 1.1).

1.1.1.2 The MHC class II region

The MHC class II region covers ~1.2Mbp at the centromeric end of the MHC extending from the HSET gene at its most centromeric end to the region just centromeric of NOTCH4 at its telomeric end (see Figure 1.1). However, the exact length of this region differs depending on haplotype, due to variation in numbers of DRB genes (Andersson *et al.* 1994; Gongora *et al.* 1997; Beck & Trowsdale 1999). It is split into two segments, the classical class II region or HLA-D region and the extended class II region (Figure 1.1). As with the MHC class I region, the class II region genomic sequence has now been resolved (Beck & Trowsdale 1999). A total of ~30 YACs, BACs, PACs and cosmids, representing a "minimum tiling path" (Beck & Trowsdale 1999; The MHC sequencing consortium 1999), were sequenced to achieve this.

There are a number of striking differences between the MHC class II region and the class I and class III regions. Firstly, in terms of isochore structure, the class II region is a Giemsa stained dark band, i.e. a relatively G+C-poor region, whereas the class III region is a Giemsa stained light band, or G+C-rich region. Quantitatively the class II region is an L1 isochore (Bernardi 1995) with an average G+C content of ~40% whilst the class I region is an H1 isochore (G+C content 45.8%) and the class III region is an H2 isochore (G+C content ~53%) (Shiina *et al.* 1999). The MHC class II region is also relatively gene poor, with one gene every 25kb, when compared to the MHC class I and III regions which contain one gene every 15.2kb and 14.3kb, respectively. Of the genes so far characterised within the class II region many are involved in antigen processing, presentation and immunity. Also, in contrast to the class I region, there are no HLA class II-like genes found elsewhere in the human genome (Beck & Trowsdale 1999).

The MHC class II region contains the classical HLA-D loci: HLA-DP, -DQ and -DR, as well as other genes encoding proteins with an immune related function. The HLA-D loci each contain at least one A and one B gene, which encode the class II α and β chains, respectively and are paired together on the chromosome. The one exception to this is the HLA-DOA (also known as HLA-DNA) and HLA-DOB genes which encode the α and β chains of the heterodimer, HLA-DO (Douek & Altmann 1997) and are separated on the chromosome by ~190kb of sequence containing ~16 genes. Some of the HLA-D loci do, however, represent pseudogenes and only the DPA1, DPB1, DQA1, DQB1, DRA, DRB1, DRB3, DRB4, DRB5, DOA and DOB genes are expressed as proteins (Bodmer *et al.* 1990). The genes encoding the class II α chains, DPA and DQA are highly polymorphic while DRA is invariant. All individuals express a DRB1-encoded β-chain, which associates with the protein product of DRA to form the heterodimic class II molecule found on the cell surface. In addition, DRβ3, 4 and 5 are expressed at the cell surface with the DR α-chain, but only at low levels and only in certain class II haplotypes (reviewed in Apple & Erlich 1996). For example, DRB4 is found only on DR4, DR7 and DR9 haplotypes, while DRB5 is found on DR15, DR16 and, rarely, DR1 haplotypes. Functional class II molecules present exogenous antigen to T helper cells (CD4+). Cell surface expression of classical class II molecules is limited to B cells, cells of myelomonocytic lineage, activated T cells and non-lymphoid cells such as endothelial and dendritic cells (Alonso *et al.* 1985). Cells expressing class II-peptide complexes are known as antigen presenting cells (APC). Extracellular proteins are internalised by these antigen presenting cells and are degraded to generate peptides within the endocytic pathway. In the endoplasmic reticulum (ER), MHC class II heterodimer molecules are assembled with the invariant chain (Ii). These are then transported to endocytic compartments where Ii is proteolytically truncated to a nested set of class II-associated

invariant chain peptides (CLIP) (Busch *et al.* 1998). Once in the endocytic pathway, HLA-DM (DM) catalyses the release of CLIP from the peptide binding groove, allowing the processed exogenous peptides to bind to the MHC class II molecule. The peptide-MHC class II complex is subsequently transported to the cell surface, where it can be recognised by CD4+ helper T cells, leading to the initiation and propagation of an immune response.

It is thought that the HLA class I and class II genes are evolutionarily related and X-ray crystallography studies have revealed that they encode proteins with a very similar three dimensional structure (Brown *et al.* 1993 and references therein). Interestingly, the HLA-DMA and -DMB genes within the class II region are more divergent than the other class II genes (about 30% identity at the protein level). They show about equal levels of similarity to the classical class I and class II sequences and are thought to have diverged around the same time that the MHC class I and class II genes diverged (Beck & Trowsdale 1999). As mentioned above, the DM gene products form a heterodimer which is involved in the loading of the class II complexes with antigenic peptide. However, as well as catalysing the release of CLIP, DM is also thought to act as a chaperone and a peptide 'editor' for the class II molecule (van-Ham *et al.* 1997; Busch *et al.* 1998). It binds to the class II molecules during peptide exchange and is thought to stabilise the complex, thus preserving the peptide binding site (Denzin *et al.* 1996; Kropshofer *et al.* 1996; Vogt *et al.* 1997), and also may act to favour the presentation of high affinity binding peptides by facilitating additional rounds of peptide exchange (Kropshofer *et al.* 1996). The role of the HLA-DOA -DOB heterodimer, DO, in antigenic peptide presentation appears to be as a negative modulator of DM (Denzin *et al.* 1997; van-Ham *et al.* 1997). DO stably associates with DM in the lysosome inhibiting the catalytic action of DM at the level of class II peptide loading.

Also located in the classical class II region is a cluster of genes that encode proteins involved in the processing of antigens to generate peptides which associate with HLA class I molecules. The TAP genes, TAP1 and TAP2, along with two LMP genes, LMP2 and LMP7 are located between the DM and DQ loci (Parham 1990). The protein products of the TAP genes form a heterodimer (TAP-A) which is localised in the ER membrane where it mediates peptide transport from the cytosol into the ER and facilitates peptide loading onto HLA-class I molecules (Ortmann *et al.* 1994; Suh *et al.* 1994). The LMP genes encode components of the proteosome which is thought to process cytoplasmic proteins by proteolytic digestion to generate peptides suitable for binding to class I molecules (Kelly *et al.* 1991; Glynne *et al.* 1993; Goldberg 1995). Apart from numerous pseudogenes, the only gene in the class II region that encodes a protein not involved in the immune response is RING3, which is expressed ubiquitously and is a homologue of the Drosophila gene, female sterile Homeotic gene (fsh) (Beck *et al.* 1992).

Genes within the extended MHC class II region appear to be more varied in function than the classical class II region. Just centromeric of the HLA-DP genes is the human alpha 2(XI) collagen gene (COL11A2) (Figure 1.1) which is a member of the type XI collagen family, a fibril forming collagen (Lui *et al.* 1996). The Retinoid X receptor B (RXRB) gene is a member of a novel subgroup of nuclear hormone receptors implicated in retinoid-mediated gene regulation (Hoopes *et al.* 1992; Endo *et al.* 1997) which in mouse appears to be regulated by tumor necrosis factor (Sugawara *et al.* 1998). The HSACM2L gene is a highly conserved gene with homology to the yeast SAC2 protein (Walter & Gunther 1998), and closely linked to this is the Rps18 gene which encodes the 40 S ribosomal protein S18 (Walter & Gunther 1998). 180 kb centromeric of the DP genes is the TAPBP (TAPASIN) gene (Herberg *et al.* 1998). TAPBP has recently been shown to be required for antigen

presentation through MHC class I molecules by forming a trimeric complex with TAP1 and TAP2 (TAP-A) facilitating the association of TAP with class I molecules (Li *et al.* 1999). The DAXX gene, just centromeric of TAPBP, is a novel signalling protein that binds specifically to the death domain of the Fas cell surface receptor (Yang *et al.* 1997; Herberg *et al.* 1998) and forms part of a signalling pathway involved in apoptosis (Yang *et al.* 1997; Chang *et al.* 1998). Lastly, at the most centromeric end of the class II region is the HSET (KNSL2) gene, a member of the kinesin superfamily and the human homolog of the murine Tctex-7 gene (Tripodis *et al.* 1998).

1.1.2 The MHC class III region

The MHC class III region spans ~900 kb between the class I and class II regions (Figure 1.1 and Figure 1.2). It is considered to be among the most gene dense regions of the human genome with on average one gene every 14.3kb. It is also one of the most complex regions with examples of overlapping genes, such as YA and YB that overlap with P450c21A and P450c21B (Bristow *et al.* 1993), and genes within genes, such as XB-S which is within the TN-X gene (Tee *et al.* 1995). In addition, there are large numbers of alternatively spliced genes, for example: Allograft Inflammatory Factor-1 (AIF-1) (Neville & Campbell 1999 and Chapter 4 of this Thesis), 1C7 (Gruen & Weissman 1997 and Neville & Campbell 1999 and Chapter 4 of this Thesis), Leukocyte specific transcript-1 (LST1) (Gruen & Weissman 1997; Neville & Campbell 1997 and Chapter 3 of this Thesis) and Palmitoyl-protein thioesterase-2 (PPT2) (Aguado & Campbell 1999; Soyombo *et al.* 1999).

Since the MHC class III region does not contain any classical HLA genes there has been some debate as to what to call this central region of the MHC. It has been argued that the

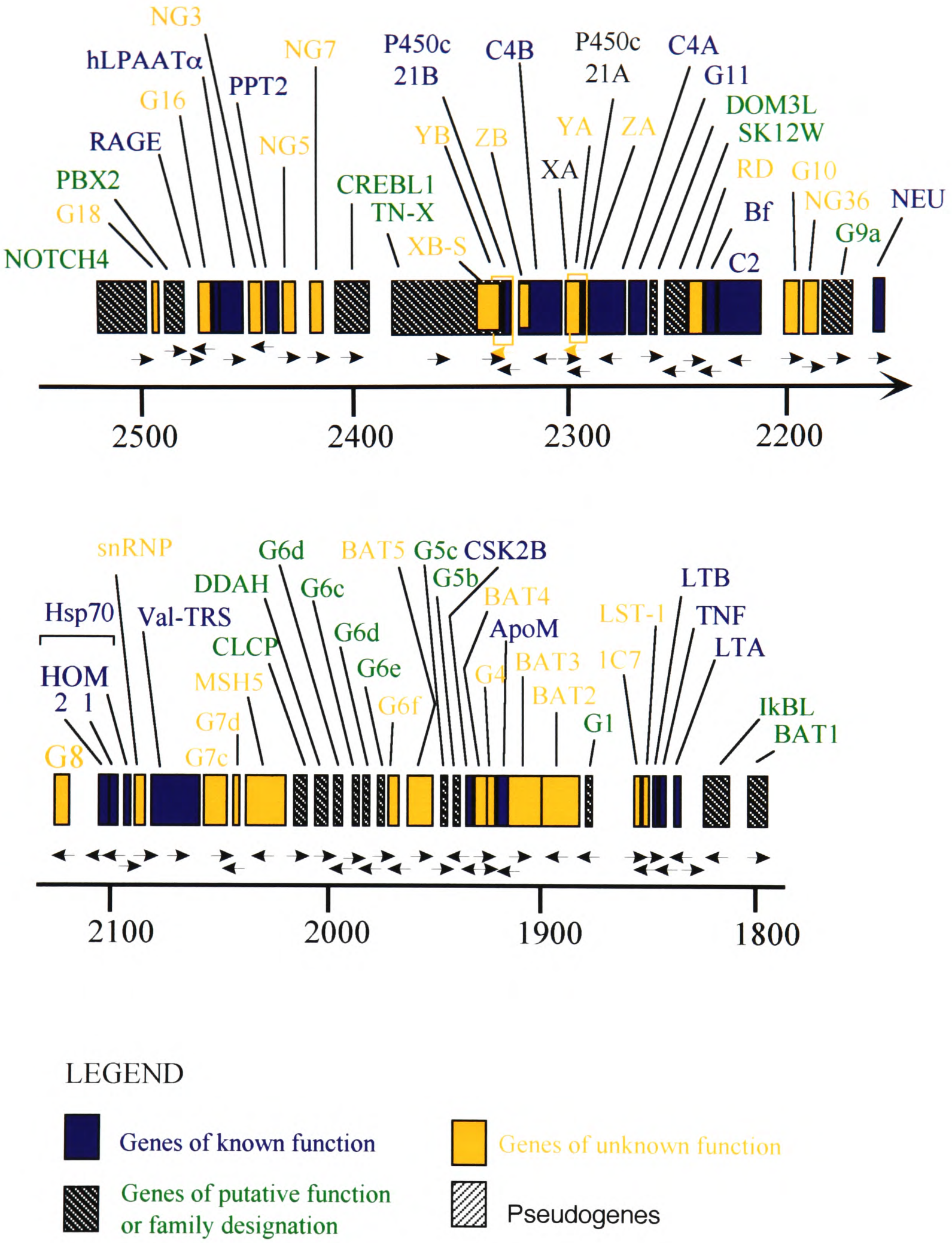


FIGURE 1.2 Map of the human MHC class III region.
The position of genes are indicated by boxes and the direction of transcripts, where known, is indicated by an arrow. The map is adapted from Aguado (1997).

class III region has no functional connection with the MHC class I and class II regions and was trapped between its neighbouring regions, by chance, during the evolution of the MHC (Klein *et al.* 1983). However, with the increasing amounts of data generated over the last decade this view is evolving (see section 1.1.3). It is now becoming increasingly apparent that, in common with the MHC class I and II regions, the class III region contains many genes that encode proteins with at least putative involvement in immune and inflammatory responses (Aguado *et al.* 1996).

1.1.2.1 The Complement genes

There are four genes that encode components of the complement system within the MHC class III region; these are C2, Factor B (Bf), C4A and C4B. The complement system is the principal effector mechanism of humoral immunity and its activation leads to cell lysis, opsonisation, clearance of immune complexes, regulation of B cell responses and anaphylactic and chemotactic responses (reviewed in Law & Reid 1995). The complement system can be activated by three pathways in response to foreign antigen. The proteins C2 and C4 are involved in the classical (Campbell & Law 1992) and lectin pathways (Suankratay *et al.* 1998) of complement activation, while Bf is involved in the alternative pathway (Pangburn 1992). C4 is encoded by two genes, C4A and C4B, the protein products of which differ by only four amino acids, but nevertheless have very different covalent binding activities (Dodds *et al.* 1985; Yu *et al.* 1986). That C2 and Bf encode functionally-related proteins and are separated by only 421 bp suggests that they arose by gene duplication of an ancestral locus. This may also be the case for C4A and C4B which lie ~10 kb apart, approximately 30 kb centromeric from Bf. C4A and C4B are highly polymorphic (Yu 1991) while C2 and Bf show more limited polymorphism (Cross & Thomson 1990).

1.1.2.2 The heat shock protein genes

The class III region contains three genes encoding heat shock proteins of the 70 kDa family (hsp70): HSP70-1, HSP70-2 and HSP70-HOM (Milner & Campbell 1990). The HSP70s are a highly conserved family of proteins, found in all species, and are expressed both constitutively and in response to a variety of stress factors. HSP70-1 and -2 both encode the major heat-inducible HSP70 and their expression is upregulated in response to heat shock. However, sequence variation in their 5'- and 3'-untranslated regions may result in differential expression in response to other stress factors, such as oxidative injury and inflammation (Moseley *et al.* 1993). All HSP70s can bind unfolded proteins and peptides and act as chaperones in the synthesis, folding, assembly, translocation and degradation of proteins during normal cellular processes and following stress. Thus, the HSP70s have been suggested to have a role in antigen processing and have also been implicated in the transport of peptides during antigen processing in tumour cells (Pierce 1994; Srivastava *et al.* 1994).

1.1.2.3 The Leukocyte Antigen 6 Superfamily gene cluster

Recent characterisation of a 30kb segment of DNA between the HSP70 genes and the TNF gene cluster has revealed a cluster of five genes encoding new members of the leukocyte antigen 6 (Ly-6) superfamily: G5b, G5c, G6c, G6d and G6e (Ribas *et al.* 1999, G. Ribas-Despuig, personal communication). This group of leucocyte antigens were first identified in mouse and are usually 70-80 amino acids in length with 10 conserved cystine residues (Shevach & Korty 1989). In addition, they are all GPI anchored proteins. Members of the Ly-6 family are differentially expressed in several hemopoietic lineages, especially in T lymphocytes (Rock *et al.* 1989). They appear to function in signal transduction and cell activation and are thought to be important in leukocyte maturation (Rock *et al.* 1989;

Shevach & Korty 1989). Interestingly, the known members of the Ly-6 family in mouse are all encoded by genes within a single segment of chromosome 15, a region linked to susceptibility to radiation induced lymphoid malignancy (LeClair *et al.* 1987; Huppi *et al.* 1988). However, these do not appear to be orthologs of the human MHC-linked genes.

1.1.2.4 RAGE

The RAGE gene is found at the centromeric end of the MHC class III region and encodes a 'receptor for advanced glycosylation end products of proteins' (RAGE), which is a member of the immunoglobulin superfamily of cell surface molecules. It shares significant amino acid sequence similarity with MUC18, N-CAM and the cytoplasmic domain of CD20 (Neeper *et al.* 1992). 'Advanced glycosylation end products of proteins' (AGEs) are none-enzymatically glycosylated proteins that accumulate in vascular tissue as a result of ageing and this occurs at an accelerated rate in diabetes. AGEs are not thought to be the natural ligand for RAGE. Instead, RAGE is thought to be a cell adhesion molecule or growth factor receptor. This is supported by evidence that RAGE contributes to the enhanced adherence of diabetic red cells to endothelium and that both AGEs and antibodies to RAGE can stimulate endothelial proliferation (reviewed in Aguado *et al.* 1996).

1.1.2.5 hLPAATα (G15)

The hLPAATα gene (G15) encodes a protein which shows significant sequence similarity with the enzyme lysophosphatidic acid acyltransferase (LPAAT) from different yeast, plant and bacterial species over the entire polypeptide (Aguado & Campbell 1998). hLPAATα is 48% identical to hLPAATβ encoded on chromosome 9p34.3 (Eberhardt *et al.* 1997). LPAAT is an enzyme that converts lysophosphatidic acid (LPA) to phosphatidic

acid (PA) in lipid metabolism. Both LPA and PA play a role in intracellular signalling and inflammation, which adds hLPAATα to the list of candidate immunoregulatory proteins encoded in the class III region.

1.1.2.6 NEU (G9)

NEU is located between the complement genes and the hsp70 genes in the class III region and encodes a protein with 28% sequence identity to the *Salmonella typhimurium* sialidase. Sialidase enzymes catalyse the removal of sialic acid residues from glycoproteins and glycolipids and thus play an important role in modulating cellular events. NEU corresponds to murine Neu-1, a liver sialidase encoded in the S region of the mouse H-2 complex (Milner *et al.* 1997). Recently, Neu-1 has been shown to be required for the regulation of IL4 production during T cell activation (Chen *et al.* 1997b) which has important implications for the regulation of the Th-1/Th-2 balance during a T cell response. NEU is likely to have a similar role in humans and thus provides another example of a putative immune system related protein encoded within the class III region.

1.1.2.7 SK12W (SKIV2L/ G11a)

SK12W (Lee & Cotanche 1995; Albertella *et al.* 1996) lies between the Bf and C4A genes and encodes a protein with significant similarities to the yeast Superkiller 2 (SKi2) protein (Liang *et al.* 1996). SKi2 confers resistance to killer toxin viruses in yeast cells by blocking the translation of non-polyadenylated or non-capped mRNAs (Widner & Wickner 1993). It is possible then that one function of SK12W is as a mediator of antiviral effects towards RNA viruses within the cell (Gruen & Weissman 1997).

1.1.2.8 The G1 – BAT1 interval, a cluster of genes involved in the inflammatory response.

The G1 - BAT1 segment of the MHC class III region is the focus of this Thesis (Chapter 3). At the start of the project the transcripts mapped to this region included: G1, Leukocyte specific transcript-1 (LST-1(B144)), Lymphotoxin β (LTB), Tumor necrosis factor (TNF), Lymphotoxin α (LTA), NB6, I-Kappa B-like protein (IKBL) and BAT1. In addition, during the course of our investigation two further transcripts were identified by other parties: the Allograft Inflammatory Factor-1 (AIF-1) and 1C7. AIF-1 was mapped to a region telomeric of BAT2 (Utans *et al.* 1996) and 1C7 between LST-1 and LTB (Nalabolu *et al.* 1996). Most striking about the genomic interval described above is the proportion of genes encoded within it that are connected with the inflammatory response.

Most notable and well studied within the G1-BAT1 interval is the LTB/TNF/LTA gene cluster (for full review see Guss & Dower 1995). TNF is one of the major mediators of the inflammatory response and is synthesised by many cell types, including: monocytes/ macrophages, lymphocytes and fibroblasts (Guss & Dower 1995). It is also known to be involved in the pathogenesis of IDDM both in humans and in animal models (Tisch & McDevitt 1996), in RA (Brennan *et al.* 1992; Maini *et al.* 1995) and in sensitivity to cerebral Malaria (Wilson *et al.* 1997; McGuire *et al.* 1999; Wattavidanage *et al.* 1999). The biological roles of LTA and LTB are less well understood. In general, both LTA and TNF exist as homotrimers, bind with near identical affinity to the same pair of cell surface receptors, p55 and p75TNFR, and display a similar spectra of activities *in vitro*, although LTA has a 50-fold less potent cytotoxic effect than TNF (Schuchmann *et al.* 1995). However, in contrast to TNF, LTA is specifically expressed by activated lymphocytes. LTB is a type II integral membrane glycoprotein that associates with LTA to form

heterodimers, targeting it to the cell surface. This LTA-LTB complex can then induce activation of NFκB in certain cell lines by binding with the LTB receptor (Gruen & Weissman 1997). Then, in turn, NFκB is capable of activating the expression of a great variety of inflammatory response related genes (Mackay *et al.* 1996). Unlike LTA, TNF does not associate with another protein for cell surface expression, but is a type II membrane protein itself and can also be secreted. In addition, the LTA homo- or hetero-trimer shows much greater stability and displays a 30-fold greater mitogenic effect than TNF. The possibility has been raised, therefore, that despite the sequence similarities between TNF, LTA and LTB they may have distinctly different roles in immune regulation (Aguado *et al.* 1996).

Of the other genes found within the G1-BAT1 region, a putative function had previously been assigned to two: IKBL and BAT1. IKBL (NFKBILI) is a putative member of the IκB family of proteins (Albertella & Campbell 1994) which regulate the NFκB family of transcription factors. As mentioned above in relation to the LTA-LTB complex, NFκB family members are important for the inducible expression of a wide variety of immunity related genes, including cytokines and cytokine receptors (Baeuerle & Henkel 1994; Thanos & Maniatis 1995; Aguado *et al.* 1996). IKBL is, therefore, another strong candidate for involvement in the inflammatory response. BAT1 (D6S81E) encodes a putative nuclear RNA helicase of the DEAD family and may have a role in the translation of closely-linked genes (Peelman *et al.* 1995) again suggesting a tentative involvement in inflammation.

The remaining genes: G1, AIF-1, LST-1 (B144), NB6, and 1C7 (Figure 1.2) are of unknown function. Some indication of their functions may, however, be inferred from the

cells in which they are expressed. G1, AIF-1, LST-1 and 1C7 all appear to be expressed exclusively in myeloid cells, suggesting their involvement in the immune response. G1 is only expressed at significant levels in cell lines representing monocytes/macrophages and T cells (Olavesen *et al.* 1993). Analysis of its encoded amino acid sequence has also revealed two potential EF-hand type calcium binding domains. This suggests a possible function as a calcium binding protein and perhaps a role in an intracellular pathway during activation or stimulation of T cells (Olavesen *et al.* 1993). AIF-1 (Allograft Inflammatory Factor 1) appears to be specifically expressed in monocyte and lymphocyte cell lineages (Utans *et al.* 1996). AIF-1 (for full review see Chapter 4), as its name suggests, was originally identified due to its elevated levels of expression in experimental cardiac transplants and is up regulated by IFN-γ. LST-1 is the human homolog of the mouse B144 gene (Holzinger *et al.* 1995) and appears to be specifically expressed in B cells and macrophages (Tsuge *et al.* 1987). Like AIF-1, it is induced by IFN-γ (Holzinger *et al.* 1995), although no protein has been observed as yet *in vivo*. The regulation of LST-1 expression is clearly complex, as a large amount of alternate splicing has been observed (Neville & Campbell 1997 and Chapter 3). 1C7 has been found to be expressed only in spleen (Nalabolu *et al.* 1996) and, like LST1, has been observed to exist in alternately splice forms (Gruen & Weissman 1997; Neville & Campbell 1997, Chapter 4). Lastly, the partial NB6 transcript was isolated from a JY cDNA library (B cell), although no clear details of expression pattern were revealed from Northern blot analysis (Albertella *et al.* 1996).

1.1.2.9 MHC class III region genes with no obvious immune related function.

The genes described above represent only a small proportion of the genes found within the MHC class III region. Much further characterisation remains to be carried out before any

conclusions can be drawn about the collective role of the MHC class III region in immunity. However, it is clear that a number of genes in this region encode proteins that are not, or are unlikely to be, involved in the immune response (reviewed in Aguado *et al.* 1996). P450c21B, encodes the adrenal enzyme steroid 21-hydroxylase (Higashi *et al.* 1986; White *et al.* 1986; Miller 1988), which is involved in the mineralocorticoid and glucocorticoid biosynthesis pathways. NOTCH4 is a homologue of the *Drosophila* Notch gene, which regulates a variety of events and is particularly important during early development (Fortini & Artavanis-Tsakonas 1993). The protein encoded by XB (G12) has been termed tenascin-X (TN-X) and is a member of a growing family of extracellular matrix proteins (Chiquet Ehrismann *et al.* 1994; Speek *et al.* 1996). The nucleotide sequence of the XA gene is ~99% identical to XB, but XA contains premature stop codons, so it is unlikely to encode a protein. The CSK2B (CKIIβ/G5a) gene encodes the casein kinase II (ckII) β subunit (Albertella *et al.* 1996). CSK2 has been linked to the regulation of cell growth and metabolism as many of its substrates (such as insulin, epidermal growth factor and a variety of cellular and viral oncoproteins) play key roles in cell growth. The Val-TRS (G7a) protein shows 48.3% identity with the valyl-tRNA synthetase (Trs^{Val}) of *Saccharomyces cerevisiae* and functional studies on recombinant Val-TRS have confirmed that it is the human Trs^{Val} (Hsieh & Campbell 1991). The G9a (BAT8) gene encodes an ANK repeat containing protein and may be involved in protein-protein interactions (Milner & Campbell 1993). The C-terminal region of G9a (which binds zinc *in vitro*) shows ~33% amino acid sequence identity with the C-terminal regions of *Drosophila* trithorax and its human homolog HRX, both of which are involved in the regulation of transcription during development (Milner & Campbell 1993). The G11 gene, adjacent to C4A, is transcribed into two alternatively spliced transcripts which encode protein products with serine/threonine protein kinase activity (Sargent *et al.* 1994; Gomez Escobar *et al.* 1998).

The PPT2 gene is a secreted glycosylated protein with S-thioesterase activity with the highest activity observed for acyl groups palmitic and myristic acids (Aguado 1997; Aguado & Campbell 1999). The PBX2 (G17) gene encodes a homeodomain containing protein, with significant similarity to PBX1 which is involved in t(1;19) chromosomal translocations in acute pre-B-cell leukemias (Aguado & Campbell 1995). Further copies of PBX2 also appear to be present on chromosomes 1 and 3, although the copy on chromosome 3 may be a processed pseudogene (Aguado & Campbell 1995).

1.1.3 Evolution of the MHC region

With the accumulation of data from sequencing and mapping studies on the MHC, we are slowly getting closer to understanding the evolution of this complex region. The structure of the MHC has been, with a few exceptions, strongly conserved throughout the evolution of most higher vertebrates (for reviews see, Kronenberg *et al.* 1994; Trowsdale 1994) and in all the amphibian, bird and mammalian species studied, the MHC class I, class II and class III regions are linked. Of the three sub-regions, the class III region is most conserved, the class I region is most divergent and the class II region lies somewhere between the two (Parham 1999). In mouse, the best studied of the higher vertebrates after man, the MHC is termed the Histocompatibility-2 (H-2) complex and is located on chromosome 17. The class II and class III regions are roughly colinear with those of the human MHC with almost identical occurrence, and order, of orthologous genes. However, the mouse class I region is considerably divergent from the human (Amadou *et al.* 1995) and contains many more class I-like genes, the precise functions of which are not all clear.

There are, however, a few situations where this rule of MHC conservation does not apply. The chicken MHC, which has been described as being a "minimal essential MHC"

(Kaufman *et al.* 1999) is one such example. This region contains 19 genes over 92kb and lacks many of the usual MHC genes including the LMP and class II α chain genes. Also, the ordering of the three regions is unusual, with the class I and class II regions next to each other and the class III region closely linked. It has been pointed out, however, that the chicken is considered a rather ancient form of bird and Southern blot analysis of pirching birds (house finch, red winged blackbird and scrub jay) suggests that their MHC regions are larger and more complicated (Edwards *et al.* 1999). Another example of unusual MHC organisation is seen in teliost fish which show no linkage between the class I and class II regions (Binguilac *et al.* 1997). This is not the case with cartilagenous fish, however, which are considered the most primitive organisms to have MHC class I and class II genes. As in humans, the class I and class II regions of these fish are linked. It is thought, therefore that the chicken and the teliost fish are the exceptions and that the ancestral MHC region organisation is similar to that of humans.

There appear to have been a number of chromosomal recombination events during the evolution of the human genome and there are at least three paralogous regions containing clusters of genes similar to the genes in the MHC (Kasahara 1999). These are found on chromosomes: **1q21-q25**, (e.g. RXRB/RXRG, TNF/OX40L, HLA class I/CD1); **9q33-q34** (e.g. RXRB/RXRA, C4A/C5, LTB/CD30L, TUBB/TUBB2) and **19p13.1-p33** (e.g. RING3/HUNKI, C4A/C3, LTB/CD70, MHC class I/FCGRT). However, so far no classical HLA class II-like genes have been found in the human genome outside the MHC class II region.

Based on the available cross species data and recombination data, the MHC as a complete set of genes seems to have arisen during the emergence of the jawed vertebrates

approximately 500 million years ago (Kasahara 1997). No MHC class I genes have been observed in the jawless fish or lower orders. The first chromosomal duplication event involving the MHC genes is also thought to have occurred at about the time of the emergence of the jawed vertebrates as MHC gene paralogs have been identified in both the teliost and cartilagenous fish (Kasahara 1997; Hughes 1998). The origin of the MHC class III region appears to predate these events since Trachtulec and colleagues (1997) have identified genes encoding PBX-like, NOTCH-like and tenascin-like genes within a 5.8Mb region of chromosome III of *Caenorhabditis elegans*. There is as yet, however, no evidence of any MHC class I or class II genes in *C. elegans*. It may well be, therefore, that many of the genes now present in the MHC class III region were the original inhabitants of the genomic segment that is now known as the MHC region (Klein & Sato 1998) and that the MHC class I and class II regions are relative newcomers (Kasahara 1999). With this ancient ancestry it is even more interesting that the MHC class III region has remained so well conserved throughout the animal kingdom, most strikingly between man and mouse, suggesting that there may have been some selective pressure to maintain this region intact.

1.1.4 The MHC and disease

The classical HLA class I and class II genes are highly polymorphic, both in terms of the number of alleles so far observed and the degree of nucleotide diversity amongst these alleles. Some of the class III region genes are also polymorphic, notably the C4 genes (Campbell *et al.* 1990). In addition, as the amount of DNA sequence data and STS (sequence tagged sites) data increases a large number of non-coding polymorphic markers are being identified, for example microsatellite markers (Abraham *et al.* 1993; Grimaldi *et al.* 1996; Hsieh *et al.* 1997). The presence of these polymorphic genes and markers

throughout the MHC has allowed close examination of recombination events occurring between them, in many diverse populations.

Analysis of recombination events within the MHC has led to the observation that particular alleles of different MHC genes tend to remain associated far more frequently in populations than expected by chance. This phenomenon is known as linkage disequilibrium and is particularly pronounced in the MHC. The result is a higher than expected occurrence of particular combinations of alleles, in some cases spread across the whole MHC. These combinations of alleles are called extended (or ancestral) haplotypes (AH). Some of these extended haplotypes are associated with susceptibilities to certain diseases (Tiwari & Terasaki 1985; Degli Esposti *et al.* 1992a; Price *et al.* 1999). The single most significant disease associated AH for the MHC is 8.1, which includes the HLA-A1, C7, B8, C4AQO, C4B1, DR3 and DQ2 alleles. The 8.1 AH is carried by most Caucasians and in one study was shown to be carried by 88% of Australians (Degli Esposti *et al.* 1992a), although this drops to about 57% in an Italian population. This haplotype is notable in its association with a wide range of diseases (Price *et al.* 1999), many of which can be described as immunopathological conditions. It has been associated with impaired survival after HIV infection (Keet *et al.* 1996) and has been shown to predispose carriers to IDDM (Hanifi Moghaddam *et al.* 1998), coeliac disease (CD) (Hall *et al.* 1996), systematic lupus erythematosus (SLE) (Christiansen *et al.* 1991), dermatitis herpetiformis (Wilson *et al.* 1995), common variable immunodeficiency (CVID) and IgA deficiency (IgAD) (Volanakis *et al.* 1992; Schroeder *et al.* 1998), MG (Degli Esposti *et al.* 1992b) and scleroderma (Kallenberg *et al.* 1981). It could also be argued though that, as the 8.1 ancestral haplotype is so common within the Caucasian population, it is not surprising that such a large number of disease associations have been observed and that it is variations

between individuals within the 8.1 extended haplotype that predispose to the conditions listed above. Associations have also been found between RA and DR1 and DR4, ankylosing spondylitis and the B27 allele (Baech *et al.* 1997), psoriasis and Cw6 (Jenisch *et al.* 1998), narcolepsy and DQB1*0602 and DR2 (Juji *et al.* 1984) and seronegative myasthenia gravis and DR9 (Chiu *et al.* 1987). Remarkably, there are as many as several hundred diverse diseases that can be shown to be associated with the MHC (Tiwari & Terasaki 1985; Dawkins *et al.* 1999). However, genetic mapping of disease genes in the MHC is hampered by the existence of linkage disequilibrium across the region. Consequently, in some diseases, the associated HLA alleles may be a marker for the true disease-associated gene lying in close proximity on the same haplotype, or may account only partially for the HLA association. For example, SLE is associated with the HLA-haplotype B8-DR3 (8.1 AH), but may be due to deficiency of the complement component C4 associated with the B8-C4AQ0-C4B1-DR3 haplotype (Arnett & Reveille 1992). Therefore, although allelic variants of the class II, and to a lesser extent class I, molecules account for some of the disease associations mentioned above, there is evidence to suggest that the class III region may also contain disease susceptibility genes (Degli Esposti *et al.* 1992b; Lechler 1994; Mizuki *et al.* 1995; Plevy *et al.* 1996; Hsieh *et al.* 1997; Lie *et al.* 1997; Hanifi Moghaddam *et al.* 1998; Schroeder *et al.* 1998; Singal *et al.* 1998).

1.1.4.1 Disease association within the MHC class III region

As mentioned above, due to linkage disequilibrium across the whole of the MHC, diseases so far associated with MHC class I or class II alleles may well be, at least partly, due to disease susceptibility genes within the MHC class III region. Indeed, for some of the genetically more straightforward conditions, susceptibility loci within the MHC class III region have already been determined. One such example is congenital adrenal hyperplasia

(CAH) which is caused by deficiency in the adrenal enzyme steroid 21-hydroxylase, encoded by the P450c21 gene (reviewed in Morel & Miller 1991). This deficiency is caused by deletion of the P450c21B gene or by mutations resulting in lack of P450c21B mRNA and leading to the accumulation of 17-hydroxyprogesterone which would otherwise be converted to 11-deoxycortisol by the P450c21. Duplication and deletion of the C4 genes, C4A and C4B, is relatively common, although complete deficiency of C4 is rare and correlates with severe immune complex disease. Partial deficiency of C4 is thought to be associated with increased susceptibility to systemic lupus erythematosus, scleroderma and primary biliary cirrhosis (Hauptmann *et al.* 1993). Other examples of the involvement of class III-encoded proteins in disease include observations of increased levels of hsp70 expression in autoimmune diseases such as scleroderma and SLE, upregulation of Casein kinase II activity in the neurofibrillary tangles in the brains of Alzheimer's Disease (AD) patients, and the implication of genetic aberrations in the sialidase-encoding NEU gene implicated in sialidosis (Carrillo *et al.* 1997; Milner *et al.* 1997; Pshezhetsky *et al.* 1997).

1.1.4.2 The significance of the G1 to BAT1 interval.

In addition to the conditions mentioned above, susceptibility loci for IDDM, IgA/CVID and MG have recently been pinpointed to the telomeric end of the MHC class III region, as described below.

In the case of IDDM it is clear that individuals heterozygous for HLA-DRB1*03, DRB1*04 are most susceptible to this disease (Price *et al.* 1999). There have also been suggestions of an independent association between IDDM and some HLA-DP alleles (Balducci Silano & Layrisse 1995), but this was not supported by Lie *et al.* (1997). In

addition, amino acid substitutions at positions 28 and 57 of HLA-DQA1 and HLA-DQB1, respectively, have been implicated in susceptibility to and protection from IDDM (Cucca & Todd 1996). However, Kelly *et al.*(1985) showed that the relative risk associated with HLA-DR3 was greatest if MHC class III region alleles characteristic of the 8.1 haplotype were also present. Furthermore, recombinant mapping identified two HLA-DRB1*04 patients who had 8.1 AH characteristics between TNF and HLA-B only (Moghaddam *et al.* 1997). This suggests that there may well be a disease susceptibility locus for IDDM within the telomeric end of the MHC class III region.

CVID and IgAD are thought to be variations of the same disorder, involving B-cell differentiation, and appear to have both genetic and environmental components. These conditions are most strongly associated with the 8.1 AH (Schaffer *et al.* 1989) and analysis of affected and unaffected members of families has allowed the mapping of susceptibility loci to the MHC class III region. Moreover, a recent publication by Schroeder *et al.*(1998) described a study of a large family, including eight members with CVID or IgAD. This led to the identification of crossover events indicating that susceptibility genes lie between the G1 gene at the telomeric end of the MHC class III region and HLA-A at the centromeric end of the class I region.

Recombination mapping of myasthenia gravis-affected families has strongly implicated the 8.1 AH and, more specifically, the region between HLA-B and TNF in a immuno-regulatory role (Degli Esposti *et al.* 1992b).

The rate of loss of CD4+ T-cells in HIV patients has been found to be associated with the 8.1 AH (Keet *et al.* 1996). More specifically, it has been suggested that the association of

the 8.1 AH with HIV is due to the TNF gene family at the telomeric end of the MHC class III region, as production of TNF is elevated in HIV patients and increased TNF production is associated with the 8.1 AH (Price *et al.* 1999). However, elevated TNF levels are a feature of many conditions, and may well be a symptom of HIV rather than a cause.

1.2 PROJECT RATIONALE AND BACKGROUND

It is clear from evidence in the literature, as highlighted in the previous sections, that the MHC is of significant biomedical, immunological and evolutionary importance. It is equally clear that although the class I and class II regions historically define the MHC the class III region also plays a significant role in the immune response. Therefore, no analysis of the MHC is complete without inclusion of the class III region. Of particular interest to us is the telomeric end of the MHC class III region. Preliminary evidence suggests a large proportion of the genes encoded therein are involved in the inflammatory response and, significantly, recombinant mapping data has suggested that a number of disease susceptibiliy loci fall within the region encompassing the TNF gene cluster. Although a large amount of work has already focussed on the cytokines TNF, LTA and LTB, there may be other candidate disease susceptibility genes closely linked to these. Therefore, it is essential to fully characterise this region, to identify all coding and regulatory sequences and to assess their contribution to disease susceptibility. This is the basis of the work described here.

1.2.1 Molecular mapping of the MHC class III region

The molecular characterisation of the class III region has so far involved the physical mapping of the region by pulsed field gel electrophoresis (PFGE) and the construction of a contig of genomic clones, comprising both cosmids (Sargent *et al.* 1989a; Spies *et al.*

1989a; Kendall *et al.* 1990; Fukagawa *et al.* 1995) and YACs (Abderrahim *et al.* 1994) (for review see Campbell 1993). This analysis established the location of the complement and TNF genes in the class III region and the physical distances between them (Carroll *et al.* 1987; Dunham *et al.* 1987). Most of the new genes subsequently located in the class III region were discovered by analysing the cloned genomic DNA using a variety of different techniques. This included the mapping of CpG islands with rare cutting restriction enzymes using PFGE (Kendall *et al.* 1990), zoo blot analysis (Sargent *et al.* 1989a), Northern blot analysis (Sargent *et al.* 1989a; Spies *et al.* 1989a), cDNA selection on immobilised cloned genomic DNA (Fan *et al.* 1993), exon trapping (Burfoot & Campbell 1994) and the direct screening of cDNA libraries using cosmid genomic DNA inserts (Spies *et al.* 1989b; Kendall *et al.* 1990). This latter technique has proved to be the most productive and in addition to isolating cDNA clones corresponding to the CpG-island associated genes, has also led to the isolation of cDNAs corresponding to genes that are more restricted in their pattern of expression (Aguado *et al.* 1996 and references therein).

1.2.2 Why sequence the MHC class III region ?

The above approaches have been responsible for the mapping of a large number of genes. However, not all genes can be detected using the above methods which is one of the main reasons for sequencing this region. DNA sequence analysis represents the finest resolution and the most informative approach to gene mapping when combined with gene prediction software and/or EST information. This has been aided by significant advances in DNA sequencing technology since the birth of the Human Genome Project in the 1980s. Dye-primer or dye-terminator labelling of sequencing template extension products with base-specific fluorescent dyes has meant that the 4 separate base-specific reactions can be combined and loaded into a single gel lane, thereby increasing the number of samples

which can be loaded per sequencing gel. The development, by Applied Biosystems, of the latest ABI 377 automated DNA sequencing hardware and supporting software has meant that the length of sequence read per clone has increased significantly from ~300 nucleotides with manual radioactive sequencing to greater than 800 nucleotides with present fluorescent chemistries, cycle sequencing protocols and automated technologies. Taken together, these advances have resulted in enormous improvements in sequencing turnover, read length, base calling accuracy and resolution. This means that the completion of sequencing projects, such as that described here, is becoming increasingly feasible within relatively short time scales and to an exceptionally high standard.

In addition to hardware and sequencing chemistry advances, there have been significant advances in gene prediction software and in the development of increasingly powerful similarity search engines. This has greatly enhanced the ability of computer programs to identify coding regions that could not be identified by previous methods, due to the transient or specific/inducible nature of their expression. These programs are especially useful when combined with the incredible resource generated by the human EST sequencing project as part of the Human Genome Project. The EST databases are expected to represent a large proportion of the transcripts expressed within the human genome. These, when combined with genomic sequence, allow not only the precise mapping of a transcript onto the genome but also the elucidation of gene structure and the localisation of intron/exon boundaries. This is particularily important when differentially spliced forms of a gene are expressed, as appears to be the case at the telomeric end of the MHC class III region. Furthermore, to understand the incredibly complex way in which genes are regulated it is essential that the untranscribed regions are also sequenced, to identify promoter and enhancer/suppresser regions. This is also relevant when looking for disease

susceptibility loci in that the aberrant expression of genes can be caused by mutations within their regulatory regions. In the case of the TNF gene a polymorphism identified within the promoter region (position –308) is associated with a number of conditions, including sensitivity to cerebral malaria (McGuire *et al.* 1994).

Genomic sequence analysis allows the identification of polymorphic markers such as microsatellites which are valuable in the fine mapping of disease susceptibility loci. In the case of Behçet's disease there was a known association with the HLA-B51 allele. Sequence analysis of the MHC class I region resulted in the identification of further microsatellite markers. This allowed the position of the susceptibility gene to be narrowed down to a 46 kb segment between MICA and MICB containing five genes, of which MICA is a strong candidate (Shiina *et al.* 1999).

1.3 AIMS

The aims of this project were:-

1. To add to the genomic sequence data already available for the MHC class III region, by sequencing the cloned genomic DNA covering the ~82kb segment of DNA between the G1 and BAT1 genes using the new fluorescent based sequencing technologies.

2. To analyse the DNA sequence generated, to precisely map the genomic structures of known transcripts and to identify any distinctive features within the sequence including regulatory and repetitive elements.

3. To identify and confirm the existence of any further potential coding regions, analyse their expression pattern at the RNA level and identify sequence features that will aid in the elucidation of the funtion of predicted protein products.

4. To analyse the genomic regions encompassing any new coding regions to identify any polymorphisms which might be useful in the identification of any candidate disease susceptibility loci.

Chapter 2

CHAPTER 2: MATERIALS AND METHODS

2.1 GENERAL SOLUTIONS, BUFFERS AND MEDIA

Ampicillin stock

20 mg/ml ampicillin (Sigma) in 1 M $NaHCO_3$ (filter sterilised)

Chloroform/Isopropanol

Chloroform and Isopropanol mixed at a ratio of 24:1 (v/v)

6% Denaturing polyacrylamide sequencing gel

6% (v/v) Acrylamide (Amresco) and 50% (w/v) Urea (de-ionised with mixed base resin and filtered) in 1x TBE, polymerised with 0.1% (v/v) N,N,N',N'-tetramethylene diamine (TEMED) and 1/200 volume 10% (w/v) ammonium persulphate (APS)

10x End fill buffer

0.5M Tris pH7.5, 10mM DTT, 0.1M NH_4SO_4, 0.05M $MgCl_2$, 1mg/ml BSA

Glucose TE

50mM Glucose, 10mM EDTA, 25mM Tris, pH 8.8

Glycerol loading dyes

30% (v/v) glycerol, 0.1% (w/v) bromophenol blue, 0.1% (w/v) xylene cyanol, 5 mM EDTA, pH 7.4

10x Hogness solution

40% (v/v) glycerol, 10 mM $MgSO_4$, 36 mM K_2HPO_4, 13 mM KH_2PO_4, 20 mM Tri-sodium citrate, pH 7.4

Hybridisation mix

50% (v/v) formamide, 1 M NaCl, 50 mM Tris-HCl, pH 7.4, 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

IPTG

100 mM IPTG in water (filter sterilise)

L broth/2x TY plates

L broth or 2x TY with 1.5% (w/v) Bacto agar (Difco)

L broth media

1% (w/v) Bacto tryptone (Difco), 1% (w/v) NaCl, 0.5% (w/v) yeast extract (Oxoid)

Lysis buffer

150 mM Glucose, 25 mM Tris, pH 8, 10 mM EDTA

5x M9 salts

0.45M Na_2HPO_4, 0.11M KH_2PO_4, 0.04M NaCl, 0.09M NH_4Cl

Filter sterilise

M9 Minimal Media agarose plates

1.8% Bacto agar (Difco)

Autoclave then add 25% (v/v) 5 x M9 salts (filter sterilised) and 0.25% (w/v) glucose (Filter sterilised)

10x MOPS buffer

0.2 M MOPS, pH 7, 10 mM EDTA, 50 mM sodium acetate

10x PCR reaction buffer

67 mM Tris-HCl, pH 8.8, 16.6 mM $(NH_4)_2SO_4$, 0.1% (v/v) Tween-20

PCR-SSP 10x base buffer.

670mM Tris-base pH8.9, 166mM ammonium sulphate, and 1% Tween 20.

PCR-SSP TDMH buffer.

16.9 ml 10x Base buffer, 0.4ml each dCTP, dGTP, dATP and dTTP from 100mM stock (Promega), 13.5ml 25mM $MgCl_2$, 33.4ml distilled water. Ingredients were mixed in a sterile vessel and stored in appropriate aliquots at -70˚C.

PCR-SSP DNA loading buffer.

300ml Glycerol, 250ml 2xTBE, 450ml dH_20, 0.25g Orange G disodium (Sigma)

Phenol/Chloroform

Equal volumes of Phenol and Chloroform/Isopropanol (24:1 (v/v))

Pre-wash solution

1 M NaCl, 1 mM EDTA, 50 mM Tris-HCl, pH 7.4, 0.1% (w/v) SDS

RF1 buffer

100 mM RbCl, 50 mM $MnCl_2$, 30 mM KOAc, 10 mM $CaCl_2$, 15% (v/v) glycerol, pH 5.8 (filter sterilised)

RF2 buffer

75 mM $CaCl_2$, 10 mM RbCl, 15% (v/v) glycerol, 10 mM MOPS, pH 6.8 (filter sterilised)

RNA loading buffer

50% (v/v) formamide, 5.9% (v/v) formaldehyde, 4% (v/v) glycerol, 0.005% (w/v) bromophenol blue in 1x MOPS buffer (qv)

Spermidine stock

25.46 mg/ml spermidine (Sigma) in 10 mM Tris, pH 7.4

20x SSC

3M NaCl, 0.3M NaCitrate

10x TBE

0.89 M Tris, 0.89 M Boric acid, 20 mM EDTA

TE

0.1 mM EDTA, 10 mM Tris, pH 7.4

Top Agar

1.5% (w/v) Nutrient Agar (Difco)

2x TY media

1.6% (w/v) Bacto tryptone (Difco), 0.5% (w/v) NaCl, 1% (w/v) Yeast Extract

X-Gal

2% (w/v) 5-bromo-4-chloro-3-indolyl β-galactopyranoside (X-Gal, Nova Biochem) in dimethylformamide.

2.2 BACTERIAL CELL CULTURE

2.2.1 Culturing Bacterial Cells

All solutions for bacterial culture were either steam-sterilised by autoclaving or filter-sterilised by passage through a 0.22 μm filter (Millipore). *Escherichia coli*, transformed with plasmids carrying the AmpR gene, were grown from single colonies or 10 μl of glycerol stock, in L broth or 2x TY media with ampicillin (100 μg/ml) at 37˚C, overnight (unless otherwise stated) in a shaking incubator (150-200 revolutions per minute). Colonies were selected on L broth plates supplemented with 50 μg/ml ampicillin.

2.2.2 Glycerol Stocks (storage of bacterial cultures)

50 μl of 10x Hogness solution was added to 450 μl of overnight culture, shaken vigorously and stored at -70˚C.

2.2.3 Preparation of Competent Bacterial Cells

Competent cells of the *E.coli* strain TG1 cells were prepared as follows. An overnight culture of TG1 cells, grown in 2x TY broth (10 ml), was diluted 100-fold in a 1 litre conical flask with 100 ml of 2x TY broth and cultured for 2-3 hours until the OD_{600} was ~0.5. The cells were harvested by centrifugation (1000 g; 15 minutes; 4˚C) and the pellet was gently re-suspended in 33 ml of RF1 buffer, incubated on ice for 15 minutes and centrifuged as above. The pellet was then gently re-suspended in 8 ml of RF2 buffer and incubated on ice for 15 minutes. Finally, the cell suspension was aliquoted into pre-chilled 1.5 ml tubes, frozen on dry ice and stored at -70˚C.

2.2.4 Transformation Procedures

2.2.4.1 Chemical transformation of plasmid DNA

Plasmid DNA (~10 ng or half the product of one 10µl ligation) was mixed with 0.5 ml of thawed competent TG1 *E.coli* cells (section 2.2.3), incubated on ice for 20 mins and then heat shocked at 37°C for 5 mins. 200 µl of L broth was added and the samples were incubated for a further 30 mins at 37°C. LA plates were prepared for Blue/White colony selection by spreading a mix of 30µl of IPTG and 40µl of X-Gal evenly over the surface of each plate then allowing the plates to dry. After incubation the transformed TG1 *E.coli* cells were spread onto the plates and incubate at 37°C overnight. Circular plasmid vector (10 ng) was used as a positive control for transformation, and non-transformed cells as a negative control.

2.2.4.2 Chemical transformation of M13mp18

M13mp18 DNA (~10 ng or half the product of one 10µl ligation) was mixed with 200 µl of thawed competent TG1 *E.coli* cells (section 2.2.3). This was then incubated on ice for 45 mins, heat shocked at 43° C for 2 mins then cooled on ice for a further 5 mins. The transformed cells were then added to 3ml of melted Top Agar equilibrated to 42° C together with 30µl of IPTG, 40µl of X-Gal and 1.5ml of an overnight culture of TG1 cells. The agar was poured evenly over a 2x TY plate, allowed to set, then incubated at 37° C over-night. Replicative form M13mp18 (10 ng) was used as a positive control for transformation and non-transformed cells as a negative control.

2.3 NUCLEIC ACID TECHNIQUES

2.3.1 Organic Extraction

Protein and other organic molecules were extracted from nucleic acid samples by the addition of an equivalent volume of phenol/chloroform to the aqueous solution. The

mixture was vortexed, centrifuged at 12000 g for 5 min and the upper aqueous layer was transferred to a fresh tube. The aqueous layer was extracted a second time with an equal volume of chloroform/isopropanol.

2.3.2 Recovery of DNA from Aqueous Solution

DNA was precipitated from aqueous solution by the addition of 2.5 volumes of ethanol and 0.1 volume of 3 M sodium acetate, pH 4.8. The solution was incubated at -20°C for 15 mins and centrifuged at 12000 g for 12 mins. The pellet was washed with 70% (v/v) ethanol and centrifuged at 12000 g for 5 mins. The pellet was air-dried and re-suspended in sterile distilled H_20 or TE.

2.3.3 Restriction Enzyme Digestion

With the exception of *Bam*HI (Amersham), all restriction enzymes were obtained from Promega. Restriction digests were carried out in buffers recommended by the supplier. Spermidine (5 mM final concentration) was added to digests performed in buffers containing more than 50 mM NaCl. Analytical digests were carried out in 10 µl reactions using ~500 ng DNA and 5 U enzyme for 2 hours at 37°C. The entire digest was then run on an analytical agarose gel (section 2.3.4.1) to resolve the restriction fragments. To prepare fragments for cloning, large scale digests of ~5 µg DNA were carried out in 50 µl reactions using 25 U enzyme and the fragments were resolved on a preparative agarose gel (section 2.3.4.2).

2.3.4 Agarose Gel Electrophoresis

2.3.4.1 DNA agarose gel electrophoresis

Electrophoresis of DNA samples was carried out in 0.8% (w/v) agarose gels (SeaKem HGT, Flowgen) unless otherwise stated. Gels were made in 1 x TBE and contained

0.001% (w/v) ethidium bromide. DNA samples were loaded in 0.1 volume of glycerol dyes unless otherwise stated. Electrophoresis was carried out at 80 mA in 1x TBE running buffer and DNA was visualised using an ultraviolet (UV) transilluminator. The size of DNA fragments was estimated by running the 1 kb ladder size standard (Gibco) in one lane of each gel.

2.3.4.2 Preparative DNA agarose gels

Preparative gels were run in exactly the same way as above but with 10mm wide sample loading wells instead of the 2mm wide wells used for an analytical gel.

2.3.4.3 RNA agarose gel electrophoresis

RNA was fractionated on a slab gel (20 x 20 x 0.5 cm) prepared by melting 1% (w/v) HGT agarose in 1x MOPS buffer and adding 37% (v/v) formaldehyde to the gel at 55˚C to give a final concentration of 1.9% (v/v) formaldehyde. 10µg of total RNA was re-suspended in 5µl RNase-free water. The samples were denatured by heating at 65˚C with 25µl RNA sample buffer for 5 min and chilled on ice for 2 min. Ethidium bromide was added to the samples, to a final concentration of 0.03% (w/v), before loading. Samples were electrophoresed overnight at 45 mA in 1x MOPS buffer and visualised on a transilluminator.

2.3.5 Recovery of DNA from Agarose Gels (Gel Purification)

DNA bands were excised from low melting temperature gels containing 1.5% to 3% (w/v) agarose (SeaKem GTG, Flowgen), using a clean razor blade. Exposure of the gel to UV light was kept to a minimum to avoid DNA damage. DNA was extracted from the agarose gel slices using either the QIAquick Gel Extraction Kit (QIAGEN) or the Prep-A-Gene DNA Purification Kit (BIO-RAD), following the supplier's instructions. DNA was eluted

in 15-50 µl of sterile distilled H_2O or 10 mM Tris, pH 7.4, according to the supplier's instructions.

2.3.6 Determination of DNA Concentration

The concentration of DNA samples was calculated by measuring the OD_{260} of a dilution of the sample in water. One absorbance unit is equivalent to 50 µg/ml of double-stranded DNA and 33 µg/ml of single stranded DNA, using a spectrophotometer with a path length of 1 cm.

2.3.7 Preparation of Plasmid or Cosmid DNA

2.3.7.1 Alkaline lysis method

A 10 ml overnight culture (section 2.2.1) was centrifuged (15 min; 1000 g; 4°C) and the supernatant discarded. The cells were re-suspended in 200 µl of lysis buffer and transferred to a 1.5 ml tube. 400 µl of fresh 1% (w/v) SDS / 0.2 M NaOH was added to the cell suspension, the sample was mixed by inversion and left to stand for 2-3 mins. The bacterial genomic DNA was precipitated by the addition of 300 µl of 3 M NaOAc (pH 4.8), followed by mixing then incubation at room temperature for 2-3 mins. Cell debris and genomic DNA were pelleted by centrifugation (12000 g; 5 mins) and the supernatant was removed to a fresh 1.5 ml tube. This step was repeated once more to ensure complete removal of genomic DNA. To precipitate the vector DNA 600 µl of isopropanol was added to the supernatant, mixed by inversion and placed at -20°C for 10 mins. The plasmid DNA was recovered by centrifugation at 12000 g for 15 mins and the pellet was re-suspended in 200 µl of 0.3 M NaOAc (pH 7). An equal volume of phenol/chloroform was added, and the sample was vortexed and centrifuged (12000 g; 5 mins; room temperature) for 5 minutes. The aqueous layer was transferred to a fresh 1.5ml tube and was extracted a second time with an equal volume of chloroform/isopropanol. The

aqueous layer was then removed to another 1.5 ml tube and the DNA was precipitated with 200 µl of isopropanol and recovered as above. The pellet was washed in 70% (v/v) ethanol, air dried and re-suspended in 25-50 µl TE. RNA was removed by incubation (30 min; 37°C) with heat inactivated RNase A (Sigma Type III-A) at a final concentration of 20 µg/ml. The DNA was stored at -20°C.

2.3.7.2 Wizard miniprep DNA purification (Plasmid only)

The Wizard Plus SV minipreps DNA Purification System (Promega) was used as an alternative method to isolate plasmid DNA from 5 ml cultures, according to supplier's instructions.

2.3.8 DNA Cloning Techniques

2.3.8.1 Cloning into Plasmid vectors

2.3.8.1.1 Preparation of plasmid vectors

10 µg of plasmid DNA was digested to completion with the appropriate restriction enzyme (section 2.3.3). Digested DNA was then de-phosphorylated (section 2.3.8.1.2). The linearised vector DNA was electrophoresed on an agarose gel, excised and gel purified (section 2.3.5).

2.3.8.1.2 Dephosphorylation

Vector DNA that had been digested with restriction enzymes was de-phosphorylated in order to prevent re-ligation of the DNA. This was achieved using 1 U of shrimp alkaline phosphatase (Boehringer Mannheim) per 100 ng of plasmid DNA. Dephosphorylation was carried out in the supplier's buffer and incubated at 37°C for 10 mins. The enzyme was then inactivated by incubation at 65°C for 15 min.

2.3.8.1.3 Ligation

Ligations were performed in 10 µl volumes, at 16°C for 12-14 hours. The ligation mix contained, on average, a ratio of 1:3 of vector to insert DNA (5-25 ng vector) and 1 U of T4 ligase (Amersham), in the supplier's ligation buffer. As a control for the efficiency of de-phosphorylation, a ligation reaction was also set up without insert DNA.

2.3.8.1.4 Transformation

As in section 2.2.4.1.

2.3.8.1.5 Identification of recombinant clones

Both pBluescript and pGEM recombinant clones were identified by blue/white selection. Insertion of DNA into the multiple cloning site disrupts the *lac*Z gene within these vectors, which encodes β-galactosidase. Cells transformed with pBluescript were spread onto agar plates containing X-Gal and IPTG as described in section 2.2.4.1 to induce expression of *lac*Z. Colonies transformed with vector containing inserts stayed white, and were either screened by colony PCR (section 2.3.9.6) or by isolation of plasmid DNA (section 2.3.7) followed by restriction enzyme digestion (section 2.3.3). Vector without insert gave rise to blue colonies. It has been observed, however, that the expression of a functional *lac*Z gene product is not always disrupted by the ligation of an insert within the coding region. Therefore, when an unexpectedly a low number of white colonies was seen a small number of blue colonies were also picked for further analysis.

2.3.8.1.6 pGEM -T Easy vector system (Promega) for the cloning of PCR products

The pGEM -T Easy vector system (Promega) utilises the fact that most thermostable polymerase enzymes used during the *in vitro* amplification of DNA fragments (PCR – section 2.3.9) non-specifically add a single deoxyadenosine to the 3' end of the template

being amplified, thus creating an overhanging adenosine to both ends of the amplification product. The pGEM -T Easy vector is cut with *Eco*R V and 3' terminal thymidines are added to complement the overhanging adenosines of the amplification product. Therefore, PCR products can be directly ligated in to the pGEM -T Easy vector.

The pGEM -T Easy vector system was used according to supplier's instructions to clone and isolate the multiple products amplified during the RT-PCR analysis (section 2.3.10) of the 1C7 and ATP6G genes (Chapter 4). Positive clones were identified by colony PCR (section 2.3.9.6) in a 96 well format with template specific oligonucleotide primers (listed in Chapter 4).

2.3.8.2 Cloning into the M13mp18 bacteriophage vector

2.3.8.2.1 Preparation of M13mp18 vectors

10 μg of replicative form M13mp18 DNA (donated by Dr B. Aguado) was digested to completion with *Sma*I (section 2.3.3) and de-phosphorylated (as in section 2.3.8.1.2). The linearised vector DNA was electrophoresed on an agarose gel and gel purified (section 2.3.5)

2.3.8.2.2 End Filling

To produce blunt ended fragments for cloning into the *Sma*I-cut M13mp18 the DNA to be ligated was 'end filled'. The end fill reaction contained ~500 ng of DNA, 10U Klenow, 10U T4 DNA polymerase, 0.5 mM dNTP, 2.5 mM MgCl2, 1x End fill buffer. This was incubated at room temperature for 1 hour before organic extraction (section 2.3.1) and DNA recovery (section 2.3.2).

2.3.8.2.3 Ligation

Ligations were performed using 5-20 ng of cut and phosphatased vector and 5-50 ng of prepared insert DNA in 1x T4 DNA ligase buffer (Promega) with 0.5-1U T4 DNA ligase in a total volume of 10µl, at room temperature for ~5 hours. *Alu*I-cut DNA was ligated into the vector as a positive control for the ligation reaction and cut/de-phosphorylated M13mp18 DNA was used as a negative control.

2.3.8.2.4 Transformation

As in section 2.2.4.2

2.3.8.2.5 Identification of recombinant clones

Incubation of 2x TY plates overlaid with the top agar containing TG1 cells resulted in the growth of a 'bacterial lawn'. TG1 cells transformed with the M13mp18 vector have a greatly inhibited growth rate since the cellular machinery is being used to reproduce the Phage DNA. This gives rise to 'plaques', i.e. areas of what looks like no growth in the bacterial lawn. Clear/white plaques were selected as containing recombinant M13mp18 clones, and single stranded DNA was synthesised as described in section 2.3.8.2.6. Uncut replicative form M13mp18 DNA was used as a positive control for plaque formation and the expression of the *Lac*Z gene (blue plaques). M13mp18 DNA ligated with *Alu*I fragments acted as a positive control for white plaque formation.

2.3.8.2.6 Preparation of single stranded DNA from M13mp18

White recombinant plaques were each picked into 1.25ml of 2xTY in a 10ml sterilin tube containing a 1/100 dilution of an overnight culture of TG1 cells and incubated in an orbital shaker at 37°C for 6 hours. The resulting cultures were transferred to 1.5ml microfuge tubes, spun down at 12000g for 10 minutes and 800µl of the resulting supernatant

transferred to individual wells of a 96 deep-well plate and stored at -20°C. The phage DNA was purified from the supernatants using a Vistra DNA Labstation into a final volume of 50µl. The labstation combines the techniques of Polyethylene Glycol (PEG) phage precipitation and ethanol DNA precipitation with a novel rapid separation technology (Alderton *et al.* 1992; Beck & Alderton 1993) using the power of magnetism rather than centrifugation. Small, dispersed, superparamagnetic FMP beads act as a solid phase around which the precipitate can form. On application of a magnetic field, the particles are retained on the side of the tube along with the precipitate. This permits the removal of phage supernatant. Subsequent phage lysis and supernatant removal can be performed rapidly by a robot. The single stranded DNA recovered was washed then re-suspended in 50 µl of H_2O for: sequencing, further analysis, or long term storage at -20°C.

2.3.9 The Polymerase Chain Reaction (PCR)

2.3.9.1 Oligonucleotide primer design

Oligonucleotide primers were selected using the Primer software (Stephen & Lander 1991). The template specific primers chosen were all between 17 and 24 nucleotides in length with melting temperatures predicted to be between 58°C and 62°C. Primers were only chosen if they had at least two G or C nucleotides at their 3' end (GC clamp) and if they did not show any significant similarity to repeat sequences or other sequences within the DNA databases. Primer pairs were chosen to have similar melting temperatures.

2.3.9.2 Preparation of oligonucleotide primers

Non-fluorescent oligonucleotides were generated in-house by Mrs V Cooper (Oxford Centre for Molecular Sciences) or by PE Applied Biosystems, UK. Those generated in-house were de-protected prior to use by incubation at 55°C overnight and then ethanol precipitated. The DNA concentration was determined and the primer was diluted to a

suitable concentration in TE. Fluorescently labelled oligonucleotides were obtained from Oswel DNA Service. The primers obtained from PE Applied Biosystems and Oswel were de-protected and had the DNA concentration determined as part of the service.

2.3.9.3 General amplification conditions for cosmid and plasmid DNA

Amplification products from cosmid or plasmid DNA were generated using ~5ng of template DNA in 1x supplier's PCR reaction buffer together with 0.2μM final concentration of primers and 0.2mM of each of the four dNTPs: dATP, dTTP, dCTP and dGTP. PCR reactions were carried out on either a Cetus DNA Thermal Cycler (Perkin Elmer) or a Hybaid Omnigene thermocycler using 1U AmpliTaq Gold DNA polymerase (Perkin Elmer) or Taq polymerase (Sigma) following the supplier's guidelines, and using conditions optimised for each primer pair and the particular thermocycler being used. General cycling conditions were: 94°C for 5 mins followed by 25-30 cycles of 94°C for 1 min, 56-64°C for 1 min and 72°C for 1.5 mins, this was followed by a 10 minute incubation at 72°C. Primer sequences and the amplification conditions for particular primer pairs are given in the relevant sections. After cycling, 5μl of each PCR reaction mix was run on an analytical agarose gel (section 2.3.4.1) to visualise the PCR product.

In cases where a single specific band could not be achieved by optimisation of the PCR conditions, or by modification of the methodology as outlined below, the correct band was gel extracted. The entire PCR reaction was split between 2-3 wells of a preparative agarose gel (section 2.3.4.2) and the band of the expected size was excised from the agarose and the DNA recovered (section 2.3.5). 1-2μl of the recovered DNA band was then used as a template in a PCR reaction with the same conditions as in the initial PCR reaction or using nested primers (as described in section 2.3.9.5).

2.3.9.4 Modifications to the general PCR cycling parameters

2.3.9.4.1 Hot start PCR

In some cases, low molecular weight secondary products were observed, caused by primer dimerisation during the initial stages of the PCR reaction. To minimise this, hot start PCR was carried out: reactions were incubated at 85°C for 5 minutes and Taq polymerase (in 5µl of 1 x PCR reaction buffer) was then added, before entering the cycling conditions described above.

2.3.9.4.2 Touchdown PCR

In cases where non-specific amplification products were observed that could not be attributed to primer dimerisation 'Touchdown PCR' was used to try and increase primer specificity. Touchdown PCR involves the initial cycles of the PCR reaction having an annealing temperature between 8°C and 10°C above the primer melting temperature. Then, over successive cycles the annealing temperature is reduced, 1°C at a time, until the predicted optimum annealing temperature is reached. This annealing temperature is then maintained for the final 20-25 cycles. In this way there is more chance of small amounts of the specific product being synthesised within the initial cycles. Therefore, by the time the predicted optimum annealing temperature is reached the desired product is at sufficient concentration to dominate the template pool, thus increasing the overall specificity of the reaction.

2.3.9.5 Amplification of genomic DNA for direct sequencing

The amplification of regions of genomic DNA for direct sequencing was achieved using a nested PCR approach. This involves two successive rounds of PCR, with the second round using the first round product as a template, but with amplification primers that fall inside the primer sequence of the first round (see Chapter 5). The first round of amplification

was carried out as described in section 2.3.9.3 using 0.2 μM primers and 0.2 mM dNTPs with 100ng of genomic DNA as template in 25ml final volume, annealing at 58°C (60°C for PCR 1C7 1st round-2) for 30 cycles (94°C 1 min, 58°C 2 min 30 sec, 72°C 1 min). The second round reactions contained 5 μl of a 1:20 dilution of the first round PCR product in a 25μl final volume with 0.07 mM dNTPs and 0.06 μM forward and reverse M13-tailed nested primers, annealing at 58°C for 28 cycles (94°C 1 min, 58°C 30 sec, 72°C 1 min). The concentration of dNTPs and primers were reduced in the second round so that PCR reagents were not carried over into the sequencing reaction. Hot start PCR was carried out for the second round of amplification to reduce primer dimerisation (section 2.3.9.4.1). The second round PCR products were sequenced directly with the dye-terminator sequencing chemistry (section 2.3.12.2) using 3μl of the PCR reaction mix with either the forward or reverse M13 sequencing primers (forward, M13-21-5'TGTAAAACGACGGCC AGT3', reverse, M13-reverse-5'GGAAACAGCTATGACCATG3').

2.3.9.6 Amplification of plasmid DNA from E. coli cells (Colony PCR)

To screen the large number of colonies produced during the cloning of the alternately spliced transcripts of 1C7 and ATP6G (section 2.3.8.1.6) colony PCR was undertaken, in a 96 well format. Firstly, 150μl of L broth/amplicilin (100μg/ml) was added to each of the wells of a sterile 96 well tissue culture plate (FALCON MICROTEST (96), and 15μl of H_2O was added to each of the wells of a 96 well PCR plate (Hybaid). Secondly, a white colony was picked for each of the 96 wells and used to inoculate the H_2O. Then, 5μl of inoculated water was added to the media in the corresponding well of the tissue culture plate. Next, 10 μl of 2x PCR buffer containing, dNTPs (final concentration 0.2 mM) and transcript specific primers (final concentration 0.5 μM) was added to the remaining 10 μl of inoculated water. This was over-laid with mineral oil and hot start PCR (section

2.3.9.4.1) was used to amplify the transcript with 0.5U Taq polymerase per reaction. An initial 98°C soak for 10 minutes was added to the PCR reaction to ensure lysis of the cells. The inoculated media was incubated overnight at 37°C then 15µl of Hogness solution was added to each well. The 96 well culture plate was then stored at –70°C.

2.3.10 The Reverse Transcriptase Polymerase Chain Reaction (RT-PCR)

2.3.10.1 First strand cDNA synthesis

The expression of transcripts was investigated by RT-PCR using total RNA (kindly donated by Dr B. Aguado). The RNA used was from the cell lines: Raji (B cell), Jurkat 6 (T cell), Molt 4 (T cell), HL60 (monocyte), U937 (macrophage), HepG2 (hepatocyte), HeLa (epithelial), HT1080 (epithelial) and SW620 (adenocarcinoma). An aliquot of the total RNA (stored in 70% ethanol) was taken for each cell line and centrifuged at 12000 g for 5 mins. The pellet was air-dried and re-suspended in sterile dH_20. The RNA concentration was then determined. The first strand cDNA synthesis was performed in a final volume of 20µl with 1µg of total RNA using the Promega Reverse Transcription system (Promega UK), according to the supplier's protocol.

2.3.10.2 Template specific PCR

PCR primers were designed for each transcript to be amplified to give products containing more than one exon so that amplification products arising from genomic DNA contamination were easily discernible. 5 µ1 of the first strand cDNA synthesis reaction mix was used in a 25 µl hot start PCR reaction with 0.5µM of transcript specific primers and 0.5 U Taq polymerase (Sigma) and using the first strand synthesis reaction buffer as the PCR buffer. The specific reaction mixes and cycling conditions are given in the relevant chapters. Each transcript-specific RT-PCR reaction was performed in at least triplicate to allow for any variation between reactions. Control amplification reactions,

with primers derived from beta-actin, were carried out for each first round cDNA synthesis reaction as a positive control and for quantification (by eye).

2.3.10.3 Confirmation of identity of PCR product

The identities of PCR products generated during the RT-PCR reactions were confirmed either by direct dye-terminator sequencing (section 2.3.12.2) or, when multiple products were obtained, by southern blot hybridisation (section 2.3.11.1) with a transcript specific probe, or by cloning the RT-PCR products into the pGEM -T EasyVector System (Promega) (section 2.3.8.1.6) and sequencing the cloned DNA (section 2.3.12.2).

2.3.11 Hybridisation Techniques

2.3.11.1 Transfer of DNA from agarose gels onto nitro-cellulose membranes (Southern blotting; Southern, 1975)

DNA was transferred onto membranes using a capillary blotting procedure. Agarose gels containing fractionated DNA were photographed with a ruler to allow measurement of band positions and then gently washed at room temperature in the following solutions for 30 min each:

i) 0.25 M HCl, to depurinate the DNA;

ii) 0.5 M NaOH; 1 M NaCl, to denature the DNA;

iii) 0.5 M Tris-HCl, pH 7.4; 3 M NaCl, to neutralise the gel.

The blots were set up as follows. The gel was placed onto three sheets of Whatman 3MM filter paper on 1cm of sponges soaked in 20x SSC. A nitro-cellulose membrane (Hybond-C extra) was pre-wetted in 2x SSC and carefully placed on top of the gel, avoiding air bubbles. Three sheets of pre-wetted Whatman 3MM paper were placed on top of the membrane, followed by a stack of paper towels and a weight. The transfer was allowed to

proceed by capillary action for 4 to 14 hours depending on the quantity of the DNA in the band to be transferred. The DNA was fixed to the nitro-cellulose membrane by baking at 80°C for 2 hours.

2.3.11.2 Transfer of RNA from agarose gels onto nitro-cellulose (Northern blotting; Fourney *et al.* 1988)

RNA gels were photographed with a ruler to one side and then washed once in 0.5 M NaOH; 1x SSC and twice in 10x SSC at room temperature for 20 min each. The RNA was transferred onto nitro-cellulose (Hybond-C extra) membrane, pre-wetted in 10x SSC, as described in section 2.3.11.1 using 10x SSC as the transfer buffer. RNA was fixed to the membrane by baking at 80°C for 2 hours.

2.3.11.3 Transfer of bacterial DNA onto nitro-cellulose membranes

Bacterial colonies were plated onto Hybond-C nitro-cellulose filters on the surface of LA plates and allowed to grow overnight at 37°C. The filters were carefully removed from the agar plates and placed colony side up onto Whatman 3MM paper. A second nitro-cellulose filter, pre-wetted on a fresh agar plate, was carefully placed onto the bacterial colonies. A second sheet of Whatman 3MM paper was placed on top, and sandwiched between glass plates to aid the transfer of colonies onto the second membrane. Asymmetric marks were made by insertion of a needle soaked in ink through both filters simultaneously. The replicas were separated and replaced onto agar plates, colony side up. The master plate was allowed to recover for 1 hour, prior to a second replica being made as described above (utilising the same orientation marks). Replicas were allowed to recover for 2 hours at 37°C, before lysing the colonies by placing the membranes colony side up onto 3MM Whatman paper soaked in the following solutions: 10% (w/v) SDS, 0.5 M NaOH and 1 M Tris-HCl pH 7.4, for 3-5 min each. The filters were washed in 2x SSC for

20 min with gentle agitation and the DNA was fixed to the membrane as in section 2.3.11.1.

2.3.11.4 Radiolabelling of DNA (Feinberg and Vogelstein, 1984)

DNA probes were labelled using the random hexanucleotide priming method. 20ng-50ng of DNA were labelled for 1 hour at 37°C, using the Multiprime Labelling System (Amersham) with 10-50μCi α-[32P]dCTP. Unincorporated nucleotides were removed by desalting the labelled probe through a G-50 Sephadex™ nick column (Pharmacia) using TE, pH 7.4 according to the supplier's instructions. Typical specific activities obtained were 10^7-10^9 cpm/μg DNA.

2.3.11.5 Hybridisation of Radiolabelled DNA Probes

Nitro-cellulose filters were carefully laid onto nylon meshes, pre-wetted in prewash buffer. These were then rolled and inserted into Hybaid bottles and prewashed with 50 ml prewash buffer for 20-40 min at 42°C in a rotating hybridisation oven (Hybaid). 20ml of hybridisation buffer containing 100μg/ml sonicated salmon sperm DNA (freshly added) was boiled for 5 min and then cooled to 42°C. Half the hybridisation buffer (10ml) was added to the washed filters within the Hybaid bottles and returned to the oven for 1 hour at 42°C to 'pre-hybridise' the filters. Radiolabelled probes were boiled in the remaining half of the hybridisation buffer ($\geq 10^5$ cpm/ml for cloned DNA or PCR products, or $\geq 10^6$ cpm/ml for genomic DNA or RNA blots) and allowed to cool to 42°C before use. This was then added to the Hybaid bottles and the filters were hybridised for 12-24 hours at 42°C in a rotating oven.

2.3.11.6 Washing of membranes and autoradiography

Following hybridisation, filters were washed in 2x SSC/0.1% (w/v) SDS for 20 min at 42˚C followed by two high stringency washes in 0.2x SSC/0.1% (w/v) SDS for 30 min each at 65˚C. Filters were air-dried, wrapped in Saran Wrap and exposed to sensitised X-ray film between two tungstate intensifying screens at -70˚C for the required length of time. Films were developed using a Kodak ME-1 automatic processor.

2.3.12 DNA Sequencing

DNA sequencing was carried out using the Sanger dideoxy method (Sanger *et al.* 1977), with either dye-primer (2.3.12.1) or dye-terminator (2.3.12.2) fluorescence based chemistry. The sequenced samples were loaded onto a 6% (w/v) denaturing polyacrylamide gel in an ABI 377™ Sequencer (PE Applied Biosystems) and sequences were determined using the ABI Prism™ DNA sequencing software (version 2.1.2, PE Applied Biosystems).

2.3.12.1 Dye-primer sequencing used for M13 clones and M13-tailed PCR sequencing

Dye-primer sequencing was carried out using the Thermo Sequenase labelled primer cycle sequencing kit (Amersham). Dye-primer sequencing involved four separate reactions per DNA sample. Each of the four supplier's reagents, A, C, G and T, included a chain terminating nucleotide, either ddATP, ddCTP, ddGTP or ddTTP, respectively; thermostable pyrophosphatase; and Thermo Sequenase DNA polymerase. The A and C reactions were set up using ~1 μg of DNA; 1.6 μl of the A or C reagent; and 1.2 μM primer labelled with the fluorophores JOE and FAM, respectively, in a total reaction volume of 6.4 μl. The G and T reactions were set up using ~2 μg of DNA; 3.2 μl of the G or T reagent; and 1.2 μM primer labelled with the fluorophores TAMRA and ROX, respectively, in a total reaction volume of 12.8 μl. The labelled primers used were the

forward sequencing primer, M13-21 (5' TGTAAAACGACGGCCAGT 3') and the reverse sequencing primer, M13-reverse (5' GGAAACAGCTATGACCATG 3'). The reaction mixes were denatured at 95°C for 5 mins before undergoing 20 cycles of 95°C for 15 seconds and 60°C for 30 seconds using an Omnigene thermocycler (Hybaid) in a 96 well format. The four sequencing reactions for each M13 clone were then pooled, precipitated in the presence of 2.5 volumes of 100% Ethanol (Sodium acetate was not needed to aid DNA precipitation as the reaction buffer already contains enough salt), pelleted by centrifugation at 12,000g for 10 minutes and supernatant removed by aspiration. The DNA pellet was washed with 70% (v/v) ethanol, pelleted again, the supernatant aspirated off and the pellet re-suspended in 3µl of formamide loading buffer. One half of each sequencing reaction was run on a 0.2mm thick denaturing 6% (w/v) acrylamide gel between 48cm glass plates, for 10 hours, using an Applied Biosystems 377 Automated DNA sequencer. The sequence data was collected and analysed with ABI 377 dedicated software.

2.3.12.2 Dye-terminator sequencing used for M13 clones, plasmid clones, and the direct sequencing of PCR products.

Dye-terminator sequencing was carried out using Thermo Sequenase dye-terminator cycle sequencing pre-mix kit, version 2.0 (Amersham). One reaction was set up per DNA sample using 0.1-2 µg of DNA, 8 µl of the supplier's termination mix and 0.32 µM sequence specific primer in a 20 µl reaction. The sequencing mix was denatured at 95°C for 5 mins before undergoing 25 cycles of 96°C for 30 seconds, 50°C for 15 seconds and 60°C for 4 mins using a Cetus DNA Thermal Cycler (Perkin Elmer). The sequencing products were ethanol precipitated in the presence of 2.5 volumes of 100% Ethanol and 0.1 volume 3M sodium acetate, stored at –20° C for 15 minutes, then pelleted by centrifugation

at 12,000g for 10 minutes and supernatant removed by aspiration. The DNA pellet was vortexed with 70% (v/v) ethanol, pelleted again, the supernatant aspirated and the pellet re-suspended in 3µl of formamide loading buffer.

2.3.13 Cosmid Genomic Insert Sequencing Strategy

To determine the nucleotide sequence of the cosmid genomic inserts an M13 shotgun strategy was adopted (Bankier *et al.* 1987) following the sequencing methodology described above (and in more detail in Chapter 3 section 3.1.1).

2.3.13.1 Preparation of an M13mp18 Shotgun Library

Cosmid clones (5-10µg) were randomly sonicated to yield a majority of fragments of between 0.5 and 1kb. The sonicated DNA was end-filled with a mixture of T4 DNA polymerase and Klenow enzyme (section 2.3.8.2.2), then size selected on a 0.8% preparative agarose gel for fragments of between 0.5 and 1kb. The appropriate size range was excised from the agarose gel and the DNA purified with the aid of the QIAquick gel extraction Kit (Qiagen) following the supplier's protocol. The DNA was ligated into M13mp18 phage vector (section 2.3.8.2.3). One half of the ligation mixture (5µl) was used to transform 200µl of previously prepared competent TG1 cells, plated out unto 2x TY plates (section 2.2.4.2), and incubated overnight at 37°C. Well separated white recombinant plaques were picked and single stranded DNA was prepared as in section (2.3.8.2.6).

2.3.13.2 Automated Sequencing

M13mp18 shotgun clones were sequenced with dye-primer sequencing chemistries (section 2.3.12.1). 4µl of the purified single stranded M13mp18 phage DNA was added to each of the A and C reaction mixes and 8µl was added to each of the G and T reaction

mixes. Sequencing was carried out in 96 well format (see section 2.3.12.1) using the M13 universal primer (5' TGACCGGCAGCAAAATG 3') labelled at its 5' end with one of four fluorescent dye labels (JOE, FAM, TAMRA and ROX).

Individual sequence trace files generated by the ABI377 collection and analysis software were processed with the 'PREGAP' Staden suite of software on a SUN SPARK station. This adds extra information to the ABI trace files, such as base calling accuracy, as well as masking vector sequences, Alu repeats and bad sequence which would otherwise interfere with template assembly. The sequences were then assembled with the Staden assembly software 'GAP v4.0-beta4'.

2.3.13.3 Finishing and Troubleshooting

2.3.13.3.1 Sequence ambiguities

Ambiguities within the sequence such as stops or compressions were resolved by the use of dye-terminator sequencing chemistry (section 2.3.12.2). Stops and compressions are typically cause by sequence specific DNA secondary structures, particularly in G-C rich regions. The use of the alternative chemistry does not generally remove the problems but shunts the ambiguity to neighbouring sequences. Therefore, with a combination of both dye-primer and dye-terminator sequencing chemistries, plus a depth of read averaging at least 4 to 5 templates per base sequenced, and sequences in both orientations, accurate sequence can be deduced.

To resolve sequence specific problems highlighted above, suitable clones from within the region of ambiguity were chosen, then the remaining half of the purified phage DNA

corresponding to each of these clones was precipitated and re-suspended in 10µl of dH_2O. Dye-terminator sequencing was carried out as described in section 2.3.12.2.

2.3.13.3.2 Areas of single orientation read

Areas of single orientation read were resolved by either dye-terminator sequencing chemistry, following the same protocol as for resolving ambiguities (above), or by generating a double stranded template then sequencing the reverse strand. For this second approach the inserts of the chosen M13 clones were PCR amplified under the following reaction conditions: 1µl M13 clone DNA; 0.4µM M13 PCR primers: Sense 5' AGCGGAT AACAATTTCACACAGGA 3'; Antisense 5' CGCCAGGGTTTTCCCAGTCACGAC 3'; 2U of AmpliTaq Gold DNA polymerase (Perkin Elmer), following the supplier's recommended conditions, in 50µl final volume. Reactions were cycled on a Perkin Elmer Thermocycler: 95°C for 2 min; followed by 25 cycles of 95°C for 30 sec, 60°C for 30 sec, 68°C for 2 min; ending in a 68°C soak for 5 min. 2µl of PCR product was used directly as a double stranded sequencing template with the M13 reverse primer (5' GGAAACAGCTA TGACCATG 3') using either dye-primer or dye-terminator sequencing chemistries.

2.3.13.3.3 Gaps between contigs

Gaps between contigs were closed by one of three methods:

1. Where longer reads are attainable, templates that extended into the gap were re-sequenced.

2. Where templates extend into the gap but longer reads are not feasible, a double stranded template was generated and the reverse strand was sequenced (following the above method for resolving single orientation regions), thus sequencing the template from the other end.

3. Where neither of the above are possible, gaps were closed by PCR. Oligonucleotide primers were chosen from accurate and non-repetitive sequence on either side of a gap with the aid of the Primer software (Stephen & Lander 1991) (primer sequences and amplification conditions are listed in Chapter 3, section 3.2.1.1). The sequence spanning the gap was PCR amplified from 5ng of cosmid cloned DNA with 0.2μM final concentration of primers. The PCR reactions were carried out on either a Perkin Elmer Thermocycler or a Hybaid Omnigene thermocycler using 2U AmpliTaq Gold DNA polymerase (Perkin Elmer) following the supplier's guidelines and using optimised conditions for each primer pair. Then either the PCR products were sequenced directly with dye-terminator fluorescent sequencing chemistries or end-filled with T4 DNA polymerase and Klenow enzyme (section 2.3.8.2.2), phosphorylated with T4 DNA Kinase, ligated into M13mp18 (section 2.3.8.2.3) and single stranded DNA made (section 2.3.8.2.6) and sequenced with fluorescent dye-primer sequencing chemistries. The latter approach was necessary when the template to be sequenced contained long tracks of adenosine or thymidine nucleotides, as found at the border of some Alu repeat sequences, as we were unable to sequence through these regions with double stranded templates

2.3.14 Genomic Sequencing Approach To Mutation Analysis

The genomic regions around the AIF-1, 1C7 and ATP6G genes and the BAT1 promoter region were scanned for mutations by direct sequencing (Chapter 5) of 'control' DNA from HLA-heterozygous individuals. In brief, overlapping M13-tailed templates covering the genomic region were generated by nested PCR (section 2.3.9.5), using the primers described in Chapter 5 and given in Appendix IIIA. The PCR products were diluted 1:10 and 2-6 μl of the diluted PCR products were used in dye-terminator sequencing reactions (section 2.3.12.2). Sequence traces were then processed with the Staden suite of software

and assembled using the 'GAP v4.0-beta4' program, as in section 2.3.13.2. Loading the traces into the GAP software made heterozygote and homozygote mutation detection relatively easy. The software highlights ambiguities between overlapping sequences or overlapping peaks within a sequence trace. The sequence traces covering each ambiguity were then examined by eye for heterozygous mutations which appear as two overlapping peaks of approximately half the neighbouring peak heights. Sequence traces overlapping the regions under investigation, which were generated during the original sequencing of the cosmids, were also added to the GAP4 database so that further comparison could be made. The templates in which mutations were identified were re-amplified and sequenced a second time to confirm that mutations were genuine and not amplification errors .

2.3.15 PCR-SSP genotyping of polymorphisms

A PCR-SSP (sequence specific primer-PCR) assay (Bunce *et al.* 1995) was used to genotype selected polymorphisms (see section 5.3).

2.3.15.1 Primer design and preparation

2.3.15.1.1 Primer design

A consensus primer was designed to anneal to both alleles of the chosen polymorphism, at a position between 200 and 800 nucleotides from the polymorphic site. Two allele specific primers were then designed with the terminal 3' nucleotide of each primer corresponding to one of the 2 allelic variants. For each sample genotyped, two PCR reactions were set up, one for the primer with the common allele as its terminal 3' nucleotide and the other for the primer with the rare allele as its terminal 3' nucleotide. Both reactions included the common consensus primer. To control for PCR failure, which may give rise to false negative results, a pair of control primers, '63' and '64' (Appendix IIIB), were included in each reaction. The control primers amplify a 796 bp fragment from the human DRB1

gene. The PCR conditions used (sections 2.3.15.2 to 2.3.15.3) are such that each primer set only amplifies DNA corresponding to the allele represented by the 3' nucleotide of the allele specific primer.

Details of primers used are given in section 5.3 and Appendix IIIB. The PCR-SSP primers were designed as outlined below and in section 5.3.1.3 to have T_ms of 58-62°C, using the equation: $T_m = [(A+T)x2]+[(G+T)x4]$.

Primers were manufactured on a 0.2μM scale by Cruachem Ltd (Glasgow, Scotland).

2.3.15.1.2 Primer preparation

Primers were stored at -20°C at a concentration of 2mg/ml prior to their use in primer mixes. Sequence specific primers were typically used at final concentrations of between 0.002 and 0.04μg/μl, with the optimum concentration being determined by titration in each case. Primers were tested in the following manner. First, the DRB1 control primers (listed in Appendix IIIB) were empirically titrated. A dilution of control primers giving good amplification over a wide range of DNA samples, but not out-competing the generation of allele specific products, was used. The final concentration of control primers was typically 0.0024μg/μl. A control primer stock solution (without allele-specific primers), also containing Cresol red (final concentration 60 μg/ml) to aid dispensing and visualisation of small amounts of primer mixes, was made up in 50ml aliquots and frozen at -30°C. Allele-specific primers were subsequently titrated in the control primer stock as described below.

Initially, allele specific primer pairs were tested at concentrations of 0.005μg/μl, 0.01μg/μl and 0.02μg/μl. On occasions these titrations were not effective and, in these cases, primers were tested by asymmetric titration. Asymmetric titration involves keeping the

concentration of the sense primer constant whilst varying the concentration of the antisense primer, and vice versa. Asymmetric titrations were especially useful for reducing the presence of false positive bands of incorrect size relative to the expected amplicon.

Once an optimum concentration of each primer pair was ascertained, working concentration primer mixes were made up, containing the control primer stock (above), together with the allele specific primer pairs. Primer mixes were dispensed in 5µl amounts into 96 well PCR plates (Costar). The plates were pre-dispensed with 10µl of mineral oil (Sigma) to prevent evaporation of stored mixes. Once the primer mixes were dispensed the plates were sealed in polythene bags and stored at -70˚C until required.

2.3.15. 2 PCR-SSP set up.

All PCR-SSP adhered to the "TDMH method" as described by Bunce *et al.* (1995 & 1996). Each PCR reaction contained the ingredients in the proportions PM:BM = 1:1 where PM is the primer mix described above and BM is the PCR buffer mix. The PCR buffer mix contains the TDMH PCR buffer (section 2.1), Taq polymerase and genomic DNA. For each row of a 96 well plate (12 wells) a PCR buffer mix was made up, containing 118µl TDMH, 0.6µl 5unit/ml Taq polymerase (Advanced Biotechnology) and 2µl target DNA at a concentration between 0.2-0.5µg/µl. Then, 5µl of the TDMH/DNA/Taq mixture was added to each well of a thawed 96 well plate containing the primer mixes and 10µl of mineral oil, as described above (section 2.3.15.1.2), using an electronic dispensing pipette. The final concentration of ingredients (per 10µl PCR reaction) were: 67mM Tris Base pH8.8, 16.6mM ammonium sulphate, 2mM magnesium chloride, 0.01% (v/v) Tween 20, 200µM of each dNTP, between 0.1 and 0.01µg DNA and 0.15 units Taq Polymerase (Advanced Biotechnology, London). After adding the PCR mixture, the 96 well plates were sealed with a self-adhesive plastic sealer (Costar). The plates were briefly

centrifuged to ensure that all the PCR mixture was submerged under the mineral oil. The plates were dipped in paraffin oil and blotted onto tissues before insertion into the PCR machine blocks. This fine coating of oil ensured a better contact between the block and the vessel and hence ensured more reliable PCR. The concentration of the DNA was seldom measured as most PCR-SSP assays worked over a wide range of DNA samples, even for reactions generating very large amplicons. However, if a DNA sample was insufficiently concentrated to amplify consistently in all PCR reactions, more DNA was added to the PCR mixture while adding less water, thus keeping the essential 1:1 ratio. If reaction failure was considered to be due to poor quality DNA, extra Taq polymerase was added.

2.3.15.3 PCR machines and thermal cycling parameters.

The following PCR machines were used for PCR-SSP: PTC-200 (96 or 192 well machine; MJ Research Ltd), PTC-100 96v (96 well machine; MJ Research Ltd) and GeneAmp 9600 (96 well machine Perkin Elmer). Identical thermal cycling parameters were used for all PCR-SSP amplifications (Table 2.1).

TABLE 2.1 PCR-SSP thermal cycling parameters

Number of cycles	Temperature (˚C)	Duration (seconds)
1	96	60
4	96	20
	70	45
	72	25
20	96	25
	65	50
	72	30
3	96	30
	55	60
	72	90
1	20	5

2.3.15.4 Electrophoresis and photography.

Following the completion of PCR reactions 6µl of gel loading buffer (section 2.1) was added to each well. The loading buffer was added by using a 1ml plastic syringe attached to a Hamilton repeating dispenser with an 8 lane stream splitter attached to enable multi-well dispensing.

Horizontal gel electrophoresis equipment (Flowgen and Anachem) was used for all gel electrophoresis. 0.5 to 1 cm thick 1% (w/v) agarose gels (made up in 0.5 x TBE were prepared utilising gel combs of the same spacing as multichannel pipettes to facilitate gel loading. The electrophoresis running buffer was 0.5 x TBE. The entire PCR reaction/loading buffer mix was loaded into a submerged horizontal gel using 8 or 12 lane multichannel pipettes. The samples were electrophoresed for 30 minutes at 200 v/cm, or 25 minutes at 250 v/cm, producing similar separation of PCR products.

2.4 BIOINFORMATICS

The Wisconsin Package Version 9-UNIX (GCG), maintained at the University of Oxford Molecular Biology Data Centre, was utilised for the majority of the sequence analysis and database interrogation. The DNA sequences generated were screened against the EMBL, SwissProt, PDB, EMBL-EST and TIGR-EST (search@hcd.tigr.org) databases using the FastA algorithm (Pearson & Lipman 1988) and the BLAST algorithm (Karlin & Altschul 1990). In this way, known genes were positioned. Intron/exon boundaries were identified by alignment with the cDNA database entries using the DOTTER dot-matrix software. Possible new coding regions were identified by alignments with expressed tagged sites (ESTs) or by alignments of sequences from intergenic regions with the sequences of known genes where there was significant sequence similarity suggestive of homology.

Repetitive elements were identified with the aid of the RepeatMasker server (Smit, AFA.,& Green, P. RepeatMasker at http://ftp.genome.washington.edu/RH/RepeatMasker.html). Potential coding regions and other features were identified using the GRAIL exon prediction program (Uberbacher & Mural 1991) and the NIX program (http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/) from the HGMP Resource Centre, Hinxton, UK. Predictions of protein secondary structure, solvent accessibility and transmembrane regions were conducted using the 'JPred'-consensus secondary structure prediction server (http://circinus.ebi.ac.uk:8081/), the 'PredictProtein' server (phd@EMBL -Heidelberg.de), and transmembrane region predictions at the TMpred server (http://www .isrec.isb-sib.ch). The GCG program 'SIGCLEAVE' and the 'SMART' (Simple Modular Architecture Research Tool) server (http://coot.embl-heidelberg.de/SMART/) were used to identify leader peptides. Sequence motifs and protein domains were identified using a combination of the GCG program 'MOTIF', the Prosite 'Profilescan' server (http://ulrec3.unil.ch/software/PFSCANform.html), the 'SMART' server, the 'PSITE' server (http://dot.imgen.bcm.tmc.edu:9331/psspredict ion/pssp.html) and the protein localisation prediction server PSORT (http://psort.nibb.ac.jp:8800/). Multiple alignments of amino acid sequences were performed using the Clustalx software, making use of protein structure information from sequences within the PDB database wherever possible. Alignments were hand-edited using the GCG9 'SeqLab' multiple alignment editor.

Chapter 3

CHAPTER 3: THE SEQUENCING AND INITIAL ANALYSIS OF 82 KILOBASES OF DNA AROUND THE TNF LOCUS WITHIN THE MHC CLASS III REGION

3.1 INTRODUCTION.

Prior to the start of this project, the MHC class III region had been completely cloned in cosmid and YAC contigs, extensive physical mapping had been carried out and transcript maps constructed (Sargent *et al.* 1989; Spies *et al.* 1989a; Spies *et al.* 1989b; Kendall *et al.* 1990; Abderrahim *et al.* 1994). Some of the mapped genes were found to be very closely clustered, while others were more widely separated from each other. This suggested that some of the large, and as yet un-characterised, intergenic regions may contain additional genes. With this in mind, a genomic DNA sequencing project was initiated to further characterise the telomeric half of the MHC class III region. Using the cosmid contigs available (Sargent *et al.* 1989) a minimal set of overlapping cosmids, covering the region between the C2 and BAT1 genes, was selected (Figure 3.1A). There was evidence to suggest the existence of disease susceptibility loci at the far telomeric end of the MHC class III region (Kelly *et al.* 1985; Degli Esposti *et al.* 1992b; Moghaddam *et al.* 1997; Schroeder *et al.* 1998) (section 1.1.4.2). Therefore, the genomic region encompassing the cytokine gene cluster, LTB/TNF/LTA, represents a major focus of this investigation and is the subject of this Thesis. In this chapter the genomic DNA sequencing and initial analysis is described, whilst more detailed analyses of the AIF-1, 1C7 and ATP6G genomic regions are included in Chapter 4.

3.1.1 Methodology

The two overlapping cosmids selected for this study, TN82 and TN62, cover approximately 82kb of DNA around the TNF locus (Figure 3.1B). The genomic inserts of

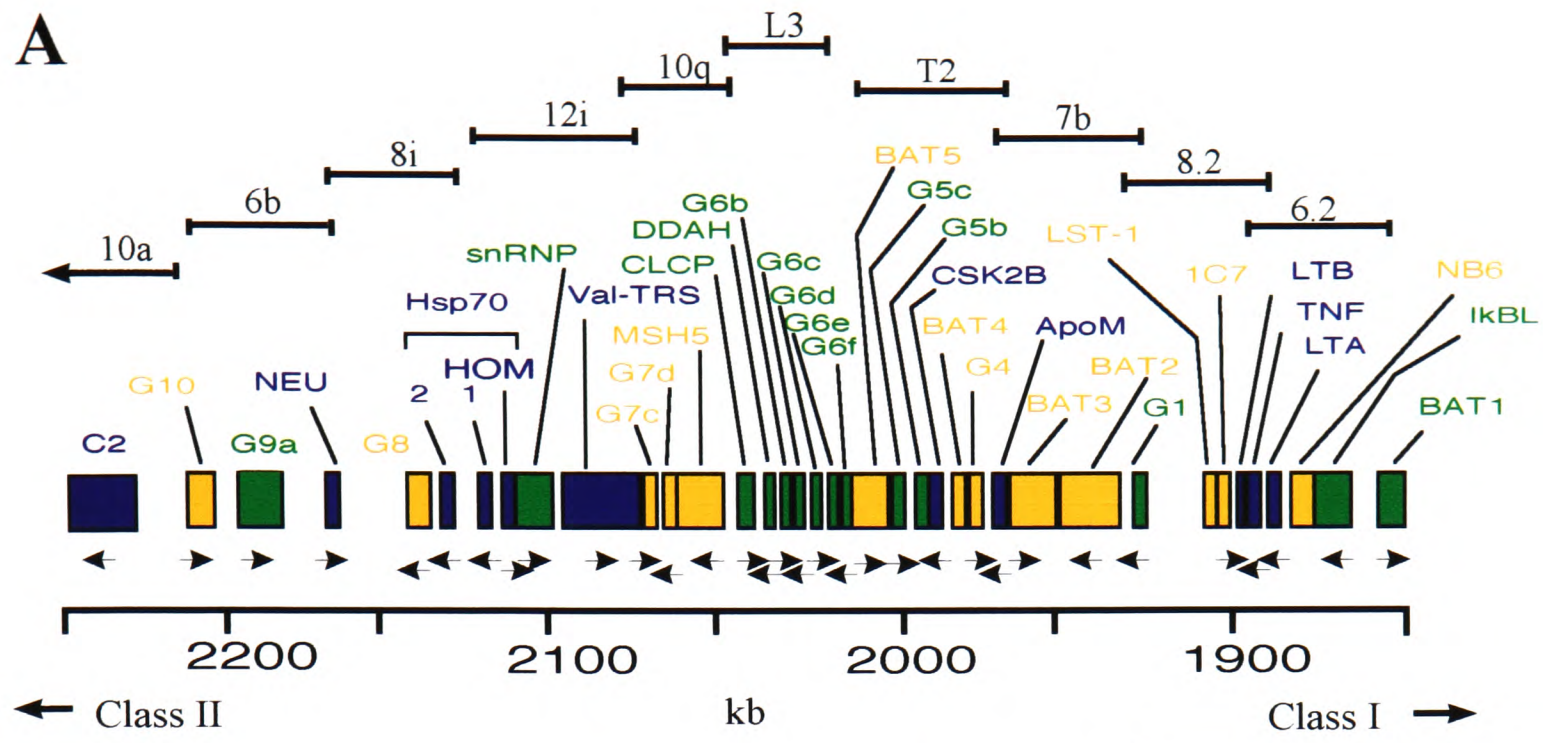


FIGURE 3.1. The telomeric half of the MHC Class III region.
A. A minimal set of overlapping cosmids covering the telomeric half of the MHC class III region between the C2 and BAT1 genes.
B. The region covered by the two overlapping cosmids TN62 and TN82 selected for sequence analysis in this thesis.
Genes are indicated by boxes, with the direction of transcription shown by arrows. Restriction sites within cosmids TN62 and TN82 are indicated in the lower part of the figure.

these two clones were prepared by *Mbo*I digestion of genomic DNA from an EBV transformed lymphoblastoid cell line derived from a consanguineous individual (homozygous for the HLA haplotype: A2, B7, DR2, C2C, BFS, C4A3, C4BQO) and were ligated into the vector pDVcos (Sargent *et al.* 1989).

3.1.1.1 M13 library preparation and sequencing

The generation of an M13 shotgun library (see section 2.3.13.1) is generally considered to be the most efficient method of preparing large fragments of DNA for sequencing (Bankier *et al.* 1987). It has the advantages that sequence determination is very fast, no prior knowledge of the target DNA is needed, clones of any size can be sequenced and the redundancy of data generated ensures a very high level of accuracy. To sequence a genomic insert, the entire cosmid clone is subjected to high frequency sound waves (sonication) which shears the DNA in a random fashion. By timing the sonication period accurately, fragments of the approximate size needed for ligation into the M13mp18 vector can be generated and further size selection can be achieved by electrophoresis (see section 2.3.13.1). The M13mp18 bacteriophage is used as a cloning vector for two main reasons. Firstly, in its replicative form it is a circular double stranded molecule and can be treated very much like a plasmid for cloning purposes. Furthermore, the mature phage particles, which are continuously produced, are packaged linear single stranded molecules which are extruded from infected *E. coli* cells into the surrounding media. Single stranded DNA is by far the best template for DNA sequencing and can be easily harvested in large amounts from the media of *E.coli* cells infected with a single recombinant M13mp18 clone. Secondly, the cloning site of the M13mp18 vector lies within a β-galactosidase gene engineered into the M13 vector, therefore, blue/white colour selection can be used to identify recombinant clones (see section 2.3.8.2.5).

The genomic inserts of cosmids TN62 and TN82 were estimated to be between 40kb and 45kb in length. Reliable sequence exceeding 600 bp in length can be obtained per lane of an ABI 377 automated sequencers and a target sequence redundancy of between 6 and 8 templates per base sequenced is required to ensure accuracy of the data generated. Therefore, ~600 random clones per cosmid were initially prepared for sequencing, the phage DNA being purified from the infected *E.coli* supernatants using a Vistra DNA Labstation (see section 2.3.8.2.6).

Initial sequencing was carried out using the dye-primer sequencing chemistry (Amersham) with the thermostable DNA polymerase ThermoSequenase (see section 2.3.12.1). This chemistry uses the Sanger chain termination method (Sanger *et al.* 1977) which relies on the synthesis of a new strand of DNA starting at a specific primer site and ending in the incorporation of a chain terminating dideoxynucleoside triphosphate (ddNTP). The relative proportions of dNTPs and ddNTPs are balanced so that, on average, there is an equal chance of a ddNTP being incorporated at each position within the desired length of sequence. A separate reaction is set up for each of the four ddNTPs (ddATP, ddCTP, ddGTP, ddTTP) and for each reaction the specific primer is labelled with a different fluorescent dye. The reaction mixes are cycle sequenced, whereby repeated cycles of thermal denaturation, annealing and extension/termination increase the levels of dye label incorporated into the newly synthesised DNA sequence and, therefore, decrease the amount of initial template required. The dye-terminator sequencing chemistry (Amersham ThermoSequenase dye-terminator cycle sequencing kit) was also utilised in this project (see section 2.3.12.2). The dye-terminator sequencing chemistry utilises the Sanger chain termination method like the dye-primer chemistry does. However, the four fluorescent dyes are incorporated into the different termination ddNTPs rather than into the primer.

There are a number of advantages in using the dye-terminator chemistry. It is more robust and can easily cope with large variations in template concentration and with double stranded PCR products. Also, the reaction is undertaken in a single tube, unlike the four reactions needed for the dye-primer chemistry. However, the disadvantages are that the read length is generally shorter and the trace file peak heights are less even than those obtained by dye-primer sequencing, leading to a greater number of miss-called bases by the ABI software. This is becoming less of a problem, however, as chemistries improve.

3.1.1.2 Sequence assembly

Sequenced templates were electrophoresed for 10 hours on an ABI 377 automated sequencer (see section 2.3.13.2), data were collected with the ABI-Prism sequencing collection and analysis software (version 2.1.2, PE Applied Biosystems) and the sequence trace files were entered into the PREGAP Staden suite of software. The PREGAP software modifies the ABI trace files by, for example, adding information on base calling accuracy and masking vector sequences, repeat sequences and low quality sequence (see section 2.3.13.2). There are a number of complications associated with the re-assembly of sequences generated using a random shotgun strategy. These include: A) cosmid and bacteriophage vector sequence will invariably constitute a proportion of the sequences, and B) the human genome contains a large proportion of repeated sequences that have spread throughout the genome during evolution, such as SINE, LINE and MIR repeats (discussed in more detail in section 3.3.2.4). All of these sequences are masked by PREGAP and are thus ignored during the sequence alignment process so that they don't interfere with correct template assembly. Sequence alignment was performed using the Staden assembly software 'GAP' (version 4.0-beta4), which not only attempts to align the sequences but also acts as a graphical interface by which the aligned sequence can be checked, the

sequence trace files viewed to identify and resolve stops and compressions and strategies designed to fill any gaps in the sequence (see section 2.3.13.3).

3.1.1.3 Sequence analysis

The Human Genome Project has prompted major advances not only in sequencing technologies and chemistries but also in the field of bioinformatics. The amount of data held within the DNA and protein sequence databases has increased at an incredible rate over the last few years and improvements in bioinformatics software have helped keep up with this flood of information. Increasingly powerful similarity search engines using the FastA (Pearson & Lipman 1988) and BLAST algorithms (Karlin & Altschul 1990) are incorporating graphical interfaces to assist visual sorting of the large amounts of data generated from a database search. These include the NCBI's BLAST database search engine (http://www.ncbi.ulm.mih.gov/cgi-bin/BLAST/nph-psi-blast) and the EBI's FastA3 search engine (http://www.ebi.ac.uk/fasta3). Also, a large number of exon and gene prediction programs have been developed to help identify potential coding regions. These include diverse approaches such as, the neural network based GRAIL program (Uberbacher & Mural 1991) and the hidden markov speech recognition algorithm based VEIL program (http://www.cs.jhu.edu/labs/compbio/veil.html) (Henderson *et al.* 1997) to name just two. No single program is 100% accurate and they each utilise different prediction methods which tend to fail on different aspects of the analysis. Therefore, an important recent development is the ability to view the results of data analysis by many programs side by side using single WWW based tools, such as NIX (http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/) for DNA analysis and JPRED (http://circinus.ebi.ac.uk:8081/) for protein secondary structure prediction. The graphical outputs of tools such as NIX and JPRED provide a consensus of information that can be used to identify

genomic regions that warrant further experimental investigation. The tools utilised in the course of this research are outlined in Chapter 2 (section 2.4).

3.2 RESULTS

3.2.1 Sequencing and Sequence Assembly

The two overlapping cosmids covering the TNF region, TN6.2 and TN8.2 (Figure 3.1), have been completely sequenced with all ambiguities resolved and gaps filled (summarised in Table 3.1). A combined total of 1757 templates over the two cosmids were entered into GAP4 databases. This number includes all PCR templates sequenced and those templates re-sequenced to resolve ambiguities and to cover single orientation reads. Of the sequences entered into the databases, after the removal of vector and bad quality sequence by the PREGAP software, over 50% were of a length greater than 500 nucleotides (nt) (Figure 3.2) with the longest read being 859 nt. Over the two cosmids a total of 121 stops and compressions were resolved and 5 gaps were closed (Table 3.1).

TABLE 3.1 Summary of the sequencing of cosmids TN62 and TN82.

Cosmid	Insert length [a]	No. of templates sequenced	Average read depth per base	Ambiguities resolved [b]	Gaps filled	GC content
TN6.2	41724	838	9.94	57	4	49.2%
TN8.2	45417	919	9.33	64	1	50.4%

[a] Insert length in nucleotides.
[b] Ambiguities represent both stops and compressions.

The deduced sequence of the two cosmids combined is 81800 nucleotides in length with an overlap between the two cosmids of 5341 nucleotides and an average depth of ~9.6 reads per nucleotide sequenced. There were no discrepancies between the sequences generated for the two cosmids across the region of overlap. The complete genomic DNA sequence of cosmids TN62 and TN82 has been deposited in the EMBL database under the accession

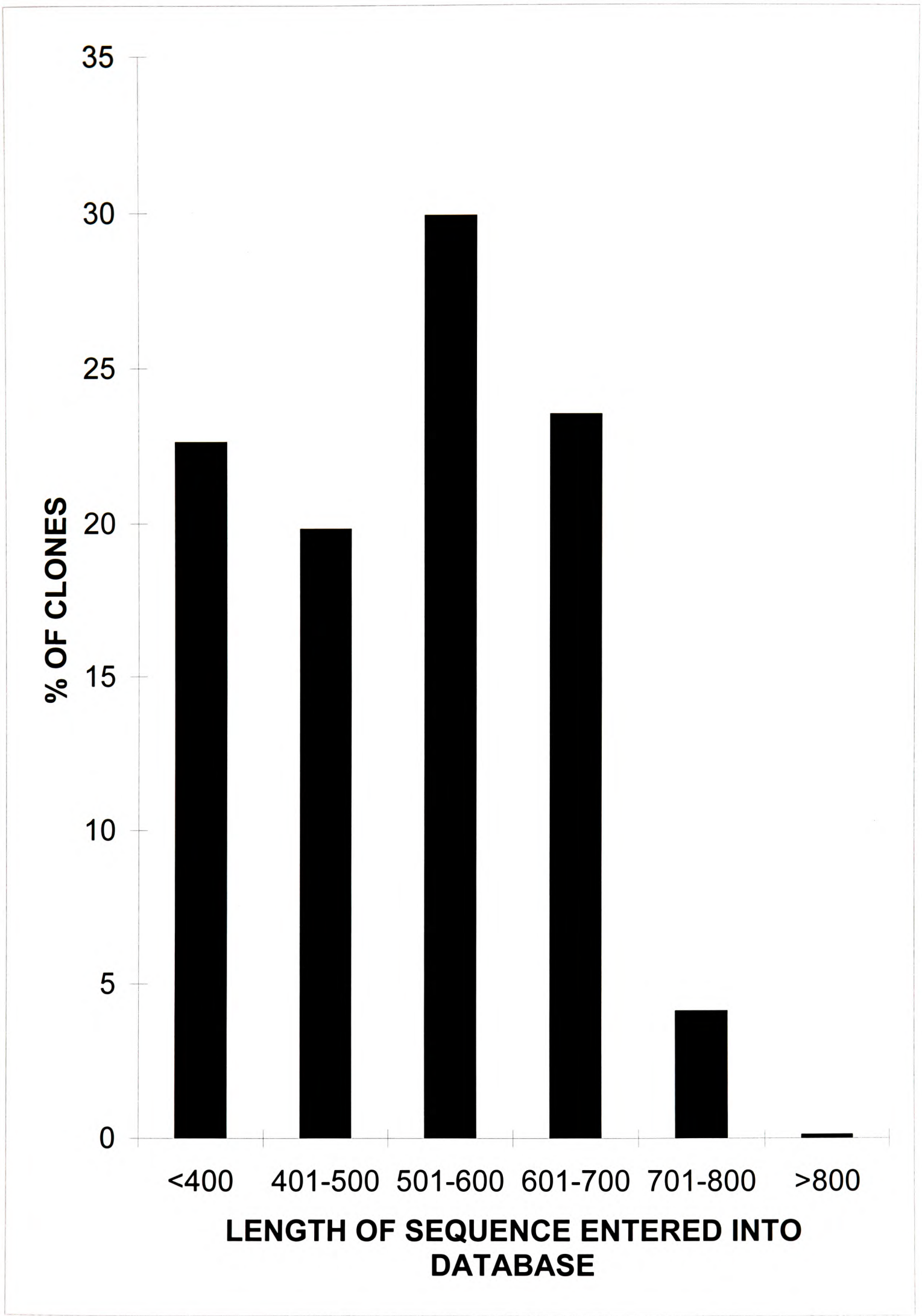


FIGURE 3. 2 Lengths of high quality and vector clipped sequence generated from 1757 randomly picked M13 templates sequenced using an ABI 377 DNA sequencer.

number HSY14768 (Appendix I). Exon positions and other sequence features are included in the annotation accompanying this database entry.

3.2.1.1 Finishing

The five gaps in the sequence of cosmids TN62 and TN82, summarised in Table 3.1 and Table 3.2, were closed by PCR. The PCR products covering the four gaps within cosmid TN62 were sequenced directly using the dye-terminator sequencing chemistry (see section 2.3.12.2). The single gap within cosmid TN8.2 was closed by sequencing single stranded templates generated from the cloning of the appropriate PCR product into the M13mp18 bacteriophage vector (see section 2.3.13.3.3). This was necessary as the gap was bordered by strings of polyAs and polyTs, both of which were found to inhibit the sequencing of double stranded templates.

All stops and compressions were resolved by re-sequencing the relevant templates using dye-terminator sequencing chemistry (see section 2.3.13.3.1). This did not necessarily remove the problem but shunted it to a different position a few bases away so that, overall, the correct sequence could be deduced. For regions of single orientation read, the initial approach was to PCR amplify the M13 insert and sequence the clone in the reverse orientation as described in Chapter 2 (section 2.3.13.3.2). However, it was subsequently found that re-sequencing the original single-stranded template using dye-terminator sequencing chemistry was just as efficient at highlighting and resolving any sequencing ambiguities as sequencing the reverse strand of a template and was less time consuming.

TABLE 3.2 Summary of gaps filled by PCR amplification to complete the sequencing of the overlapping cosmids TN6.2 and TN8.2.

PRIMER REFERENCE NAME	PCR PRIMER SEQUENCE	PRIMER [a] CO-ORDINATES	PCR PRODUCT SIZE	NO. of TEMPLATES SEQUENCED	PCR CYCLING CONDITIONS
Cos8.2-105f	5' GCGAATCTCAC AGACATTAAGC 3'	17092	855	14	0.25μM primer, 1.5mM Mg Cl_2 annealing: 59°C, 30 cycles
Cos8.2-707r	5' ATATGTACACCC ATGCAACCAC 3'	17947			
Cos 6.2-272fnest	5' CCTCCTACCACC ATCTTTAGTG 3'	54987	431	7	0.25μM primer, 2mM Mg Cl_2 annealing: 60°C, 25 cycles
Cos 6.2-NB324a	5' GCTGTAAAACCA GTAGACTTGAG 3'	55418			
Cos 6.2-29f	5' CTCATGAGTAT CTTACAGTC 3'	64806	761	8	0.25μM primer, 1.5mM Mg Cl_2 annealing: 60°C, 25 cycles
Cos 6.2-29r	5' ATAGAGGCCTTC CTAAGAGTG 3'	65567			
Cos 6.2-443f1	5' ACTCTATCTAG TCCATCTGG 3'	70558	776	13	0.25μM primer, 1.5mM Mg Cl_2 annealing: 62°C, 25 cycles
Cos 6.2-443r1	5' GGCCTCGATGT AGATGCTGG 3'	71334			
Cos6.2-443f2	5' ATGGAGGTGGA GCAATACCG 3'	73442	714	19	0.25μM primer, 1.5mM Mg Cl_2 annealing: 62°C, 25 cycles
Cos6.2-443r2	5' CAGACCATGCC AAGAAGCTG 3'	74156			

[a] The primer co-ordinates relate to the nucleotide position within the combined TN62 and TN82 sequence generated in this study.

3.2.2 Sequence Analysis.

3.2.2.1 Sequence composition and the positions of previously identified genes

As a result of the work described here the precise locations and genomic structures of the 9 known genes in the region analysed have been determined and the order of these genes has been shown to be: G1/AIF-1, 1C7, LST-1 (B144), LTB, TNF, LTA, NB6, IKBL, BAT1, centromere to telomere. The G1 (Olavesen *et al.* 1991) and AIF-1 (Utans *et al.* 1996) transcripts were found to be derived from the same genomic region, while the 1C7 gene has been shown to lie on the opposite side of LST-1 (B144) to that previously published (Nalabolu *et al.* 1996). A summary of the relative gene positions as well as other features is diagrammatically represented in Figure 3.8 and exon co-ordinates, in the context of the 81.8kb of deduced sequence, are recorded in Table 3.7.

Cosmids TN62 and TN82 have G & C nucleotide contentS of 49.2%, and 50.4%, respectively, giving an overall GC content of 49.6%. Of the 81800 nucleotides resolved, 20% represents exonic sequence and 40.5% corresponds to genes (i.e. exonic and intronic regions combined). Therefore, 59.5% of the 81800 nucleotides represents intergenic sequence, of which approximately 80% is made up of repeat sequences, i.e. repeat sequences account for 43% of the combined TN62 and TN82 sequence (see section 3.2.2.5).

3.2.2.2 Comparison with other human genomic sequences within the databases

At the point of writing this Thesis, eleven further human genomic sequence database entries (EMBL) were found to align with the G1-BAT1 genomic region. Figure 3.3 shows the positions of these database entries relative to the region described here while their precise locations and other details are given in Table 3.3.

TABLE 3.3 The co-ordinates for the genomic sequence database entries in relation to the deduced sequence of overlapping cosmids TN62 and TN82 (Figure 3.3).

Accession No. (EMBL)	ID	Genomic Region covered	Start	End	Length covered within TN62/TN82 sequence	% identity to the TN62/TN82 genomic insert sequence
AF129756	AF129756	MHC class III MSH5 to LTA	1	54254	54254	99.9
[a]Z15025	HSBAT2	intergenic BAT2 to 1C7				
U00921	HS92115	LST-1 + partial 1C7	27851	33316	5465	99.3
L11016	HLYTOXBB	LTB	34424	40728	6304	100
Z15026	HSTNFABX	TNF + LTA	38676	51960	13284	99.7
X02910	HSTNFA	TNF	41033	44675	3642	99.8
U42625	HS4262510	TNF Promoter and 5' coding seq.	43850	44446	596	99.8
X59350	HSTNFAPRO	TNF Promoter	44415	44734	319	99.1
X02911	HSTNFB	LTA	45108	48143	3035	99.8
M55913	HSTNFBA	LTA	45265	47404	2139	99.8
X59351	HSTNFB5ED	LTA 5' region	46536	48533	1997	99.9
Z15027	HSHLA1467	Intergenic LTA to IKBL	63999	74957	10958	99.7
AC004181	AC004181	LTA to MICB	50362	81800	31438	99.8
AB000876	AB000876	IKBL to MICA	57098	81800	24702	99.8

[a] The Z15025 database entry alignment stops at position 4742; the alignment should continue for a further 17.2kb into the deduced sequence. There are further regions of alignment, but these are not linear and some are in the wrong orientation. Thus no details of the Z15025 alignment are included here.

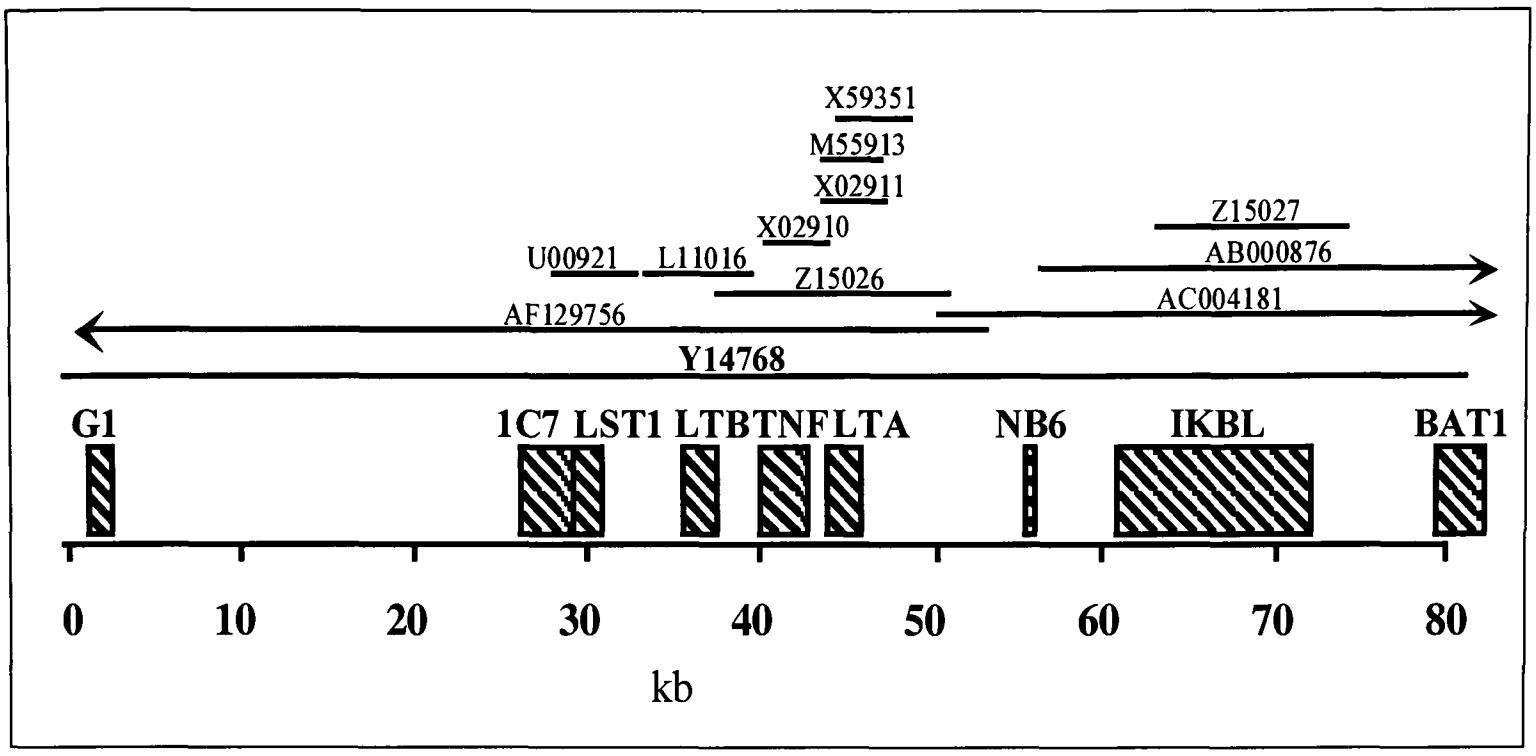


FIGURE 3.3 The relative positions of genomic sequences found within the EMBL database that overlap with the G1-BAT1 genomic region. Each entry is represented by a horizontal line and the relevant accession numbers (arrow heads indicate that the sequences extend outside the region of interest). The data reported in this Thesis corresponds to accession number Y14768.

The database entries AC004181 and AB000876 contain the sequence data reported by Guillaudeux *et al.* (1998) and Shiina *et al.* (1998). These two database entries, together with the database entry AF129756, represent the most substantial contributions to the genomic sequencing of this region, other than the work reported here.

The sequences reported by Guillaudeux *et al.* (1998) and Shiina *et al.* (1998) both overlap the telomeric end of our contig and share 99.8% identity, AB000876 over 24702 nucleotides and AC004181 over 31438 nucleotides. The database entry AF129756 overlaps the centromeric half of our contig with 99.9% identity over 54254 nucleotides. Further investigation will be needed to determine whether any of the nucleotide differences between the generated sequences are haplotype specific (see Chapter 5).

The database entries: L11016, Z15026, X02910, X02911, X59351 and M55913 cover the 17.5kb encompassing the LTB/TNF/LTA gene cluster. The entry U00921 covers the LST-1 gene and also includes part of the 1C7 sequence. Discrepancies were found between the sequence generated in this study and the other database entries, except L11016, with the percentage identity being around 99.8% in each case (Table 3.3). The database entry Z15025 aligns with the centromeric end of the TN62/TN82 sequence for 4.7kb with ~95% identity. Although the Z15025 sequence continues for a further 17.2kb, into the Alu rich intergenic region between G1 and 1C7, the alignment with our genomic sequence breaks down over this region and details are, therefore not included in Table 3.3 or Figure 3.3. It is clear that Iris *et al.* (1993) experienced considerable difficulties in sequencing this segment of DNA. This is a highly repetitive region and the low depth of read per base obtained, together with the shorter read lengths attainable using older sequencing technology would have made it extremely difficult to correctly assemble the sequence data generated. Indeed, areas of alignment with our sequence were found, but these are not in a linear order and some are in the wrong orientation. With the advances in sequencing technology and a greater depth of read in the sequencing strategy adopted here, we are confident that the sequence generated in this project is correct. This has recently been confirmed by the database entry AF129756, which also covers this region (Figure 3.3).

3.2.2.3 Analysis of known genes

3.2.2.3.1 G1

2kb of sequence encompassing the G1 gene was screened against the DNA sequence databases. This region was found to align not only with the G1 cDNA (Olavesen *et al.* 1991) but also with two further human EMBL database entries, the Allograft Inflammatory Factor-1 (AIF-1) gene (accession no U19713) (Utans *et al.* 1996) and the Interferon

gamma responsive transcript-1 (IRT-1) (accession no U95213) (Autieri & Agrawal 1998), and with a number of cDNAs from other organisms and 16 ESTs from the EST databases. These results are presented and discussed in full in Chapter 4 of this Thesis. However, from now on G1 will be referred to as a splice variant of the AIF-1 gene and the genomic region encompassing the G1 sequence will be referred to as the AIF-1 genomic region.

3.2.2.3.2 1C7

In addition to the published incomplete 1C7 cDNA sequence (Nalabolu *et al.* 1996), two I.M.A.G.E consortium cDNA clones, 685808 (EST AA262074) and 683968 (ESTs AA237100 and AA236886), both of B cell origin, were found to align within the genomic region immediately centromeric of the LST-1 (B144) gene. The complete 1C7 cDNA has been found to contain a single putative immunoglobulin domain and a membrane spanning region. These results are presented and discussed in full in Chapter 4 of this Thesis.

3.2.2.3.3 LST-1(B144)

The LST-1 genomic region was found to match both the published LST-1 cDNA sequence and fifteen single entry ESTs. These corresponded to 7 distinctly different putative transcripts representing a minimum of 9 exons, the details of which can be found in Table 3.4, Figure 3.4 and Figure 3.5 and published in Neville & Campbell (1997). Although the majority of ESTs were from liver and spleen cDNA libraries, ESTs corresponding to transcripts from blood, placenta, foetal lung, brain and heart tissue were also found (Table 3.4).

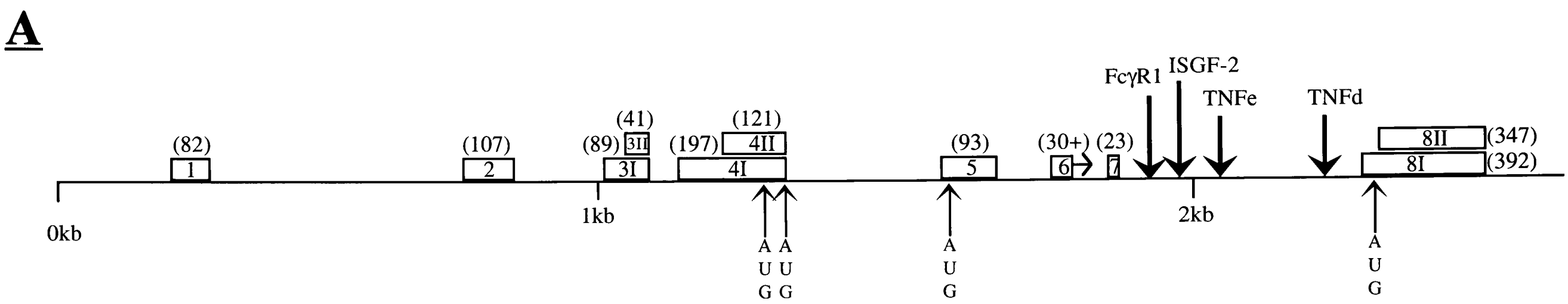


FIGURE 3.4 The genomic structure and splice variants of LST-1.

A. The LST-1 genomic structure. Boxes represent LST-1 exons from all splice variants as defined by EST and EMBL database sequences recorded in Table 3.4. Numbers in brackets represent exon length in bases, FcgR1 represents a IFN-g-activation site, ISGF-2 represents a IFN-stimulated gene factor-2 response element site (Holzinger *et al.* 1995) and TNFe and TNFd represent the positions of two polymorphic microsatellites. The arrows below the scale bar point to initiation AUG codons used in the different open reading frames (Figure 3.6).

B. Splice variants of LST-1.

Exons for each splice variant relate to those shown in Figure 3.4A. All splice variants were derived from sequences deposited within the EST databases (Table 3.4) except for Lst1-db which represents the LST-1 gene database entry. LST1-7 is represented by a single EST (W68662) (Table 3.4) which is incomplete at its 3' end; therefore, the position of the 3' end of exon 6, found in this transcript, is not known.

GCCCAGAAGT TTGAAAAGAA AGATGTTGTA ACCCTAACCC TGGAGCCGAA CTTCCTCTCC TAACAATGCT GGGGAGGAAC CCAGGCTGGG GGAGAAGTTA

Exon 1 5' start site ↓

AAGCCAGAGG AGGGGCAGGA AT|**GTCTGAGG TGGCAACACT TCTCTTCAGC CAGACAGCAC TGGCCAGTTT GGAGTCTGTC CATCCTGCAG GCCACAAGCT**

CTGG⌋GTAAGC TGGGAATGGG CAGGGACCTT GGTGGAAGGA TGGTCACACC CCAGAGTGGG GTGAAGCTAA GATGAGGGGA GGGAGAGTAT GGGTTTGAGT

TTCCCTGGGC CGTCGAGGAA TCCTCTGAGT CTCTGCTCCC CAAAGAAATT AAAGACAATT CATTTCTGTG CCCACGGCCC TTATGGCCTC CACCTGCACT

TCTGCTCCCC ACCCCCCAGA ATTCCTCTTA AACCCAGAAG GGTCCCAGTT TCCAGACCCT AGTCAGTATA TCTGGCTCTG GGGTGAAGAG AACGGCCCCC

Exon 2 5' start site ↓

TCTTCACCCT CAAACAGGAA CCAGTGGTTG GAGGGGAGGA AGTGCCTGAG GGGAAGTTAT GGGGCCCCAG ATACTCCTCC ATGCCCCAC|**T TCAGCCCTAG**

CAGCATCTGC CTGTGGGAAG CAGCTCTCCA CACCAGCCAA GGGGGCCCCC ACACTCCCGC GCTGCTCTGC GGCTCAGGGA GCAGCCCACC TGCTGG⌋GTGT

GCTGATATCA CCCTCCCTTC TTCCCCCCAG TGCCCACACC CACCCAGGCC CAGGCTCCTT CCCCTCCATC ATCCCCTTAC CAGCACCTAG AACCATCCAG

GGCTGAAAAG TCCCCTCCAA ACCACGTGGT CAGCCCAGGG CAGAGGAAAG GGCTGGGCTC TGGAGTTGGG CAGAGCTGGC CTTAAACCCC AGCTCCACCT

Exon 3I 5' start site ↓ Exon 3II 5' start site ↓

TTCTGGGATG GGTGACCTAG |**TAAAGTCCAG GCTTGAATCT CGGGTCTTTA CTTGGGCAAC GGGCACCA**|**TG ATACCCTATG TTCTGGGGAT TAGCAGTGAG**

Exon 4I 5' start site ↓ Exon 4II 5' splice site ↓

GAATG⌋GAAAG |**TGCCCAGCTC AGGGTTGGCA CATAAGGGAG GCTCCCCAGC CTGGGAACGA TTATAACAGA GGGCCCCTCA CTTCACAG**⌊**AT GAGGAACTTG**

MET 1

AGGCAAGTCA CCAGCCCCTG ATCATTTCGC CTAAAAGAGC AAGGACTAGA GTTCCTGACC TCCAGGCCAG TCCCTGATCC CTGACCTA[**AT G**]**TTATCGCGG**

MET 2

AATG[**ATG**]⌋GTA AGTAAAGTGT CTCTTGCATC TGCATAGAGA GAGTCCTGGG AGCTTAGGAA GTGATGGGGA ACAGTGATGT ATGCAGCTCA TGACTAGGTG

GACAGGCCTC TGGGGACAGC TGGTACAGGA GGGAAAGGGA CCTCACGGGA GGCCCAGAAA CCTGGTAAGA GGTGAGGTAT TAAGGTCTGG GATGGAGAAG

CTCTGAGGGT ATATTTTTCT GCCTCTAAAA CTGTTGGAGA GGGAATCTGA GAAAGCTGCA ACCAACCAGG AGGCTGGGGT ACGCTGGAGA AGGAATGGGC

Exon 5 MET 3

TTCCTAACCT TGAGCCCTCT TCCCTGAAG⌊**A TAT**[**ATG**]**TATC TACGGGGGCC TGGGGCTGGG CGGGCTCCTG CTTCTGGCAG TGGTCCTTCT GTCCGCCTGC**

CTGTGTTGGC TGCATCGAAG AG⌋GTGAGCGC TGCACTCCCT CCCTCCCCCT GCAGCAGTGC CCCCTGTGCC CCCACCCCCA CACGCTTTCC CACTGCTTTC

Exon 6 → ?

CCAG⌊**AACACT GCCTGGCCCT GGAGCCACTG GGAA**GCCAAC AGGGGAGTCC ACGCCTGCTG GTGGGGGGAG CCCGGGAGGG CCCGGGAGAA GCACAAAGGG

Exon 7

TGGGCTGTGT TGAGCTTCTT CTTTTCTTCC AG⌊**TAAAGAGG CTGGAGAGGA GCTGG**⌋GTGAG TCTGGGGACA GGGAAGGGGG AGGGCAAGAG AGATCCTGAG

TGGGTGAGTG GGGAGAAGCA TGGCTGAGCG CTGAGAGGAG GGTTGGGGAC GGGAGACAAG GAGAGAGAAA GTAGGAGCAT GAGAGAGGCA GAGAAAATCG

Microsatellite TNFe

AGGCAAAAGA GAAAGAGAAA ATGAGACAGA AACCAAGAGA AAAAGTGAGA CAGAGGATAG GAGAGACAGG GAGAAAATGA GAGTGAGAGA GACACAAAGA

GAAGAGCAAT GAAAGAGAGA GAGAGAGAGA GGCTCCAGAA CCAGGCACAG TGGCTCACGT CTGTCATTCC AGCTATCGCA AGGCTGAGGC AGGAAGATAG

Microsatellite TNFd

CTTGAGCTCA GGGGTTGAAG ACAATCCTGG ACAACATAGT GGGACTCTGT CTCCAAAGAA AAAAGAGAGA GAGAGAGAGA GAGAGAGAGA GGGAGAGAGA

GAGAGAGAGA GGGAGAGAAG TAAGAAAGGC TGGAGGTGGG AGCAGAACTC ACAGGGAAGG ATCTGACGGC ATCGCCTCCC ATCAGCACCT TCTGTCCTGG

↓Exon 8I 5' splice site MET 4 ↓Exon 8II 5' *

TCCCAG⌊**GCCC AGGGCTCCTC AGAGCAGGAA CTCCACT**[**ATG**]**CATCTCTGCA GAG**⌊**GCTGCCA GTGCCCAGCA G**[**TGA**]**GGGACC TGACCTCAGG GGCAGAGACA**

*

AGAGAGGCAC CAAGGAGGAT CCAAGAGCTG ACTATGCCTG CATTGCTGAG AACAAACCCA CC[**TGA**]**GCACC CCAGACACCT TCCTCAACCC AGGCGGGTGG**

*

ACAGGGTCCC CCTGTGGTCC AGCCAG[**TAA**]**A AACCATGGTC CCCCCACTTC TGTGTCTCAG TCCTCTCAGT CCATCTCGAG CCTCCGTTCA AATTGATCAT**

CATCAAAACT TATGTGGCTT TTTGACCTTT GAATAGGGAA TTTTTTAAAT TTTTTAAAAA TTAAAATAAA **AAAAACACAT GGCTCACCCT TCCACCCA**|CT

CTGGGGTCAA ATAGTAATTT ATTGGGTGAA TGACAGTGTT CAGGGACCCA AGCTCCCCTA ACAGCCAGAA GAGGGTATGT GTGGGCCTGG CAGGAAAGGG

FIGURE 3.5 The LST-1 genomic region from the genomic sequence generated in this study, co-ordinates 33450-30601 (reverse orientation).

Exonic regions are shown in bold, as represented in Figure 3.4. The exon numbers are underlined above the sequence. The 5' splice sites are indicated with a '⌊' within the sequence, and the 3' splice sites are indicated with a '⌋'. The transcriptional start sites of the alternate first exons of LST-1 are indicated with a '|'. The alternative 5' and 3' splice sites when falling within exonic sequence are further annotated with an arrow (↓) directly above them next to the relevant identifying text. All possible initiation AUG codons (as presented in Figure 3.6) are boxed in bold and numbered above by 'MET 1 to 4'. The three possible stop codons are boxed and indicated above by an asterisk '*'. The polymorphic microsatellites, TNFe and TNFd are indicated with wavy underlines. Also highlighted is the poly-adenylation signal (AATAAA) .

TABLE 3.4 The database sequences found to encode the different splice variants of LST-1.

Splice Variants	Accession numbers [a]	EST library
LST1-db	U00921	Blood
LST1-2	N70057, R71629, AI300316	Placenta, Lung
LST1-3	H18316, H49135	Brain, Lung, Lymph
LST1-4	H71384	Liver & Spleen
LST1-5	R71126, R91669	Placenta, Liver & Spleen
LST1-6	T79306, T84147, T95707, T95606, T79220, R01588	Liver & Spleen
LST1-7	W68662	Heart

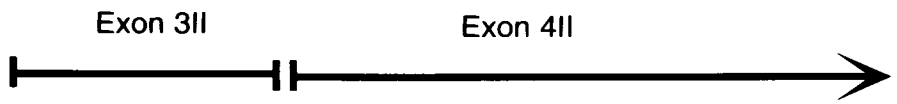
[a] All except the cDNA database entry U00921 represent EST database entries.

No termination codon could be found for the sequence of LST1-7. The most abundant transcript was LST1-6 represented by 6 ESTs of varying length. LST1-2 was represented by 3 ESTs, LST1-5 and LST1-3 by 2 ESTs, whilst LST1-4 and LST1-7 were each represented by 1 EST. Exons 1, 2, 3I, 3II, 4I and 4II appear to be alternative first exons, the exon 3 region (exons 3I and 3II) being the best represented with three out of the 7 variants containing it. All of these first exons appear to be non-coding. It is possible that exon 3II is a truncated version of exon 3I as the only difference between the two is that the exon 3II 5' sequence does not extend as far. However, both exonic regions are represented by more than one EST: exon 3I by 2 ESTs and exon 3II by 4 ESTs, the latter all with identical 5' ends. Therefore, the two overlapping exonic regions may represent distinctly different 5' start sites. Exon 8II covers the same region as exon 8I, but is 45 nucleotides shorter, having a different 5' splice site. The polymorphic microsatellites TNFe and TNFd, the IFN-γ-activated site and the IFN-stimulated gene factor-2 responsive element are all located in the last intron of LST-1 (Figure 3.5) as reported by Holzinger *et al* (1995).

It was noted by Holzinger *et al.* (1995) that there is more than one possible open reading frame within the LST-1 transcripts and this is also apparent here. There are three potential open reading frames for all the LST-1 transcripts, except LST1-6 which contains two. There are four possible different initiation AUG codons and three different stop codons in the LST-1 gene (see Figure 3.5). The initiation AUG codons from the first two overlapping open reading frames are 13 nucleotides apart and found on exon 4II (Figure 3.5). The first initiation AUG codon (MET 1) initiates an open reading frame in all splice variants. The second initiation AUG codon (MET 2) initiates an open reading frame in all but the LST1-6 variant where the splicing of exon 4II to exon 7 results in an in-frame stop codon immediately following MET 2 in this splice variant. The third initiation AUG codon (MET 3) is close to the beginning of exon 5 and, therefore, only initiates an open reading frame in the splice variants containing this exon (i.e. LST1-db, LST1-2, LST1-4, LST1-5 and LST1-7). The forth putative initiation AUG codon (MET 4) is located in exon 8I and only initiates an open reading frame in splice variants LST1-3 and LST1-6. In total there are 17 different polypeptide sequences that could potentially be derived from the LST-1 splice variants (Figure 3.6B). All these open reading frames are shown in Figure 3.6B and their relevance is considered in section 3.3.3 of this Chapter. The splice variant LST1-7 was not included in the amino acid analysis as the coding region was incomplete.

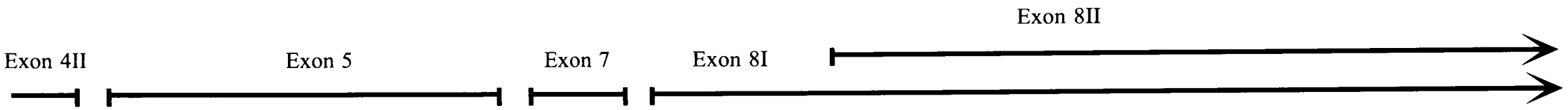
The first open reading frame (ORF1 in Figure 3.6B) is predicted to encode a membrane bound protein in all but the LST1-3 and LST1-6 transcripts, with the transmembrane domain encoded in exon 5. This was also suggested to be the case for the mouse LST-1 homolog, B144 (Holzinger *et al.* 1995). The other open reading frames presented in Figure 3.6 are all predicted to encode soluble peptides and share no significant amino acid identity with any other protein sequence within the databases.

A



MFWGLAVRND. .EELEASHQPL IISPKRARTR VPDLQASP*
MRNLRQVTSP*

B



ORF1

```
lst1-2    MLSRND. .DI CIYGGLGLGG LLLLAVVLLS ACLCWLHRR. .......... .......... ......EAAS AQQ*...... .......... ..........
lst1-db   MLSRND. .DI CIYGGLGLGG LLLLAVVLLS ACLCWLHRR. .......... .GPGLLRAGT PLCISAEAAS AQQ*...... .......... ..........
lst1-3    MLSRND. ... .......... .......... .......... .......... .GPGLLRAGT PLCISAEAAS AQQ*...... .......... ..........
lst1-6    MLSRND. ... .......... .......... .......... .VKRLERSW. .AQGSSEQEL HYASLQRLPV PSSEGPDLRG RDKRGTKEDP RADYACIAEN KPT*
lst1-5    MLSRND. .DI CIYGGLGLGG LLLLAVVLLS ACLCWLHRR. .VKRLERSW. .AQGSSEQEL HYASLQRLPV PSSEGPDLRG RDKRGTKEDP RADYACIAEN KPT*
lst1-4    MLSRND. .DI CIYGGLGLGG LLLLAVVLLS ACLCWLHRR. .VKRLERSW. .......... ......RLPV PSSEGPDLRG RDKRGTKEDP RADYACIAEN KPT*
```

ORF2

```
lst1-4         M. .IY VSTGAWGWAG SCFWQWSFCP PACVGCIEE* .......... .......... .......... .......... .......... .......... ...
lst1-5         M. .IY VSTGAWGWAG SCFWQWSFCP PACVGCIEE. .......... .......... ......RLPV PSSEGPDLRG RDKRGTKEDP RADYACIAEN KPT*
lst1-2         M. .IY VSTGAWGWAG SCFWQWSFCP PACVGCIEE. .......... .AQGSSEQEL HYASLQRLPV PSSEGPDLRG RDKRGTKEDP RADYACIAEN KPT*
lst1-db        M. .IY VSTGAWGWAG SCFWQWSFCP PACVGCIEE. .......... .AQGSSEQEL HYASLQRLPV PSSEGPDLRG RDKRGTKEDP RADYACIAEN KPT*
lst1-3         M. ... .......... .......... .......... .......... .AQGSSEQEL HYASLQRLPV PSSEGPDLRG RDKRGTKEDP RADYACIAEN KPT*
```

ORF3

```
lst1-4                 MYLRGPGAGR APASGSGPSV RLPVLAASK. .SKEAGEEL. .......... ......EAAS AQQ*...... .......... .......... .......... .......... .....
lst1-5                 MYLRGPGAGR APASGSGPSV RLPVLAASK. .SKEAGEEL. .GPGLLRAGT PLCISAEAAS AQQ*...... .......... .......... .......... .......... .....
lst1-2                 MYLRGPGAGR APASGSGPSV RLPVLAASK. .......... .......... ......RGCQ CPAVRDLTSG AETREAPRRI QELTMPALLR TNPPEHPRHL PQPRRVDRVP LWSSQ*
lst1-db                MYLRGPGAGR APASGSGPSV RLPVLAASK. .......... .RPRAPQSRN STMHLCRGCQ CPAVRDLTSG AETREAPRRI QELTMPALLR TNPPEHPRHL PQPRRVDRVP LWSSQ*
```

ORF4

```
lst1-3                                                                 MHLCRGCQ CPAVRDLTSG AETREAPRRI QELTMPALLR TNPPEHPRHL PQPRRVDRVP LWSSQ*
lst1-6                                                                 MHLCRGCQ CPAVRDLTSG AETREAPRRI QELTMPALLR TNPPEHPRHL PQPRRVDRVP LWSSQ*
```

FIGURE 3.6 Polypeptides encoded by open reading frames (ORFs) within the different splice variants of the LST-1 gene.
Lines above the amino acid sequence represent the exons (as represented in Figures 3.4 and 3.5) within which the amino acid sequence is encoded.
A. Open reading frames upstream of the main coding region. The first ORF is found in splice variants LST1-5, LST1-6 and LST1-7 only, the second smaller ORF is found in all splice variants.
B. All potential ORFs encoded within the main coding region of LST-1.

In addition to the main open reading frames and in support of the findings of Holzinger *et al.* (1995) we also found that the LST-1 transcripts encode upstream 'mini-cistrons', or small open reading frames (Figure 3.6A). All splice variants encode a 10 amino acid peptide, within exon 4II, which terminates 67 nucleotides up-stream of the first main initiation AUG codon. In addition, the splice variants LST1-5 and LST1-6 (and LST1-7) contain a longer open reading frame, starting 7 nucleotides into exon 3II, extending for 113 nucleotides and terminating 11 nucleotides upstream of the first main initiation codon. This encodes a 38 amino acid peptide with no significant similarity to any database sequences.

3.2.2.3.4 The LTB/TNF/LTA gene cluster

The genomic structure of the LTB/TNF/LTA gene cluster has already been studied extensively (Nedwin *et al.* 1985; Abraham *et al.* 1991; Messer *et al.* 1991; Browning *et al.* 1993; Iris *et al.* 1993; Zimmerman *et al.* 1996) and there is little genomic information that can be added here. The intron/exon co-ordinates for these genes, in relation to the TN62/TN82 cosmid sequence, are given in Table 3.7. In addition, a number of genomic sequences now cover this region (see Figure 3.3) and analysis of sequence differences between them may add to the known polymorphisms in this region. All discrepancies identified between the sequence generated here and other database sequences have been annotated in the database entry (HSY14768, Appendix I). Twenty seven discrepancies were identified within the TNF genomic region, one of which (a C/T transition at position 44294) is the known polymorphism TNFAp1 (Zimmerman *et al.* 1996). Also, fifteen discrepancies were identified within the LTA genomic region: a G/T transversion at position 46605 and a C/T transition at position 47076 represent the known polymorphisms, TNF8-1-3 and TNF8-1-2, respectively (Abraham *et al.* 1991). It is not possible to say yet

whether the newly identified discrepancies are haplotype specific or artefacts of DNA cloning and sequencing. However, they represent a good staring point for the identification of further haplotypic variation within this region.

3.2.2.3.5 IKBL

The 1.6kb IKBL transcript (EMBL accession number X77909) encodes a polypeptide of 381 amino acids. The structure of exons 1 and 2 has been previously reported (Albertella & Campbell 1994), but the remaining genomic structure had not been characterised (co-ordinates given in Table 3.7). Comparison of the IKBL cDNA sequence with the genomic sequence determined here showed that the IKBL gene spans 11176 nucleotides and comprises four exons of 807, 201, 276 and 125 nucleotides in length. Exons 2 and 3 are separated by a particularly large intron (intron 2) of 9189 nucleotides, which contains 20 SINE repeat elements (15 Alu and 5 MIR), 1 L1 LINE repeat element and 1 MER2 DNA element. There is one conflict in the coding region of exon 4 between the cDNA database entry and the genomic sequence reported here. A change from C-G in the database entry to G-C in the genomic sequence (positions 61431 and 61432) results in two adjacent codon changes from CAC-GAC to CAG-CAG and two corresponding amino acid changes: His-238-Gln and Glu-239-Gln. It is not possible to say, without further analysis, whether this sequence variation is haplotype specific or represtents a sequencing error.

3.2.2.3.6 BAT1

Only the first three exons of the BAT1 cDNA sequence (accession number Z37166) are contained within the sequence resolved from within the TN62/TN82 genomic sequence (Figure 3.7, Table 3.7). The remainder of this gene lies telomeric of the region contained within the TN62 cosmid clone.

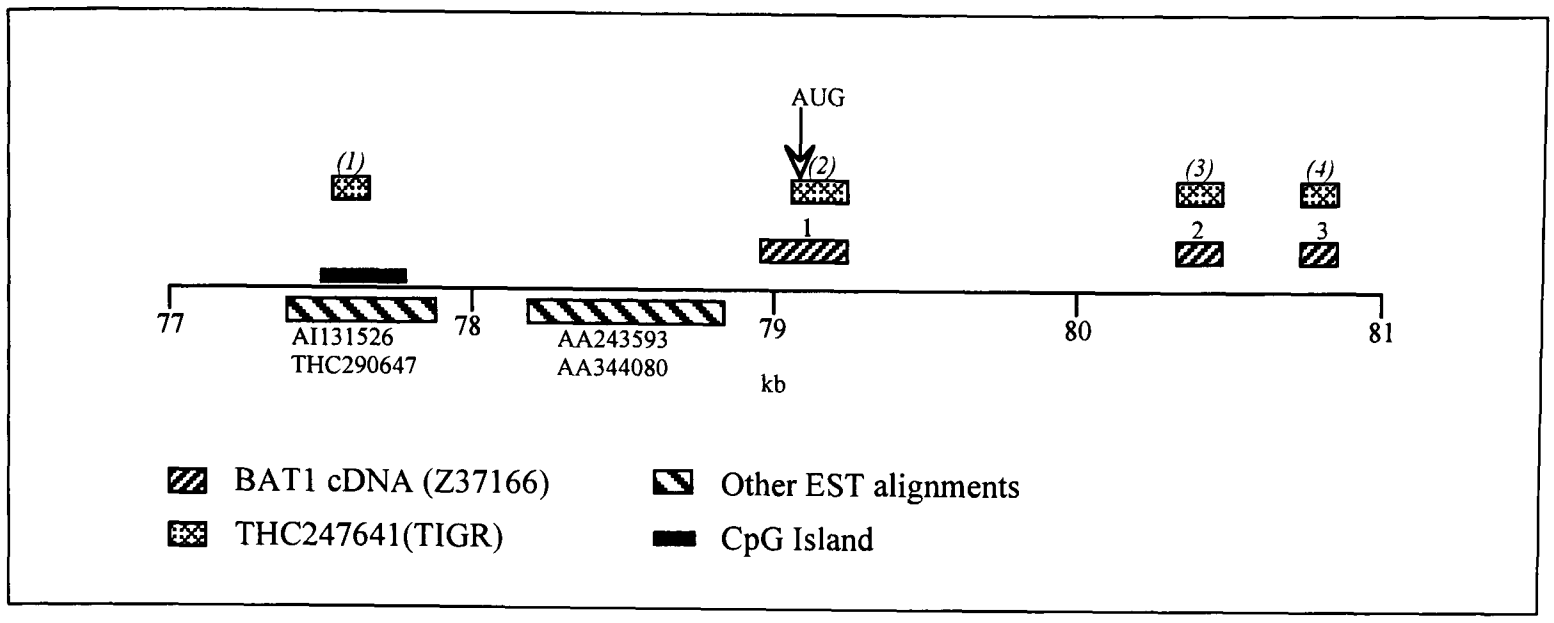


FIGURE 3.7. The BAT1 genomic region contained within the TN62/TN82 genomic sequence. Only the first three exons of the BAT1 database entry, Z37166 are encoded within the region sequenced. In addition to this database entry however, five EST sequences align immediately upstream of, and overlap with the BAT1 cDNA sequence. Of these, only THC247641 aligns with the BAT1 exonic regions. This EST contains an additional exonic region 5' of the previously reported exon 1 of BAT1 and may contain a novel exon 1 of the BAT1 gene. The scale is in kilobases and corresponds to the co-ordinates within the 81800 nucleotides of genomic sequence generated during the course of this Thesis. The numbers above the boxes correspond to exon numbers, those in *italic* and parentheses represent the new BAT1 exon numbers, according to EST THC247641. The arrow indicates the position of the BAT1 initiation AUG codon. The boxes below the scale represent regions of additional EST alignment.

Five EST sequences have been found to align within a ~2kb region upstream of BAT1 exon 1. These align over two genomic segment: AI131526, THC290647(TIGR) and THC247641(TIGR) align within the co-ordinates 77385 to 77886, whilst AA243593 and AA344080 align within the co-ordinates 78170 to 78787 (Figure 3.7). In addition, the GRAIL server predicted the presence of a CpG island between co-ordinates 77489 to 77768. THC247641 is the most interesting of these ESTs as it is divided into four exonic regions (Figure 3.7). Exons 3 and 4 of THC247641 match exactly exons 2 and 3 of the database BAT1 cDNA (Z37166), however, exon 2 of THC247641 although matching the splicing of BAT1 exon 1 at its 3' end is 122 nucleotides shorter at its 5' end, with a 5' splice site at position 79076 (compared to the reported start site for the BAT1 cDNA at position 78954) (Figure 3.7). Exon 2 of THC247641 contains the BAT1 initiation AUG codon

which is located at co-ordinate 79078, just 2 nucleotides downstream of the 5' splice site of this exon. This 5' splice boundary also falls at the 5' splice site of an exon predicted by GRAIL with an 'excellent' (e) level of confidence. Exon 1 of THC247641 is found 1416 nucleotides upstream of exon 2 between co-ordinates 77595 to 77660. Therefore, this appears to represent a previously unidentified non-coding first exon of BAT1.

The EST alignments represented in Figure 3.7 show no intron/exon structure apart from that described for THC247641. The region which aligns with ESTs AI31526 and THC290647 encompasses the new putative BAT1 exon 1 whilst the region aligning with ESTs AA243593 and AA344080 is found within the new intron 1 of BAT1.

Comparison of the genomic database entry AC004181 (see Figure 3.3 and Table 3.3) with the TN62 genomic sequence has identified a 15 nucleotide deletion in the AC004181 sequence (AGGCTGTGCTCGTGA) from co-ordinates 78900 to 78913. This discrepancy is just 40 nucleotides upstream of the database BAT1 exon 1 and falls within intron 1 of BAT1 according to the splicing pattern of THC247641 BAT1 splicing. Further work will determine whether this discrepancy is haplotype specific (see Chapter 5).

3.2.2.4 Identification of new coding regions

There are a number of ESTs and GRAIL predicted exons that align within the TN62/TN82 sequence but cannot be assigned to known genes. These are summarised in Table 3.5 and Figure 3.8 and are described below. There are three genomic regions of interest in this respect: AIF-1 to 1C7, LTA to IKBL, and IKBL to BAT1.

TABLE 3.5 The co-ordinates of additional potential coding regions of no assigned status, comprising alignment with: ESTs, known genes, and predictions by the GRAIL exon prediction program.

BASIS FOR IDENTIFICATION	ORIENTATION	START	END
GRAIL exon (e)	c	7638	7687
THC87311 (TIGR) 80% ID	c	7649	7782
AA737308, R83917, AA528414, UCRH-Human	+	8182	8719
THC127890, T93239, R89178, THC117492 all with ~80% identity	+/c	11911	12089
GRAIL exon (g)	c	12137	12220
GRAIL exon (m)	+	15665	15886
cDNA clone NB6 [a]	c	55364	56066
Est22679 (TIGR)		57155	57473
AA633002	+	58210	58386
cDNA clone 2p1 [b]		58331	58588
GRAIL exon (e)	c	58343	58491
AF150144	c	58935	60318
GRAIL exon (g)	c	59980	60110
GRAIL exon (e)	+	72271	72343
AA401769, AA399356, AA324358, (showing intron/exon divisions)	+/c	72763	74780
GRAIL exon (g)	+	73051	73132
GRAIL exon (e)	+	73401	73501
GRAIL exon (e)	+	74029	74172
GRAIL exon (g)	+	77167	77324

[a] NB6 genomic fragment identified by Zoo blot analysis (Albertella, *et al.* 1996)
[b] Mark R. Albertella-personal communication
GRAIL predictions: (e) excellent, (g) good, (m) moderate

3.2.2.4.1 The AIF-1 to 1C7 genomic region

The TIGR EST entry THC87311 was found to match exon 4II of the AIF-1 gene (see section 4.2.1 and Table 3.7) and is now known to be an EST of the G1 splice variant of the AIF-1 gene. EST THC87311 also aligns, with 80% identitiy, with a segment of DNA 3.3kb telomeric of the AIF-1 gene. This appears to represent a duplicated fragment of the AIF-1 gene.

The region between co-ordinates 8182 and 8719, as well as aligning with a number of ESTs (Table 3.5) shares 80% identity at the DNA level with UCRH-Human (accession number M36647) a ubiquinol-cytochrome-c reductase mitochondrial hinge protein. The region of alignment covers the whole cDNA for this gene, with no splicing evident. However, the TN62/TN82 region does not contain an open reading frame equivalent to that of the UCRH cDNA, due to two frame shift errors (single nucleotide deletions) in the MHC class III region sequence, at positions 8341 and 8392. After allowing for these frame shifts, the UCRH amino acid sequence shares 71% identity with the deduced amino acid sequence encoded in the MHC. The MHC-linked sequence, designated UCRHL (ubiquinlo-cytochrome-c reductase hinge protein like gene) (Figure 3.8), although not containing an open reading frame shares enough sequence similarity to represent a copy or close relative of this gene and is most probably a UCRH pseudogene.

The next EST hotspot, between co-ordinates 11911 to 12089, falls within a region identified as containing a LTR48 repeat region by the RepeatMasker server (see section 2.4). LTR48 is a member of a family of retrovirus-like repeat elements characterised by long terminal repeats (LTR) (Smit 1996a). The alignment of these ESTs within a repeat region suggests that they represent genomic contamination within the EST libraries and,

therefore, will not be investigated further here. The two GRAIL exons predicted in this region (Figure 3.8 and Table 3.5) were also discounted as they fall within Alu repeat elements.

3.2.2.4.2 NB6 genomic region between LTA and IKBL

Telomeric of the segment of DNA named NB6 (Albertella *et al.* 1996) there are three potential coding regions. Firstly, EST22679 (TIGR) shares 100% identity over 318 nucleotides with the genomic sequence 1089 nucleotides telomeric of NB6. 737 nucleotides telomeric of this region, a GRAIL exon (excellent), ESTAA633002 and a 206 nucleotide segment of a 1.4kb chimeric cDNA clone (2p1) from a JY (B cell) library (M.R. Albertella, personal communication) all align, with some overlap, within a 378 nucleotide region. Finally a region between nucleotides 58935 and 60318 aligns with EST AF150144 and a GRAIL (g) exon.

cDNA clones 120912 (ATCC) and 1143343 (I.M.A.G.E) containing ESTs EST22679 (TIGR) and AA633002, respectively were obtained and sequenced. Both were found to contain Alu repeats, as was NB6 (Albertella *et al.* 1996). In addition to this, primers (see below) were designed for RT-PCR analysis to determine whether any transcripts were encoded in the regions that align with NB6, EST22679 and 2p1. RT-PCR was carried out using RNA from Raji (B cell), Jurkat 6 (T cell), U937 (Monocyte), HeLa (fibroblast) and HepG2 (Hepatic) cell lines. No combination of the primers below generated a product, despite varying the PCR conditions.

NB6F: 5'-CTAAGTCTACTGGTTTTACAGC-3' [55397 to 55418]
EST22679F: 5'-ACTAGCCCCCTTAACGAACC-3' [57173 to 57192]
EST22679R: 5'-GTTGGGTTCAGGAACTAGACG-3' [57415 to 57395]
2p1R: 5'-AACACGTCTTGAGCGAGCC-3' [58531 to 58513]

More recently the EST, AF150144 was found to align over 1383 nucleotides between co-ordinates 58935 and 60318 with 99% identity. However, no significant similarity could be found with this sequence and any other sequence within the EMBL database and there was no intron/exon structure evident.

3.2.2.4.3 The IKBL – BAT1 intergenic region

Information gathered from GRAIL predicted exons, EST alignments (see Table 3.5) and sequence similarity of the predicted protein product of the genomic sequence generated here, with an insect V-ATPase G subunit has led to the identification of a novel gene between IKBL and BAT1 (see Figure 3.8). This gene encodes a protein homologous to V-ATPase subunits from a number of different organisms. Full details can be found in Chapter 4 of this Thesis.

3.2.2.5 Repeat sequences

A total of 138 interspersed repeat elements were identified by the RepeatMasker server, covering a total of 35817 nucleotides, or 44% of the TN62/TN82 region sequence. Of these, 34% were from the Alu family of repeat elements. A summary of the RepeatMasker output is shown in Table 3.6 and a full list of the repeat elements and their co-ordinates can be found in Appendix II. All the Alu subfamilies except Sb2 were represented within the region sequenced, the most common being the Jo and SX subfamilies with 17 and 31 elements, respectively. Of the four main divisions of Alu there were 7 free arm monomer elements (FAM, FRAM and FLAM), 12 Alu Y elements, 67 Alu S elements and 24 Alu J elements. The average distribution of Alu sequences over the TN62/TN82 region is 1.2 Alus per kb of DNA. These are almost entirely clustered within intronic or intergenic

TABLE 3.6 The RepeatMasker server output summary for the genomic sequence contained within the overlapping Cosmids TN62 and TN82.

Repeat elements[a]	number of elements[b]	length occupied	percentage of sequence
SINEs:	107	29010 bp	35.46 %
ALUs	100	28072 bp	34.32 %
MIRs	7	938 bp	1.15 %
LINEs:	14	4123 bp	5.04 %
LINE1	4	1277 bp	1.56 %
LINE2	10	2846 bp	3.48 %
LTR elements:	5	1562 bp	1.91 %
Retrov:	1	144 bp	0.18 %
DNA elements:	6	1122 bp	1.37 %
MER1_type:	1	133 bp	0.16 %
MER2_type:	4	864 bp	1.06 %
Total interspersed repeats:		35817 bp	43.79 %

The sequence was submitted to: http://ftp.genome.washington.edu/RH/RepeatMasker.html.

[a] 'SINE' stands for, short interspersed nuclear repeats, 'MIR' for mammalian-wide interspersed repeats, 'LINE' for long interspersed nuclear repeats, 'LTR' for long terminal repeats, and 'MER' for medium reiterated frequency repeats.

[b] Most repeats fragmented by insertions or deletions have been counted as one element.

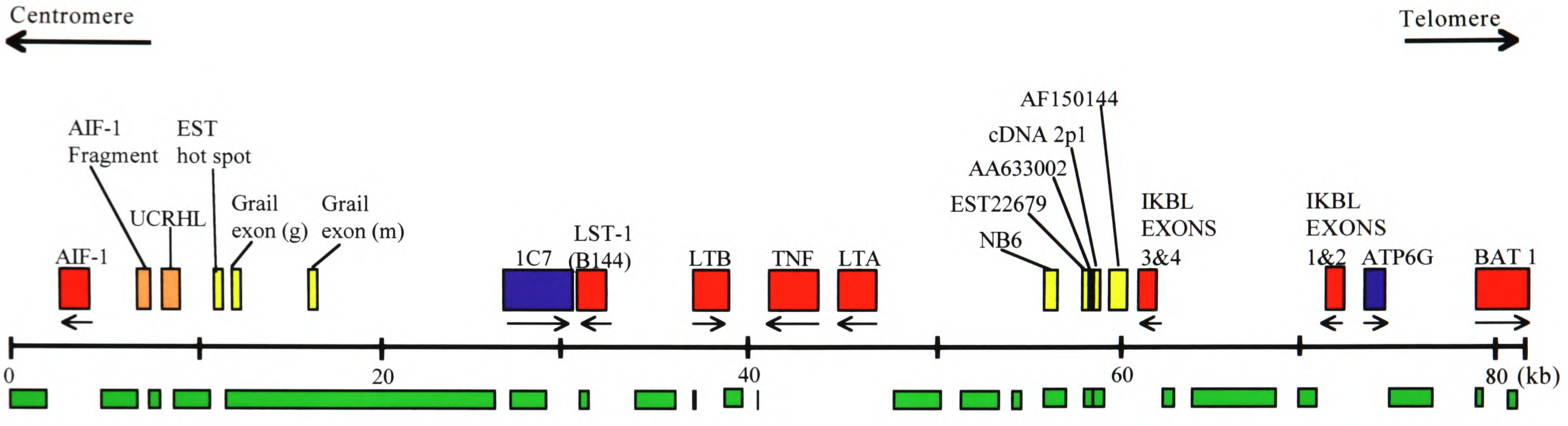


FIGURE 3.8 A map of transcripts and other features within the human genomic region covered by overlapping cosmids TN62 and TN82. The positions of the genes and other features are indicated by boxes and the directions of transcription, where relevent, are indicated by arrows. The gene exon coordinates are recorded in Table 3.7, including the AIF-1 gene fragment and the pseudogene UCRHL. The coordinates for additional potential coding regions, i.e. EST alignments and GRAIL exon predictions, not assigned to known genes, are recorded in Table 3.5. The coordinates of repeat elements identified by the RepeatMasker server are recorded in Appendix II.

TABLE 3.7 The exon co-ordinates for known genes, putative genes, pseudogenes and gene fragments found within the combined 81800 nucleotides of DNA sequence of overlapping Cosmids TN62 and TN82 (diagrammatically represented in Figure 3.8).

EXON DETAILS						GRAIL PREDICTION[f]		
Gene	**Start**	**End**	**Exon number**	**Orientation**	**Splice boundaries**	**Prediction**	**Confidence**	**Extent**
[a]AIF-1	2788	2583	6	c	AG-PolyA	√	e	5' 3'
	3260	3098	5	c	AG-GT	√	e	5' 3'
	3510	2583	4III	c	CC-GT	√	e	3'
	3592	3459	4II	c	5'end-GT	x		
	3500	3459	4I	c	AG-GT	√	e	5' 3'
	3934	3561	3II	c	AG-GT	√	e	5'
	3934	3868	3I	c	AG-GT	√	e	5' 3'
	4083	4022	2	c	AG-GT	√	g	5' 3'
	4348	4249	1	c	5'end-GT	x		
[b]G1 fragment	7649	7782	-	c	-	-		
[c] UCRHL	8182	8719	-	+	-	-		
[a]1C7	26642	26933	1	+	5'end-GT	√	e	3'
	29486	29830	2	+	AG-GT	√	e	5'
	29486	29637	2I	+	AG-GT	√	e	3'5'
	29713	29830	2II	+	AG-GT	√	e	5'
	29979	30086	3	+	AG-GT	√	e	3' 5'
	30224	30340	4I	+	AG-PolyA	x		
	30381	30711	4II	+	AG-PolyA	x		
	30436	30711	4III	+	AG-PolyA	x		
[d] LST-1	31049	30731	8II	c	AG-PolyA	x		
	31094	30731	8I	c	AG-PolyA	x		
	31668	31646	7	c	AG-GT	x		
	31797	?	6	c	AG-?	x		
	31971	31879	5	c	AG-GT	x		
	32412	32294	4II	c	AG-GT	x		
	32495	32294	4I	c	5'end-GT	x		
	32631	32496	3II	c	5'end-GA	x		
	32580	32496	3I	c	5'end-GA	x		
	32913	32805	2	c	5'end-GT	x		
	33378	33297	1	c	5'end-GT	x		

EXON DETAILS						GRAIL PREDICTION[f]		
Gene	**Start**	**End**	**Exon number**	**Orientation**	**Splice boundaries**	**Prediction**	**Confidence**	**Extent**
LTB	37185	37356	1	+	5'end-GT	√	g	3'
	37753	37798	2	+	AG-GT	x		
	37982	38053	3	+	AG-GT	√	g	5'
	38449	39054	4	+	AG-PolyA	x		
TNF	42496	41286	4	c	AG-PolyA	√	e	5'
	42845	42798	3	c	AG-GT	x		
	43079	43037	2	c	AG-GT	x		
	44051	43685	1	c	5'end-GT	√	e	5' 3'
LTA	46311	45291	4	c	AG-PolyA	√	e	5'
	46684	46579	3	c	AG-GT	√	g	5' 3'
	46878	46771	2	c	AG-GT	√	e	5' 3'
	47327	47166	1	c	5'end-GT	x		
IKBL	61589	60782	4	c	AG-PolyA	√	g	5'
	61983	61782	3	c	AG-GT	√	e	5' 3'
	71448	71172	2	c	AG-GT	√	e	5' 3'
	71985	71860	1	c	5'end-GT	x		
[a]ATPase	72763	73132	1I	+	5'end-GT	√	g	5' 3'
	72763	72978	1II	+	5'end-GT	x		
	73401	73501	2	+	AG-GT	√	e	5' 3'
	74029	74202	3	+	AG-PolyA	√	e	5' 3'
[e] BAT1	77595	77660	*(1)*	+	5'end-GT	x		
	78954	79288	1	+	5'end-GT	x		
	79076	79288	*(2)*	+	AG-GT	√	e	5'
	80336	80463	2 *(3)*	+	AG-GT	√	e	5' 3'
	80755	80847	3 *(4)*	+	AG-GT	√	e	3'

[a] For relevant details see Chapter 4 of this Thesis.

[b] G1 fragment is a duplicated fragment of the AIF-1 gene containing the G1 exon 1.

[c] UCRHL is a putative pseudogene of Ubiquinol-cytochrome C reductase hinge protein.

[d] For relevant details of LST-1 splicing see section 3.2.2.3.3.

[e] Only the first three exons from the database entry for BAT1 are represented in the sequence. The exons numbered in italic and parentheses are the exons corresponding to TIGR EST consensus THC247641 (see section 3.2.2.3.6).

[f] GRAIL prediction performance for each gene: '√' signifies GRAIL has predicted the exon and 'x' signifies it has failed to predict the exon. The 'Confidence' column represents the confidence with which GRAIL predicts the exon, 'e' signifies an excellent prediction, 'g' good, and 'm' moderate. The 'Extent' column represents the correct prediction of either 5', 3' or both splice sites..

regions (Figure 3.8), with the 17.2kb intergenic region between G1 and 1C7 having the highest density of repeats: 2.6 Alu sequences per kb.

In addition to the Alu elements fifteen further types of repeat sequence were identified by RepeatMasker, representing 38 repeat elements and 7.7kb of sequence (Table 3.6 and Appendix II).

3.2.2.6 Alignment of the Human and Mouse genomic sequences

Two murine (*Mus musculus*) genomic sequence database entries, accession numbers AF109719 and AC007080, overlap to give a contig of 151438 nucleotides in length. Contained within this contig is the region equivalent to the segment of the human MHC class III region sequenced in this Thesis. The murine AIF-1/BAT1 region is approximately 78500 nucleotides in length, i.e. just 3300 nucleotides shorter that its human counterpart, as reported here. Furthermore, the human genes identified in this segment of DNA are present in the mouse in the same order (Figure 3.9). The differences in sequence length appears to be due to differential repeat element expansion between the two species. This is most evident in the AIF-1 to 1C7 intergenic region which is 22617 nucleotides in man compared to just 7992 nucleotides in mouse. This is almost entirely due to Alu repeat element expansion in this segment of the human genome. The greatest size increase in the murine sequence, when compared to the human, is within the large intron 2 of IKBL, which is 10916 nucleotides in length in the human sequence and 15312 nucleotides in the murine sequence.

Interestingly, the 1C7 gene appears to be a pseudogene in mouse, or at least in the mouse strain sequenced. Since, although the predicted 1C7 amino acid sequences from man to

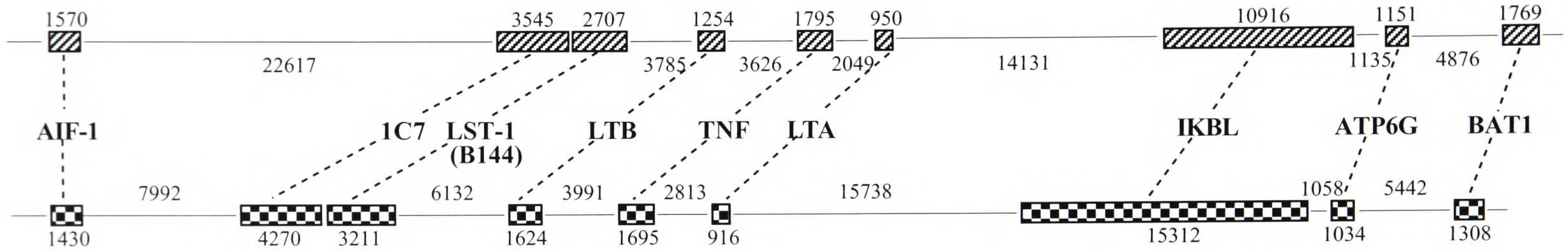


FIGURE 3.9 Comparison of the Human TN62/TN82 genomic region with the equivalent Murine genomic region.
The relative positions of the 9 genes identified within the TN62/TN82 genomic region, and their orthologs within the corresponding murine genomic region, are indicated by boxes. The upper line represents the human genes, whilst the lower line represents their murine equivalent. All numbers are in nucleotides. The numbers above or below the boxes indicate the genomic distance in nucleotides between the initiation AUG codon and the STOP codon for each gene (except for LST1(B144), where the entire gene coverage is used). The numbers above or below the lines in between the boxes indicates the intergenic distances, in nucleotides.

mouse share 64% identity, the murine sequence contains two in-frame stop codons. Also, the murine polyadenylation signal (AATAAG) deviates from the highly conserved consensus sequence (AATAAA) found in the human sequence. Furthermore, on searching the murine EST databases no ESTs could be identified corresponding to the murine 1C7 DNA sequence. However, in the rat databases, an EST (AA894045) representing a rat 1C7 homolog was identified. This rat sequence does contain the consensus polyadenylation signal found in human 1C7. However, the EST only covers the 3' half of the 1C7 sequence. Therefore, it is not possible to say whether the rat 1C7 homolog encodes a full length protein.

The putative pseudogene, UCRHL, found in the human AIF-1 to 1C7 genomic region (see section 3.2.2.4.1) could not be identified within the equivalent murine genomic region. However, the AIF-1 gene fragment (section 3.2.2.4.1) was present.

3.2.2.6.1 Alignment of the human and murine genomic sequences to aid the identification of potential coding regions in the human sequence.

No region with significant similarity to the NB6 cDNA sequence was found in the murine genomic sequence. This was also the case for: EST22679, AA633002 and the cDNA clone 2p1 (see section 3.2.2.4.2) which all border the region corresponding to NB6 in the human sequence. There were 3 regions of approximately 80% identity between murine and human sequence within the genomic region that shows similarity to EST AF150144 (see section 3.2.2.4.2). The co-ordinates of these three regions within the human TN62/TN82 sequence are: nt50477 to nt50717, nt50905 to nt50995, and nt51540 to nt51727. Also, an additional region at co-ordinates nt42226 to nt42365 was identified.

None of these regions showed significant similarity with any other sequences within the human or mouse databases.

Of the two regions with similarities to ESTs upstream of the BAT1 gene, the only region conserved in the murine sequence was that corresponding to the potential new exon 1 for BAT1 (as outlined in section 3.2.2.3.6).

3.3 DISCUSSION

3.3.1 Sequencing strategy

The sequencing of 1757 templates covering the 82kb region characterised here has resulted in an average depth of ~9.6 reads per base over the entire contig. The lack of any ambiguities over the 5341nt of overlap between the two cosmids sequenced is an indication of the high level of confidence that can be placed on the data generated. The use of different approaches for filling gaps in the sequence, covering single orientation reads and resolving ambiguities (see section 2.3.13.3) all incorporated a certain amount of replication into the data. Therefore, the average depth of sequence stated above is an elevated figure and the number of templates used to sequence future cosmids could be reduced by 10-20%.

The traditional method of resolving sequence compressions was to use Deaza GTPs or dITPs, where the chemical structure of the deoxynucleotides has been slightly modified in an attempt to overcome compressions in GC rich regions. However, it has been found here, and in other studies, that compressions and stops are more efficiently resolved by re-sequencing of templates using the dye-terminator chemistry. Therefore, by looking at the problem sequence with both dye-primer and dye-terminator chemistries, ambiguities can

be resolved. This is also the case for regions of single orientation read, where the re-sequencing of templates with the dye-terminator chemistry is cheaper, less time consuming and as informative as making a double stranded template to allow sequencing of the reverse orientation. It is now considered standard practice that from the onset of sequencing a shotgun library, ~50% of the templates are sequenced with dye-primer and 50% with dye-terminator chemistries, to minimise the finishing time.

The major problem encountered here, as with most sequencing projects, was the closing of gaps in the contigs. There appeared to be a disproportionate number of templates starting or finishing with polyA or polyT tracts, a feature associated with the ends of Alu repeats and other retrotransposable elements. As a result, gaps occurred most commonly within dense clusters of Alu repeats. The design of PCR primers to generate templates filling such gaps was extremely difficult and time consuming due to the highly repetitive nature of Alu regions. There are a number of theories as to why gaps are more frequent within Alu repeat regions. One is that the sonication is not entirely random and that relatively unstable polyA tracts are preferentially sheared, thus leaving the above mentioned gaps in the sequence. Another is that the sonication is random, but that the cloning into the M13 vector is not, due to the nicking of insert DNA, which is more likely to occur in Alu rich regions (because of the less stable polyA sequences). The best solution to this problem is to clone larger inserts into the M13mp18 vector, e.g. 1.5kb to 2kb, as compared to the 0.5kb to 1kb inserts used here, or to use a combination of insert sizes. This results in less efficient ligation reactions, but involves shorter sonication times and, therefore, causes less damage to the cosmid DNA, and makes it more likely that whole Alu elements will be cloned. In addition, the longer templates are more likely to bridge gaps than the smaller templates and then simply by sequencing from the other end of selected longer templates.

or by sequencing using internal primers, the whole template can be sequenced and gaps can be filled without the need for PCR.

3.3.2 Genomic structure, other genomic database sequences, and repeat sequences

3.3.2.1 Genomic structure

As stated in the results, the genomic structure of the known genes within the region sequenced have been precisely positioned in the order: G1/AIF-1, 1C7, LST-1, LTB, TNF, LTA, NB6, IKBL, BAT1 (centromere to telomere). In addition, a new gene, ATP6G (see Chapter 4) has been identified in the intergenic region between IKBL and BAT1 (Figure 3.8). Also, a duplicated fragment of the AIF-1 gene and a putative pseudogene, UCRHL, have been located in the AIF-1/G1 – 1C7 intergenic region. Discounting the pseudogene, the AIF-1 gene fragment and NB6 (section 3.3.5) the total number of genes in this 82kb region is now 9. This is equivalent to one gene per 8kb of DNA on average, making this among the most gene dense regions currently known in the human genome.

The average GC content of the AIF-1/BAT1 region is 49.6%. This figure falls between the average for the MHC class III region as a whole (53%), and that for the neighbouring MHC class I regions (45%). This is reasonable since the region sequenced here is on the boundary between the class III and class I regions.

3.3.2.1.1 The putative pseudogene UCRHL and a fragment of the AIF-1 gene

There appears to have been both an inter- and intra- chromosomal duplication event within the AIF-1 to 1C7 intergenic region. A region 3.3kb upstream of the AIF-1 gene shows 80% identity with exon 4II of AIF-1 (exon 1 of G1). Another region telomeric of this (Figure 3.8, Table 3.5 and Table 3.7) shares ~80% identity with ESTs and EMBL database

entries for the Ubiquinol-cytochrome-c reductase mitochondrial hinge protein (UCRH). This region appears to represent a UCRH pseudogene and has been named UCRHL. On comparing the human and murine genomic sequences (section 3.2.2.6) it is interesting to note that the AIF-1 gene fragment is present in both human and mouse genomes, whereas there is no evidence of a UCRHL ortholog in mouse. This suggests that the intrachromosomal duplication of exon 4II of AIF-1 most probably took place prior to the separation of the human and mouse lineages, whereas the insertion of the sequence corresponding to UCRHL took place after the separation of the two lineages. As the genomic region between AIF-1 and 1C7 has been the site of significant repeat expansion since the human/mouse separation, resulting in an increase in length from 8kb in mouse to 22.6kb in man (Figure 3.9), it is perhaps not surprising that, in addition to the insertion of repeat sequences, other sequences have been inserted. The AIF-1 fragment and UCRHL may well have been inserted along with repeat sequences into this unstable region. It is not possible to say at this stage whether UCRHL is the result of the insertion of a genomic region or of a spliced RNA species, as there is no published information on the genomic structure or location of UCRH. One final note, if UCRHL is found to be a pseudogene it would be a rare occurrence within the MHC class III region, indeed it would be one of only two or three pseudogenes so far identified in this region (including P450c21A and XA), unlike the MHC class I and II regions that contain many pseudogenes (see Figure 1.1).

3.3.2.2 Variation between the different genomic sequences within the databases

Comparison of the genomic sequences within the databases that overlap with the sequence generated here reveals an average level of sequence variation in each case of about 2

substitutions in every 1000 nucleotides. This is close to the previous estimated genome wide variation of about 1 nucleotide in 1000.

One of the next challenges for the Human Genome Project will be the understanding and mapping of sequence variation between different human genomes. As the number of genomic sequence database entries increases, the analysis of overlapping sequences for single nucleotide polymorphisms (SNPs) will become a major source of information for sequence variation databases. Also, as the technology improves in the field of high throughput and automated SNP screening, these SNPs will be invaluable in the characterisation of complex diseases such as IDDM and RA.

3.3.2.3 Conservation between Human and Mouse

The human and mouse genomes are well conserved throughout the MHC class III region. The only significant differences between the two species over the region examined here relate to the rate and location of repeat element expansion. Although the overall genomic regions of man and mouse are of a comparable size (81.8kb and 78.5kb, respectively) the focal points of repeat expansion differ. The significance of these difference, if any, is unclear, as they appear essentially to be random.

Comparison of the human and mouse genomic sequences will be referred to in a number of sections within this Thesis. As the sequencing of the human and murine genomes (as well as those of other model organisms) nears completion, such comparisons will not only shed light on the evolution of different species but will also aid in other areas, such as in the identification of coding regions and regulatory elements. In this investigation, interspecies sequence comparison has been a useful resource for confirming the existence of potential

coding regions. We have found that known amino acid coding regions within the sequence examined share approximately 80% to 85% identity between man and mouse, at the DNA level. This degree of conservation is rarely found within the non-coding regions. Therefore, comparison of orthologous sequences can add valuable information to the gene hunting arsenal in very much the same way that Zoo Blot analysis is used to confirm the integrity of cDNA inserts. In the case of the NB6 genomic region, the lack of a conserved region in mouse suggests that NB6 and the ESTs mapping in close proximity to it (section 3.2.2.4.2) do not represent a new gene or genes. This is supported by the inability to detect an RT-PCR product from these genomic segments in this study (see section 3.2.2.4.2). In contrast, evidence suggesting the existence of a novel exon 1 for the BAT1 gene in man is supported by the presence of a conserved region at the same position in the mouse genome (section 3.2.2.3.6).

3.3.2.4 Repeat sequences

Over a third of the human genome consists of interspersed repetitive sequences, consisting mainly of degenerate copies of transposable elements (Smit 1996a). These repeat sequences can be divided into four main groups: short interspersed nuclear elements (SINE), long interspersed nuclear elements (LINE), long terminal repeats (LTRs) and DNA elements.

SINEs are retroposons of between 100 and 300 nucleotides which have an internal RNA polymerase III promoter site, typically derived from a structural RNA such as a tRNA, and usually end in a A- or T- rich tail or a short simple repeat (Smit & Riggs 1995). They are found in most vertebrates and invertebrates and are among the most numerous repeat elements. In the mammalian genome SINEs are divided into two groups: mammalian-

wide interspersed repeats (MIRs) and Alu repeats. MIRs are found in all mammalian genomes, including marsupials and are thought to have amplified before the mammalian radiation. Alus on the other hand are only found in primates and are the most numerous retroposon repeat elements found in the human genome. Alu elements can be further subdivided into distinct lineages. The first Alu progenitors are thought to be the free-arm monomers (FAM), derived from 7SL RNA: from these the free right arm Alu monomers (FRAM) and free left arm Alu monomers (FLAM) evolved (Quentin 1992). These gave rise to the first progenitors of the classic dimeric Alu elements by the fusion of a FLAM with a FRAM element. The dimeric Alu elements, which occur throughout the human genome, are represented by the oldest J Alu elements, the more modern S Alu elements and the youngest Y elements.

The LINE repeat elements, which are between 6 and 8kb in length and make up at least 15% of the human genome (Smit & Riggs 1995), were active before marsupials divided for other mammals. LINE elements encodes a reverse transcriptase and other proteins necessary for retroposition plus, in some cases, an internal polymerase II promoter. Like the SINEs, they contain a variable length insertion duplication site and a poly-A or simple repeat tail. It has been suggested that SINE mobility is dependent on proteins encoded by LINEs, furthermore, SINEs and LINEs do appear to be equally dispersed throughout the human genome (Smit & Riggs 1995).

LTR elements are retrovirus-like elements of between 300 and 1000 nucleotides which carry their own transcriptional regulatory sequences. The human genome contains a variety of endogenous retroviral elements, estimated to total about 50,000 copies (Smit & Riggs 1995). LTRs are members of a family of retroviruses called HERV-L which are

responsible for the many polymorphic CA dinucleotide repeats found within the human genome. Other LTR lineages include: MaLRs (mammalian LTR-retroposons) some of the MER (medium reiterary repeats) elements (e.g. MER21), and other retroviral elements. In total the LTR elements make up about 5% of the human genome.

Lastly, about 1.5% of our genome consists of DNA transposable elements, characterised by short inverted terminal repeats. These elements move by excision and reintegration (Smit & Riggs 1996b) with the aid of an internally encoded transposase. The most common of these are the Mariners, but a number of MERs (medium reiterary repeats) have also been found to be DNA elements.

One striking feature of the genomic region sequenced in this Thesis is the distribution and high density of Alu repeats (also noted by Iris *et al.* (1993)) and other retroposons within non-coding regions, to the extent that any significant breaks in the clusters of repeats are likely to contain exonic sequences (highlighted in Figure 3.8). The high density of Alu repeats in this region support the suggestion, based on fluorescent in-situ hybridisation (FISH) data, that GC rich, Giemsa light bands (such as 6p21.3) are in general Alu rich (Korenberg & Rykowski 1988).

The telomeric end of the MHC class III region contains examples of the early Alu progenitors, (FAMs, FLAMs, and FRAMs), as well as the oldest of the dimeric lineages (the J-Alu subfamily) (Appendix II). This suggests that the invasion of Alu elements into this region of the genome was an ancient event. Furthermore, the ratio of S-Alu(new) to J-Alu(old) in this segment of DNA is 2.8, which is close to the global human genome S/J ratio of 3.00 (Jurka & Smith 1988). This indicates a random distribution of S-Alu and J-

Alu sequences within this segment of the MHC class III region. In addition, the apparent random mixing of the old J-Alu subfamily with the new S-Alu and Y-Alu subfamilies within the dense clusters of Alu repeats adds weight to the suggestion that new insertions may preferentially target existing Alu repeats (Slagel *et al.* 1987), i.e. as new Alu families have appeared they have been inserted into regions already rich in Alu. This would go some way to explaining the uneven expansion of non-coding regions, for example the huge expansion within the human AIF-1 to 1C7 intergenic region which has expanded by over 300%, when compared to mouse, since the human and rodent lineages divided (see section 3.3.2.3) as compared to the LTA to IKBL intergenic region that has remained roughly the same size between mouse and man (15.9kb and 14.1kb, respectively) (see Figure 3.9).

3.3.3 Alternative splicing of the Leukocyte specific transcript-1(LST-1) gene

The LST-1 gene is the recently identified human homolog of the mouse B144 gene (Holzinger *et al.* 1995), which was mapped to the equivalent position in the mouse genome about 10kb centromeric of TNF in the H2S:H2D interval (Tsuge *et al.* 1987). B144 appears to be specifically expressed in B cells and macrophages and encodes a mRNA of ~800 nucleotides. However, no protein product has been isolated and the published 400 nucleotide cDNA of B144 is incomplete at the 5' end (accession no M18187).

Work on the human homolog of B144, LST-1 by Holzinger *et al* (1995) and de Baey *et al* (1996) has shown that, like B144, LST-1 is predominantly expressed in monocytes and also at lower levels in some T-cells (de Baey *et al.* 1996). In monocytes, expression is induced by IFN-γ: an IFN-γ-activated site and an IFN-stimulated gene factor-2 responsive element have been located within the last intron of LST-1 (Holzinger *et al.* 1995), suggesting an immune related function. Five alternatively spliced forms of the LST-1

mRNA had been previously reported which potentially encode four different protein isoforms of 97, 66, 54 and 28 amino acids. These were reported to be either globular or membrane anchored, depending on the splice variant (de Baey *et al.* 1996). The globular form was reported to be predominantly expressed in IFN-γ induced monocytes while the membrane-bound form was predominantly expressed in stimulated T-cells. However, despite this there is still no clear picture as to the actual amino acid sequence of LST-1 and no protein has been isolated *in vivo*.

The findings reported here further support the existence of multiple splice forms of LST-1 and potentially increase the number of isoforms occurring, from four to seven. This was determined by comparison of our genomic sequence with EST database entries and highlights the wealth of data available from these sources. Shortly after our findings were published (Neville & Campbell 1997), de Baey *et al* (1997), experimentally confirmed many of these splice variants by cloning RT-PCR and RACE products of LST-1. de Baey and co-workers (1997) also identified an additional first exon, which was not found in any of the EST database sequences or predicted by GRAIL and an additional last exon, differing from exons 8I and 8II in Figure 3.4 at its 5' splice site only. In total, de Baey and co-workers (1997) identified 9 splice variants (LST1/A to I). Their LST1/D corresponds to LST1-3 here, LST/G to LST1-5, LST1/H to LST1-6, and LST1/I to LST1-4. Transcripts equivalent to LST1-2 and LST1-7 were not identified by de Baey *et al* (1997) nor was the original LST-1 database entry, U00921. The splice variants LST1/A, LST1/B, LST1/C, LST1/E and LST1/F, identified by de Baey *et al* (1997) were not identified in this investigation from the EST databases. However, all of the open reading frames identified by de Baey *et al* (1997) are represented in the splice variants identified from the EST databases and LST1-2 and LST1-7 (if the latter is a real splice form) encode isoforms not

identified by de Baey *et al* (1997). Despite the large number of putative splice variants of LST-1, the fact still remains that no protein product has as yet been isolated either *in vitro* or *in vivo* and, until it is, no assumptions can be made as to how many of these splice variants are functional.

The elucidation of the amino acid sequence of LST-1 is further complicated by the findings reported here (Figure 3.6) and by Holzinger *et al.* (1995) and de Baey *et al.* (1996), i.e. that more than one potential open reading frame exists. At the DNA level the murine B144 cDNA and the human LST-1 cDNA show 63 % identity. However, at the protein level different LST-1 open reading frames (ORFs) show similarity with different regions of the published predicted B144 amino acid sequence, i.e. there is a frame shift within the B144 sequence. Holzinger and co-workers (1995) suggested that ORF 1 (Figure 3.6B) is most likely to be translated as it follows the first AUG rule and is, therefore, situated in the most favourable context for translation initiation. This reading frame also shows the longest region of similarity with the predicted B144 amino acid sequence over the putative transmembrane region. To confuse matters, however, the LST-1 amino acid sequence in the PIR database (PIR:g1127546) discussed by Holzinger, *et al.* (1995) is that of ORF 2 (Figure 3.6B). This also shows similarity with the B144 amino acid sequence, but over a different region to the predicted product of ORF1, and is predicted as a globular protein. Lastly, it has been found here as in previous investigations (Holzinger *et al.* 1995; de Baey *et al.* 1996; de Baey *et al.* 1997) that none of the LST-1 polypeptides share significant overall similarity with other proteins in either the Genbank, Swissprot, PDB or PID databases.

With the aim of further clarifying the sequence of the murine B144 transcript for comparison with LST-1 we have searched the published B144 cDNA sequence (Tsuge *et al.* 1987) against the murine EST database to identify any B144 transcripts. We have also compared it to the recent murine genomic sequence database entry covering this entire region (Accession number AF109719). Figure 3.10 shows a multiple alignment of all the B144 ESTs identified with the published B144 sequence (mmb144a in Figure 3.10) and with the assembled exonic regions from the murine genomic sequence (m6282mm-b144 in Figure 3.10). There are a number of sequence differences between the B144 ESTs and cDNA sequences, the most significant of which is a 'G' deletion in the published B144 sequence at nucleotide position 194 (Figure 3.10) (also identified by de Baey *et al.* 1997). This single deletion in the published DNA sequence causes a frame-shift in the B144 amino acid sequence at the point where the LST-1 amino acid sequence of ORF 1 (Figure 3.6) and the published B144 amino acid sequences deviate from each other and is also the point where the alignment with ORF2 starts. The consensus sequence generated from the alignment in Figure 3.10 encodes a single long open reading frame that encodes a polypeptide that matches the ORF1 of the human LST1 splice variant LST1-5 over its entire length with 56% identity (Figure 3.11). The fact that the corrected mouse sequence has now been shown to have a single open reading frame that corresponds to the first open reading frame in the human sequence seems to support the statement by Holzinger and co-workers (1995) that ORF1 is the most likely ORF and that LST-1 is a membrane bound protein.

Interestingly, the mouse sequences available seem to represent two different allelic variants of B144. Recent work by de Baey et al. (1997), which also identified the discrepancy in the original B144 database entry included the cloning of two different splice

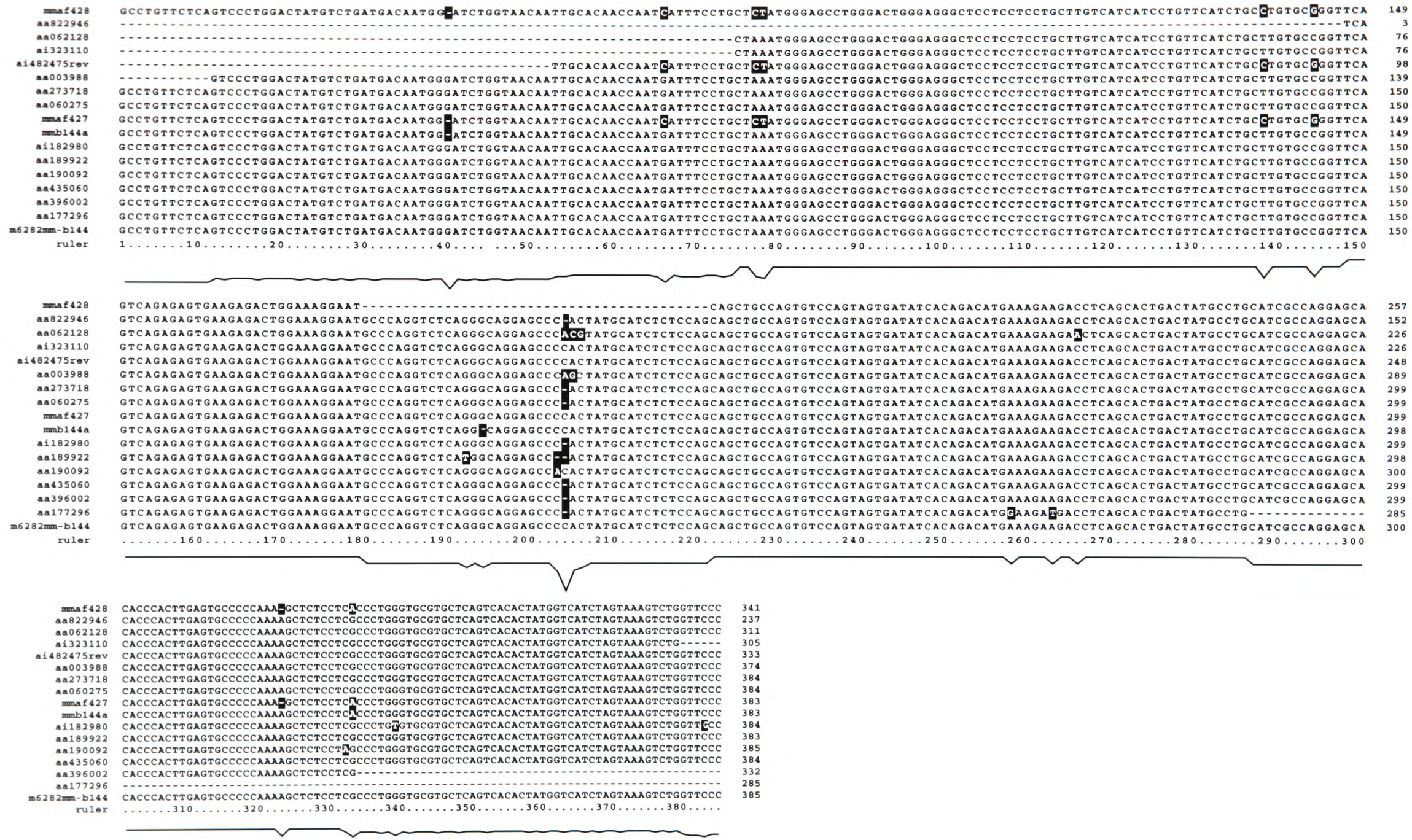


FIGURE 3.10 Multiple alignment of the Murine B144 gene expressed sequences held within the databases. Sequence mmb144a is the original B144 database entry, accession number M18187; sequences mmaf427 and mmaf428 are the cDNA clones published in de-Baey, *et al* (1997); sequence m6282mm-b144 is taken from the Murine genomic sequence database entry, accession number AF109710; all other sequences are ESTs from the Murine EST databases. Black boxes represent nucleotide differences and the line below the multiple alignment indicates alignment quality.

variants of the murine B144. These are represented by mmaf428 (accession number AF000428) and mmaf427 (accession number AF000427) in Figure 3.10 and, together with EST AI482475, differ from the consensus in Figure 3.10 (m6282mm-b144) at positions 66 (G/C), 77-78 (AA/CT), 138 (T/C) and 144 (G/C), as well as having a 'G' deletion at position 40. These differences cause three internal amino acid changes (Figure 3.11). The deletion/insertion at nucleotide position 40 (Figure 3.10) also results in the use of different initiation AUG codons in the different variants due to a frame shift. The amino acid sequences of the two identified variants of B144, together with the human LST-1 (LST1-5) ORF1 amino acid sequence are shown aligned below in Figure 3.11.

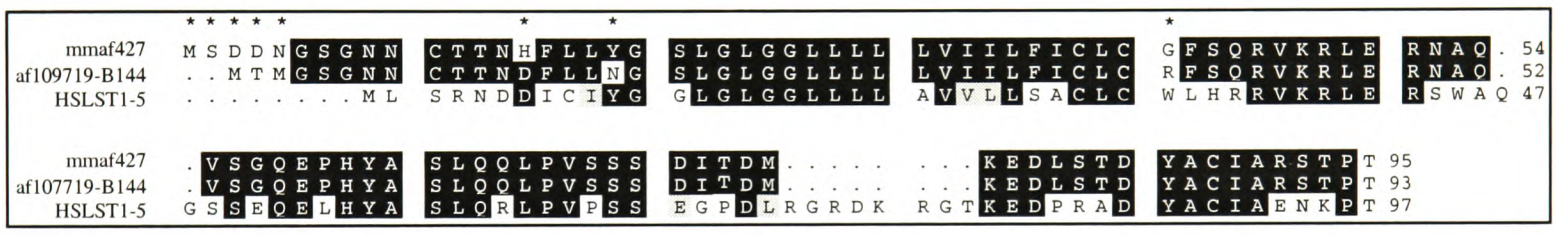


FIGURE 3.11 Multiple alignment of the amino acid sequence of the two identified variants of murine B144 with the first open reading frame (ORF1) of human LST1-5 splice variant, as presented in Figure 3.6 (section 3.2.2.3.3). mmaf427 is the B144 clone m13r, published by de Baey, *et al* (1997), af109719-B144 is the deduced amino acid sequence from the murine genomic database entry AF109719 (as in the sequence m6282mm-b144 in Figure 3.10) and HSLST1-5 is the human homolog of B144 as encoded by ORF1 of the LST1-5 splice variant. The black boxes represent identity and the grey boxes similarity. The asterisks (*) above the sequence highlight the positions of the amino acids that differ between the two murine variants (mmaf427 and af109719-B144).

The absence of any experimental data on a protein product for LST-1 makes the confident assignment of a putative amino acid sequence a lot more difficult for this gene than for its mouse homolog. As shown in Figure 3.6, in contrast to the mouse B144 cDNA, there is no single dominant open reading frame for LST-1. In fact, there is a potential open reading frame in all three frames of the DNA sequence. If one looks at the relative positions and consensus sequences around the four different candidate initiation AUG codons, however,

ORF3 and ORF4 are unlikely to be strong contenders. ORF4 starts 71 nucleotides downstream of the first ORF. Kozak (1995) has shown that potential initiation AUG codons such a long distance from the 5' proximal start codon, no matter how poor the consensus is around the first AUG, are rarely used. ORF3, although only 21 nucleotides downstream of the first potential start site, shows no conformance to the Kozak consensus sequence other than the invariant AUG triplet (Kozak 1984) (Table 3.8) and is also unlikely to be used. The determination of a consensus sequence around the initiation AUG codon has been the focus of considerable work over the last 2 decades, (Sargan *et al.* 1982; Kozak 1984; Kozak 1986; Cavener 1987; Kozak 1991; Iida & Masuda 1996). Kozak (1984) proposed an optimal sequence for translation initiation of: GCC(G/A)CCAUGG, around the invariant AUG codon (underlined). This has largely been supported by a later study of using 699 eukaryotic sequences (Kozak 1991) and using quantitative methods (Iida & Masuda 1996), although there is deviation from this consensus in many transcripts to a greater or lesser degree. The most highly conserved residue is at the –3 position (i.e. 3 nucleotides upstream of the AUG codon), where there is a purine (usually A) in 97% of the eukaryotic genes tested by Kozak (1991). Translation initiation is negatively affected by a substitution at this position. The +4 'G' nucleotide is conserved in 46% of eukaryotic genes, although the importance of this residue in translation initiation has been questioned (Cavener 1987). Of the 23 genes included in the database entry AF129756 (excluding LST-1) which covers most of the MHC class III region (although not all the genes are annotated) only two of the genes, LTA and IKBL, have neither the –3 purine or the +4 'G'. It is important to note though that neither of these two genes have more than one significant open reading frame, therefore, the weak consensus sequence is not in competition with another initiation AUG. However, the deviation from the consensus may well reduce the translation efficiency of such genes. The percentage of the 23 MHC class

III genes with the –3 purine is 78% in contrast to the 97% found by Kozak (1991) and most have 'G' rather than 'A'. However, the sample size here is much smaller than that used by Kozak: n=23 compared to n=699.

Of the two most likely open reading frames for the LST-1 splice variants (i.e. ORF1 and ORF2), that with the downstream AUG (ORF2) fits the Kozak consensus best (see Table 3.8) having the key -3 purine residue. In contrast, ORF1 has neither the -3 or +4 consensus residues. The second initiation AUG of the putative variant LST1-3, conforms most strongly to the consensus (Table 3.8), having both the key residues, a purine at -3 and G at +4, due to its splice pattern. Use of this initiation codon would, however, violate the first AUG rule. This 'leaky translation' is rare, but not unheard of and examples include a large number of critical regulatory genes such as growth factors and cytokines (Kozak, 1991). It has been shown that initiation at a second, downstream open reading frame can occur if the first start site has a less than optimal consensus sequence (Kozak, 1995), as is the case here, but only if the two AUG codons are in close proximity. In the case of the findings by Kozak (1995) the two AUG codons were only four nucleotides apart, whereas in LST-1 they are fourteen nucleotides apart.

For two of the alternatively spliced transcripts, LST1-4 and LST1-5, the downstream open reading frame (ORF2) although having the strongest adherence to the Kozak consensus sequence results in a significantly shortened polypeptide (Figure 3.6, Table 3.8) due to an in-frame stop codon at the start of exon 7. For the same reason, LST1-6 has no ORF2. It is interesting to note that LST1-6 is the most highly represented in the EST database (Table 3.4). This suggests that ORF1 of LST1-6 may be functional despite having a weak translation initiation signal.

TABLE 3.8 Details of the LST-1 gene open reading frames.

Splice Variants	ORF1			ORF2			ORF3			ORF4		
	length[a]	state[b]	Translation[c] initiation seq	length	state	Translation initiation seq	length	state	Translation initiation seq	length	state	Translation initiation seq
LST1-db	59	TM	G.C...**AUG**.	84	S	G..**A**..**AUG**.	103	S	.....**AUG**.		-	-
LST1-2	44	TM	G.C...**AUG**.	69	S	G..**A**..**AUG**.	88	S	.....**AUG**.		-	-
LST1-3	28	S	G.C...**AUG**.	53	S	G..**A**..**AUGG**.		-	-	63	S	.CC**A**C.**AUG**.
LST1-4	82	TM	G.C...**AUG**.	32	S	G..**A**..**AUG**.	44	S	.....**AUG**.		-	-
LST1-5	97	TM	G.C...**AUG**.	32	S	G..**A**..**AUG**.	59	S	.....**AUG**.		-	-
LST1-6	66	S	G.C...**AUG**.	-	-	-		-	-	63	S	.CC**A**C.**AUG**.

[a] Length is in amino acids.
[b] TM: Protein with predicted transmembrane region, S: Protein predicted to be soluble
[c] The translation initiation sequence consensus is based on the Kozak (1991) consensus sequence GCC(A/G)CCAUGG. Only the bases conforming to this consensus around the initiation codon are shown with other positions indicated with dots (.). The most important bases are underlined and in bold.

It is clear from the literature that the Kozak consensus sequence is not always strictly adhered to. However, the important point here is not the extent to which the sequence around the start codon conforms to the consensus, although this may well effect translation efficiency, but the existence of another reading frame in close proximity with a stronger initiation consensus. Evidence does suggest that the first open reading frame of LST-1 is a good candidate as it agrees with the mouse open reading frame, it follows the first AUG rule, and LST1-6 which does not encode ORF2 is the most common splice variant in the EST databases. However, the close proximity and more favourable context of the second ORF means that this cannot be ignored, especially in light of evidence that leaky translation does occur under conditions not unlike these, where the first AUG deviates from the consensus at the –3 or +4 positions. It is possible then that the LST-1 mRNAs are bifunctional, encoding two different proteins, with the membrane bound ORF1 translated at a much lower efficiency than the soluble ORF2 due to its weaker translation initiation signal.

To complicate matters further we found, as did Holzinger *et al* (1995), that the LST-1 transcripts have upstream 'mini-cistrons'. All splice variants have one decapeptide mini-cistron, which terminates 67 nucleotides 5' of the ORF1 initiation AUG, but has very weak agreement with the Kozak translation initiation consensus. LST1-5, LST1-6 and LST1-7 have an additional open reading frame terminating just 11 nucleotides 5' of the ORF1 initiation AUG. This mini-cistron encodes a 38 amino acid polypeptide, also having a very weak agreement with the Kozak translation initiation consensus. One would imagine that, despite having weak Kozak sequences, such elements may reduce the efficiency of translation of LST-1 if not completely block it, even if they are far enough away to allow re-initiation. This is supported by increasing evidence in the literature (Namen *et al.* 1988; Mueller *et al.* 1990; Waterhouse *et al.* 1990; Arrick *et al.* 1991; Kozak 1991) relating to proteins that include proto-oncogenes, transcription factors, signal transduction factors and many proteins involved in the immune response. One theory is that it would be disadvantageous if such regulatory proteins were translated at high levels and that the upstream 'mini-cistrons', when present, act as translational down-regulators.

The complicated repertoire of AUG-burdened 5' sequences in LST-1 may result in a range of translational efficiencies, thus regulating the expression of the different splice variants. It may also be that these transcripts represent precursor mRNAs. There is evidence for the occurrence of incompletely spliced transcripts in lymphocytes (Froussard *et al.* 1988; Kozak 1991) and it has been found that post transcriptional processing improves following lymphocyte activation (Cohen *et al.* 1990) suggesting that these precursor mRNAs are translated at only very low levels until they are needed.

In conclusion, these results are a good example of the amount of information that can be generated from the analysis of genomic DNA sequence alone, making use of the increasing number of ESTs being entered into the databases. However, this approach has its obvious limitations. No assumption can be made regarding the completeness of the EST sequence without further cDNA sequence analysis. Even with a complete open reading frame, informative untranslated regions may be missing. Neither can it be assumed that all expressed transcripts encode functional proteins. Also, some of the EST entries may represent incompletely spliced transcripts which could encode cryptic peptides. With the confusing array of potential open reading frames no firm conclusions regarding the encoded protein sequence of LST-1 can be drawn and any extrapolation to biological function is tentative without further experimentation. However, the inducibility of LST-1 by IFN-γ, its expression pattern and its possible post-transcriptional regulatory strategies suggests that LST-1 may have a role in the immune response. Also, the potential for tight regulation of the LST-1 gene, together with the small size of any encoded polypeptide and its relative proximity to the TNF gene cluster, adds strong weight to the hypothesis that this gene may encode a novel immune regulatory protein. Lastly, LST-1 presents a very interesting candidate for further study in relation to post-transcriptional events such as alternate splicing of the pre-mRNA, initiation AUG usage and the relevance of 5' burdened sequences.

3.3.4 A novel splice variant of the BAT1 gene

The BAT1 gene encodes a putative nuclear RNA helicase of the DEAD family of ATP-dependant RNA helicases. Members of the RNA helicase family are involved in diverse cellular functions such as initiation of translation, RNA splicing, ribosome assembly, spermatogenesis, cell growth and cell division.

BAT1 covers ~10kb of genomic sequence, is divided into 10 exons and encodes a protein of 428 amino acids. It is highly conserved between different organisms: human BAT1 shows 95.6% identity with the porcine BAT1 gene (Peelman *et al.* 1995) and orthologs in rat (p47), *Drosophila* (WM6) and *C. elegans* (Ce08102) have been suggested (Nair *et al.* 1992; Peelman *et al.* 1995). Of the 10 exons constituting the BAT1 cDNA sequence (accession number Z37166) only the first three are represented in the sequence generated within the course of this project (Figure 3.7, Figure 3.8 and Table 3.7).

The TIGR database EST entry THC247641 appears to represent a novel splicing event at the 5' end of BAT1. Comparison with the genomic sequence suggests the presence of an additional exon 1416 nucleotides upstream of the previously identified BAT1 transcription start site. This exon is 480 nucleotides in length and is a consensus sequence of the overlapping ESTs: AA326097 and T60343. Of particular interest is that this new exon 1 falls within the GC-rich region predicted to be a CpG island and seems to be non-coding. Such a GC-rich 5' untranslated sequence would most likely be highly structured and may throttle translation, thus presenting a possible means of regulating gene expression (Kozak 1991). Examples of genes with GC-rich 5' UTRs include: growth factors such as IGF-II (Irminger *et al.* 1987), transcription factors such as NF-κB (Meyer *et al.* 1991), signal transduction factors components such as cAMP-dependent PKCα (Maldonado & Hanks 1988), receptors for ligands such as oestrogen (Green *et al.* 1986) and proto-oncogenes such as *int-2* (Grinberg *et al.* 1991). Many of the above genes are known to be expressed at low levels or would be expected to need tight regulation. In a study by Kozak (1991) 699 mRNA sequences were examined (see also section 3.3.3), of which only 19% contained GC-rich 5' UTRs, 40% of these fell into the categories mentioned above and a further 30% derive from cytoskeletal or housekeeping genes (Kozak 1991). Among these

GC-encumbered genes some were found to express alternative mRNAs lacking the GC-rich 5' UTRs, by utilising different promoter elements and thus different exonic sequence at their 5' ends. It is possible to imagine that this could also be the case for BAT1: the non-GC encumbered 5' UTR of database entry Z37166 using one promoter and the splice variant found with EST entry THC247641 using a promoter further upstream, and thus splicing in an additional, and GC-rich, exon (Figure 3.7). The latter would be expected to be translated less efficiently due to its, hypothesised, highly structured 5' UTR interfering with the translation machinery. The reduced efficiency of translation of mRNAs with GC-burdened 5' UTRs has been shown experimentally in a number of cases (Kobilka *et al.* 1987; Dixon *et al.* 1989; Leung *et al.* 1989; Muller & Witte 1989; Manzella & Blackshear 1990). If this novel BAT1 exon does exist in the context of a mature mRNA *in vivo* and, if it does have an effect on translation efficiency, this raises a question as to the significance of CpG islands being commonly found in close proximity to the 5' ends of housekeeping genes; genes which are generally recognised to be expressed at low levels (Kozak 1991).

In addition to THC247641 four other ESTs aligned upstream of the database BAT1 sequence (Figure3.7) None of these show any evidence of splicing and all are of a relatively large size when compared to the BAT1 exons. It is most likely that these are either incompletely spliced fragments of BAT1 or genomic contamination. This is further supported by the lack of conservation over these regions between the human and mouse genomes, except for the region corresponding to the novel exon 1 of BAT1 (section 3.2.2.6.1). This, yet again highlights the problems of using EST data alone. As the number of EST database entries grows it may be found that an increasing proportion of them represent precursor mRNAs or incompletely spliced fragments. Therefore, the above

hypothesis on BAT1 splicing and regulation must be considered with caution in the absence of experimental data.

3.3.5 The hunt for potential new coding regions

As outlined in section 3.2.2.4, three regions have been the focus of intensive analysis in an attempt to identify further potential coding regions, i.e. AIF-1-1C7, LTA-IKBL and IKBL-BAT1 intergenic regions. This analysis has resulted in the identification of a fragment of the AIF-1 gene and the putative pseudogene UCRHL in the AIF-1-1C7 intergenic region, together with the identification of an additional exon and thus the elucidation of the complete 1C7 gene sequence (see Chapter 4 of this Thesis). Analysis of the IKBL-BAT1 region has led to the identification of an ATPase sub-unit (see Chapter 4 of this Thesis) and an additional exonic region associated with BAT1. Further EST alignments or exon predictions within the AIF-1C7 region and IKBL-BAT1 intergenic regions have either fallen within repeat sequences, have no other supporting evidence, or are not conserved between the human and murine sequences and subsequently have all be discounted.

The LTA-IKBL intergenic region encompasses the NB6 DNA segment (Albertella *et al.* 1996). NB6 was first identified by Zoo blot analysis and was added to the transcript map of the MHC class III region as the positive Zoo blot result suggested that it was an evolutionarily conserved sequence and thus was thought to be part of a novel gene. However, no signal was ever detected on a Northern blot for this sequence (Albertella *et al.* 1996). Initial analysis hinted at a novel coding region encompassing NB6. Figure 3.8 clearly shows a number of potential coding sequences within this region, represented by ESTs or cDNAs, in addition to two GRAIL predicted exons (Table 3.5). Also, within the genomic region that aligns with EST AF150144 there are a number of regions conserved

between the human and mouse. However, despite the tantalising evidence there is no consensus of information to indicate the presence of a novel gene in this region. Apart from three small segments within EST AF150144 none of the EST sequences are conserved in mouse and this is also the case for NB6. Furthermore, none of the ESTs show any intron/exon structure or align with predicted exons and there is no consensus between the different prediction methods incorporated into the NIX interface (Figure 3.12). Also, many of the ESTs and cDNAs that align with this region contain Alu or other repeat sequences and no signal could be detected by RT-PCR. The lack of an RT-PCR product does not rule out the existence of a gene in this region. It may be very specifically or transiently expressed and thus would be unlikely to be detected in the experiments undertaken. An example of this comes from our analysis of the IRT-1 transcript, a splice variant of AIF-1 (discussed in Chapter 4) that we were unable to detect in RT-PCR experiments, but which has been shown to be transiently expressed (Autieri & Agrawal 1998). Also, the lack of a consensus between the exon prediction software does not rule out the existence of a coding region: the software was also unable to detect any exons in the LST-1 gene. However, the combination of all the negative evidence presently available suggests that these EST and cDNA alignments are most probably genomic contamination within the EST databases and cDNA libraries. Therefore, we conclude that with the resources available at present there is no evidence to suggest further coding regions within the sequence analysed.

3.3.6 Tools and databases: Gene hunting strategies summarised.

3.3.6.1 The use of EST databases.

The results presented here are an example of the amount of information that can be generated from database interrogation of genomic DNA sequence alone, making use of the

increasing number of ESTs being entered into the databases. The identification of a putative V-ATPase subunit, the completion of the 1C7 coding sequence and the identification of multiple splice variants of LST-1 have all been significantly aided by alignments with EST sequences. However, this approach has its limitations. No assumption can be made regarding the completeness of the EST sequence without further cDNA sequence analysis. Even with a complete open reading frame, informative untranslated regions may be missing. Neither can it be assumed that all expressed transcripts represent functional proteins. Also, some of the EST entries may represent incompletely spliced transcripts which could encode cryptic polypeptides and many may turn out to be genomic contamination. Therefore, this approach to sequence analysis needs to be backed up with more rigorous expression studies before any conclusions can be drawn.

3.3.6.2 Exon prediction analysis.

By using the positions of known exons as a positive control, the results presented in Table 3.7 highlight the value of exon prediction programmes such as GRAIL, in the search for new genes. With the exception of LST-1 that had no exons predicted, GRAIL was successful in identifying the existence and general position of all the known genes and correctly located at least 50% of the exons of each gene. Therefore, its value is obvious, but its inability to predict LST-1 also shows it is not infallible. It is perhaps not surprising, however, that the program failed with LST-1 in view of the complicated repertoire of alternatively-spliced transcripts (section 3.3.3). One would imagine that for such a large number of splice variants to exist there would need to be a certain amount of ambiguity in the splice site recognition sequence. Therefore, GRAIL may not be able to predict such exonic regions. The accuracy of GRAIL will improve, however, as more genes are entered

into the databases, as it is through these entries that its neural network learns what constitutes an exon. Indeed, its ability to predict correctly intron/exon boundaries has improved, even during the course of this project. For example, on the first submission of the IKBL gene region to the GRAIL server, the general position of the gene was predicted, but in the wrong orientation and with no splice junctions correctly positioned. With no changes to the sequence, a second submission of the same region to GRAIL, 6 months later, came back with IKBL predicted in the correct orientation, with three out of the four exonic regions located, two of which had correct splice site predictions.

A major problem with exon prediction programs is the identification of false positives and this is where the value of the consensus approach such as in NIX, can be seen. In Figure 3.12, exon prediction outputs, EST alignments and the protein database BLAST outputs from the NIX software are graphically displayed. All of the exon prediction methods used in NIX generate a certain amount of false positive and false negative data. In general though, the existence of a gene becomes obvious when comparing all the methods together, visually, especially when combined with EST data. Clusters of positive predictions from the majority of the different programs are only apparent at the positions of known genes, highlighted with red boxes in Figure 3.12.

In summary then, it is clear that no single method can be relied on for the confident identification of new genes. Instead, a combined approach is best adopted to compensate for the limitations of each method. A good example of this is the identification of the putative V-ATPase G subunit (Chapter 4), where a combined analysis using GRAIL, EST alignments and homology searches led to the localisation of this novel gene.



FIGURE 3.12 The NIX graphical output for the TN62/TN82 genomic sequence.
NIX is a WWW based tool (http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/) that combines many sequence analysis and gene prediction programs into a single interface. Only the gene prediction outputs (Fex, Hexon, MZEF, Genemark, GRAIL, Genefinder) and EST, swissprot and trembl BLAST search results are displayed in this figure. The positions of known genes are indicated with red boxes. All annotation above the scale relates to the forward DNA strand and all annotation below the scale relates to the reverse DNA strand.

Chapter 4

CHAPTER 4: ANALYSIS AND CHARACTERISATION OF THE 1C7, AIF-1 AND ATP6G GENOMIC REGIONS.

4.1 INTRODUCTION

Computer based sequence analysis and database interrogation represent informative initial steps in the characterisation of genes and the protein products they encode. As software technologies improve these approaches are becoming increasingly powerful. However, extensive experimental work is needed to follow up any initial findings.

Three regions within the TN62/TN82 sequence generated in this Thesis (Chapter 3) were of particular interest. These were: the AIF-1 genomic region, encoding a multiply spliced candidate disease susceptibility gene; the 1C7 genomic region encoding a potential new member of the immunoglobulin superfamily; and ATP6G, a novel ATPase subunit identified in this Thesis. This chapter focuses on the further characterisation and expression analysis of these three genomic regions to assess whether transcripts derived from them can be detected at the RNA level, and to identify any sequence features that may help to define the functions of the encoded proteins. Therefore, in addition to the computer based analysis of these regions, their expression pattern was evaluated by the reverse transcriptase polymerase chain reaction (RT-PCR) across a number of cell lines of both immune and non-immune origin (as described in section 2.3.10).

4.2 RESULTS

4.2.1 The G1/AIF-1 genomic region

2kb of sequence encompassing the G1/AIF-1 genomic region was screened against the DNA sequence databases. Two different human, four different rat and one pig EMBL cDNA database entries, as well as 13 human, one mouse and two fish ESTs, aligned over this genomic region with significant similarity.

4.2.1.1 Alignments with human database entries

The two human EMBL entries were: the Allograft Inflammatory Factor-1 (AIF-1) gene (accession no U19713) (Utans *et al.* 1996) and the Interferon gamma Responsive Transcript-1 (IRT-1) (accession no U95213) (Autieri & Agrawal 1998). The 13 human EST entries (accession nos. W58116, W67117, R71716, N47817, W67362, T79488, T69387, N32593, N32605, W67118, W21034, AA091585, THC167143(TIGR)) fell into two groups: ESTs W21034, AA091585 and THC167143(TIGR) matched the G1 cDNA (Olavesen *et al.* 1991) exactly, whilst the remaining 10 ESTs were all found to be partial transcripts of AIF-1. The G1, AIF-1 and IRT-1 cDNA sequences are 500, 639 and 1235 nucleotides in length and encode polypeptides of 93, 147 and 132 amino acids, respectively (Olavesen *et al.* 1991; Utans *et al.* 1996; Autieri & Agrawal 1998). All three transcripts appear to be splice variants of the same gene. Therefore, to avoid confusion, and to reiterate the statement made in section 3.2.2.3.1, the AIF-1, G1 and IRT-1 transcripts will be referred to as splice variants of the AIF-1 gene and the genomic region encompassing these sequences will be referred to as the AIF-1 genomic region. The G1 transcript aligns with the genomic sequence over 3 exons in the reverse orientation, with an initiation AUG codon at genomic position 3492 and an in-frame stop at position 2706. The AIF-1 transcript aligns with the genomic sequence over 6 exons in the reverse orientation with an initiation AUG codon at position 4274 and an in-frame stop at position

2706, and the IRT-1 transcript aligns over 3 exons in the reverse orientation utilising the same initiation AUG codon as G1 (position 3492) and having an in-frame stop codon at position 3096. In total two different initiation AUG codons are utilised (AIF-1 utilising the first initiation AUG codon, whilst G1 and IRT-1 utilise the second (Figure 4.1)) and two different in-frame stop codons (IRT-1 utilising STOP-1 and AIF-1 and G1 utilising in-frame STOP-2 (Figure 4.1)). The co-ordinates of the exons shown in Figures 4.1 and 4.2, are summarised in Table 3.7 (see Chapter 3). The AIF-1 and G1 transcripts share their last two exons (Figure 4.2, exons 5 & 6) as well as the sequence within exon 4I of AIF-1. However, the 5' sequence of G1, encoded by exon 4II in Figure 4.2, extends into intron 3 of AIF-1 for 93nt and AIF-1 has a further 3 exons upstream. The IRT-1 transcript shares exon 2 of AIF-1 as its first exon, then exon 3I of the AIF-1 transcript as the 5' sequence of its second exon. However, the second exon of IRT-1 (exon 3II in Figure 4.2) contains additional 3' sequence, extending into intron 3 of AIF-1 by 307nt and overlapping with the 5' end of the first exon of G1 (exon 4II) by 33 nt. The third and final exon of IRT-1, exon 4III, includes the sequences of exons 4I, 5 and 6 of AIF-1 but is unbroken by intronic sequence. The open reading frame of the AIF-1 transcript spans all 6 of its exons and the G1 transcript all three of its exons. However, the dominant open reading frame of the IRT-1 transcript is entirely contained within its last exon, the first two exons being non-coding.

4.2.1.2 Alignment of database entries from non-human organisms

The five sequences from non-human organisms within the databases that closely matched the AIF-1 genomic region were: AIF-1 homologs from pig (P81076), mouse (AA016714), rat (P55009) (Chen *et al.* 1997), zebrafish [AA495202 (partial cDNA sequence)] and Carp [C88427 (partial cDNA sequence)]; a rat transcript varying from rat AIF-1 over its first 279 nucleotides and 8 amino acids (D82069); and the rat Balloon angioplasty responsive

TTTCTAATCT GCATTTCACA CCAGGCTCCC CCTCCTTCCT GGGGTGCTAG TGTCAGCAGA ACCTGATGGG GAAGTGAGGT

Exon 1 start site

CTGGGAGGCA GAGGAGGAAG GAATGAGGGG AAAGGGGAAG TTTGGGAGGA AGGCTTCTGA GAAGACTGGT G|**GGAGAGAAG**

GAGAGCCTGC AGACAGAGGC CTCCAGCTTG GTCTGTCTCC CCACCTCTAC CAGCATCTGC TGAGCTATGA GCCAAACCAG

GGATTTACAG G⌊GTAGGGAGG GTGGGATAGG CAGCGGCATT AGATCGGAGG AATGAGATGG ACAGACCTGG GCTGTGGGCT

AGGAGGGCAG TCAGCTGGCC TAGGGTAGCC CGGGCTGGTG TCAGGGTAAG GAGAGGAAGG GAGGGATGAG GGCTGATTAA

Exon 2

TTTTTTTCAC CCCACAG⌊**GAG GAAAAGCTTT CGGACTGCTG AAGGCCCAGC AGGAAGAGAG GCTGGATGAG ATCAACAAG**⌋G

TAGAAGGAAG AACTAAGGGG GCAGAGCCAG GGGGATGGGG CGTGGATGGG GAGGGCCTAC CCTGGCTCTT ATTTTCCCCT

Exon 3I and 3II — Exon 3I 3′ splice site↓

CCATAG⌊**CAAT TCCTAGACGA TCCCAAATAT AGCAGTGATG AGGATCTGCC CTCCAAACTG GAAGGCTTCA AAG**⌋**GTGAGGG**

GGAAACTGTA GGCGGTGGAG ACAGGGCTGG GGGTAGGAGG GTTAGGATTT CCACAAGAAC AAGGCAGGAA CAGCAGAGAT

AAAAAGTTTA CTTTTGTGGT AGCAAAAGGG GAACCTGCCT TTATTGCCCT CCTGCCACAC TGCGGTCCCT TTCCCGGGCC

TGCCTCTCTC AGCATCCCCT CTAGCTCCTT ACACCCTAGC GGGGCCCCTC AACTCCCCAA CCCCACTTCC TCTGCCTGCC

Exon 4II 5′ start site↓ — ↓Exon 3II 3′ splice site

CCTCCTCCTC CTTCCACGTT GTCTCCT|CCA CCTAGCAGTT GGTTGGCAAC CCCTTCCTCA⌋GTCCCCTGCT GAAAACCCTC

Exon 4III 5′ splice site↓ — ↓Exon 4I 5′ splice site

CAGTCAGCGC TTATCCCTTC TGCTCTCTCC C⌊CTCACCCAG⌊AGAAATACATG GGAGTTTGAC CTTAATGGAA ATGGCGATAT

↓Exon 4I and 4II 3′ splice site

TG⌋**GTGAGAAA CGGGTGATTT GCGGGGGCAG GGTGGTGTGC AGGCCTAAGA AGACAGAGGT CTCTCCTACA TGCTCCATTC**

CTCATGATTT GGGAGGGGGC CCACCTACCA CAGTGGGAGG AAGGAGAATG GGGATGCGGA AGTGGGAGAG GAGAGAGAGG

↓Exon 5 5′ splice site

GTCTCCCCAC CTTCTCCCCA TCCCCATCCT CTGCCCCCAG⌊ATATCATGTC CCTGAAACGA ATGCTGGAGA AACTTGGAGT

CCCCAAGACT CACCTAGAGC TAAAGAAATT AATTGGAGAG GTGTCCAGTG GCTCCGGGGA GACGTTCAGC TACCCTGACT

Exon 5 3′ splice site↓*1

TTCTCAGGAT GATGCTGGGC AAGAGATCTG CCATCCTAAA AAT⌋GTGAGTG TCAATTTCCA ACCTCCCCTG TACTTACCTG

TTTTCTCCTC CCCCATCCCT ACCCTTGTCC ACAGGCTCAA CATTTCTACA CGTTGCCCAT CATCCCTTCT TCCATCCTTA

GAGGGACCCT TCCAAGGTCC CGACCCCATC CCTATCCATA GTCCTGGTCC CCAGAAACTC CAACCCCTGC CCTTCCTCTT

CCCCCTTCCA CCCTCACATC CCCATCCCCT TCTAGCCTTT CCTAGCACCC TATGATTTAT TCCCTTGAGA GGAGTGTTCC

↓Exon 6 5′ splice site

CTGATCCCTG TGCCTCTTCC CATCTCAACC AG⌊GATCCTGA TGTATGAGGA AAAAGCGAGA GAAAAGGAAA AGCCAACAGG

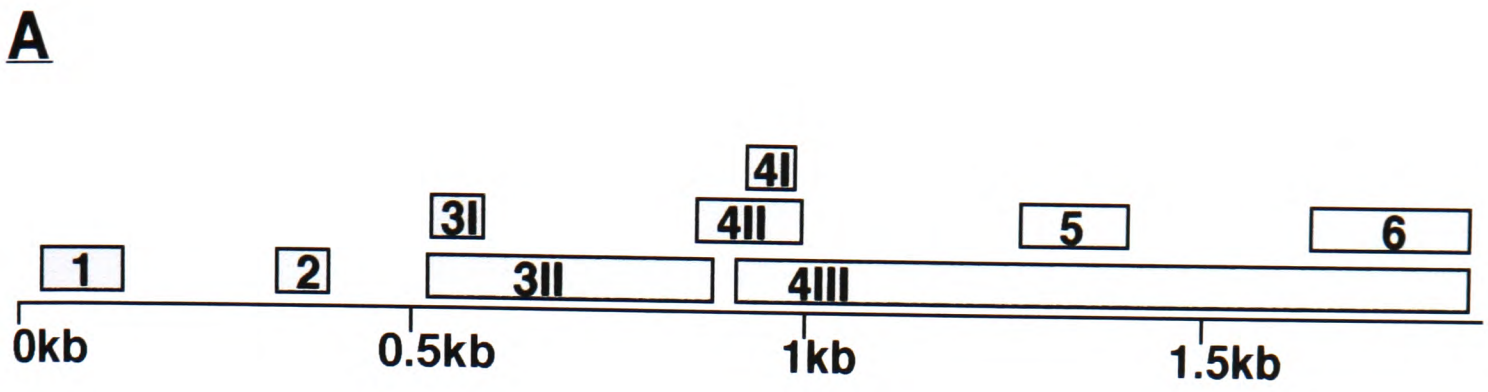
*2

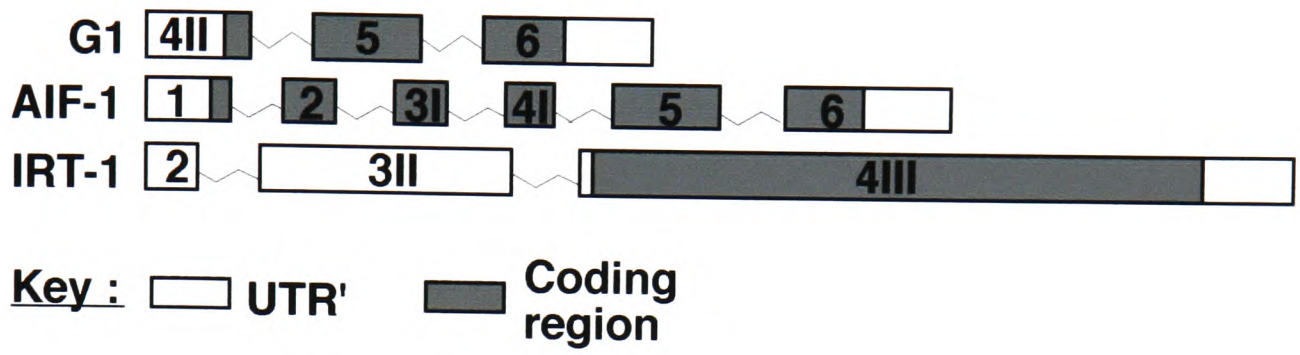
CCCCCCAGCC AAGAAAGCTA TCTCTGAGTT GCCCTGATTT GAAGGGAAAA GGGATGATGG GATTGAAGGG GCTTCTAATG

ACCCAGATAT GGAAACAGAA GACAAAATTG TAAGCCAGAG TCAACAAATT AATAAATTA CCCCCTCCTC CAGATCAAGT

CAGCTTAGTT TTTATTTGGG TGATTTTTTT CCTGGGTTTG GGAAGGAGAG ACAGGTCTTG AGGGAAAGGT GGCAAGGATT

FIGURE 4.1 The AIF-1 genomic region corresponding to co-ordinates 4500-2501 (reverse orientation) of the TN62/TN82 genomic sequence.
Exonic regions, as represented in Figure 4.2, are shown in bold. The exon numbers and information about the splice sites are underlined above the sequence. The 5' splice sites are indicated with a '⌊' within the sequence, and the 3' splice sites are indicated with a '⌋'. The transcriptional start sites of AIF-1 and G1 are indicated with a '|'. Where alternative 5' and 3' splice sites fall within exonic sequence these are further annotated with an arrow (↓) directly above them next to the relevant text. The two initiation codons are boxed in bold, the first is utilised in the AIF-1 transcript and the second in the G1 and IRT-1 transcripts. The two stop codons are boxed, indicated with a '*' above and numbered. Stop codon 1 is the termination codon for the deduced IRT-1 amino acid sequence and stop codon 2 is the termination codon for the deduced G1 and AIF-1 amino acid sequences. The polyadenylation signal is also highlighted (AATAAA).





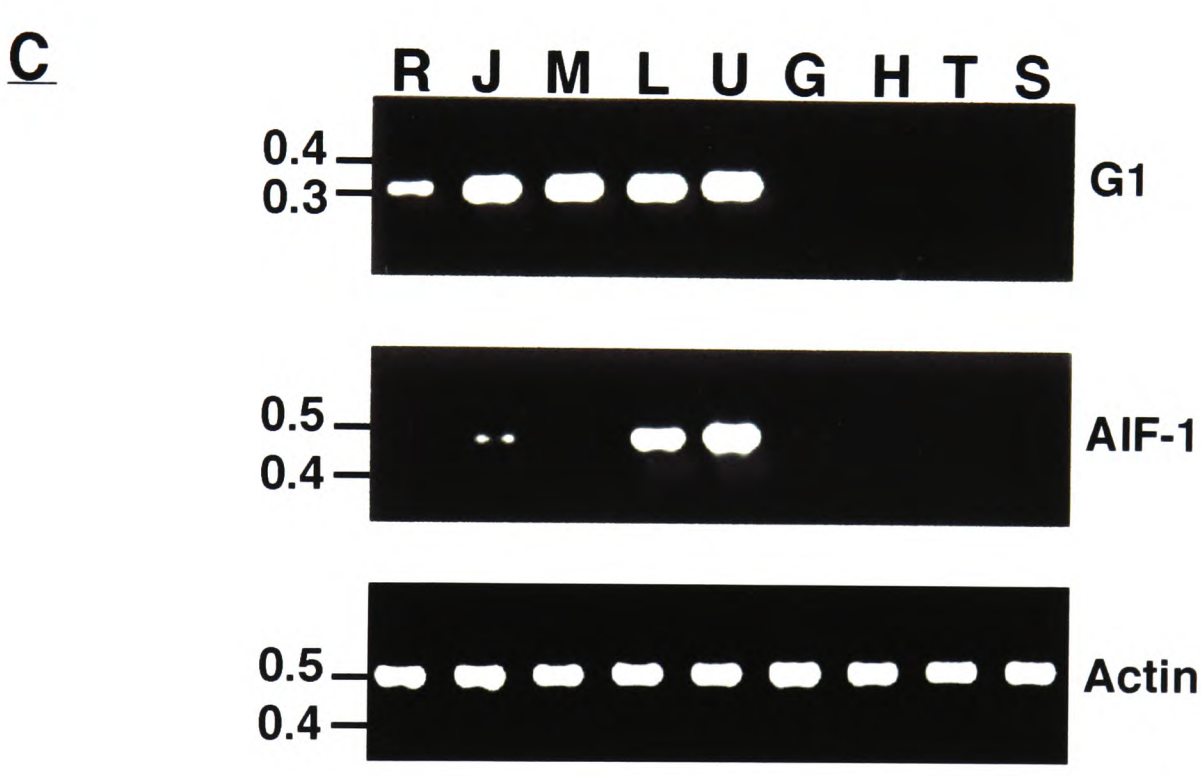


FIGURE 4.2 AIF1 splice variants.
The AIF-1 genomic structure (**A**), and the exon usage of the human transcripts G1, AIF1 and IRT-1 (**B**), were determined by comparison of the genomic sequence obtained in this study with previously described cDNA sequences.
C. Transcript profiles of G1 and AIF-1 were determined by RT-PCR. The cell lines used were Raji (R), Jurkat 6 (J), Molt 4 (M), HL60 (L), U937 (U), HepG2 (G), HeLa (H), HT1080 (T), SW620 (S) (see section 2.3.10.1). The primer pairs G1f - G1AIF1r, AIF1f - G1AIF1r, and ACTINf1 - ACTINr, were used in the analysis of G1, AIF1 and beta-actin, respectively (See Table 4.1). The sizes of mol. wt. markers (kb) are shown on the left of each panel. The RT-PCR products obtained using control primers from beta-actin are shown in the bottom panel.

transcript (BART-1) (U33471) (Autieri *et al.* 1996b) that differs yet again in its splicing pattern. The latter two rat transcripts (D82069 and U33471) appear to share exonic sequence, and possibly coding sequence, 5' of the first exon of AIF-1. Equivalent sequence was not identified in the corresponding region of the human genomic sequence, although this does not rule out the possibility that this alternative exon 1 sequence falls 5' of the sequence generated here. Furthermore, the BART-1 transcript contains an additional 444 nucleotides of internal sequence: a 158 nucleotide insert at TN62/TN82 genomic position 4021 between exons 1 and 2; and a 286 nucleotide insert at TN62/TN82 genomic position 3867 between exons 3I and 4I. The latter insertion encodes a putative transmembrane region predicted by the TMpred server (also reported by Autieri *et al.* 1996b), between amino acid residues 75 to 91 of BART-1, N-terminal of the G1 and IRT-1 protein sequences. The rat cDNA database entry BART-1 was found to contain a number of sequence discrepancies when compared to the rat AIF-1 transcripts (U17919, AB000818, U10894 and I22424), including 3 frame shift errors within the proposed coding region. After editing the BART-1 database sequence to correct for the identifiable frame shift errors, the length of the predicted BART-1 polypeptide deduced by us is 15 residues shorter (215 amino acids) than U33471 and differs from the database sequence over its first 26 amino acids and last 22 amino acids, although it still encodes a putative transmembrane region. A multiple alignment of the splice variants of the human AIF-1 gene and the orthologous proteins identified in other species, excluding BART-1, is presented in Figure 4.3 and the multiple alignment from Figure 4.3, together with the two deduced amino acid sequences for BART-1 is shown in Figure 4.4.

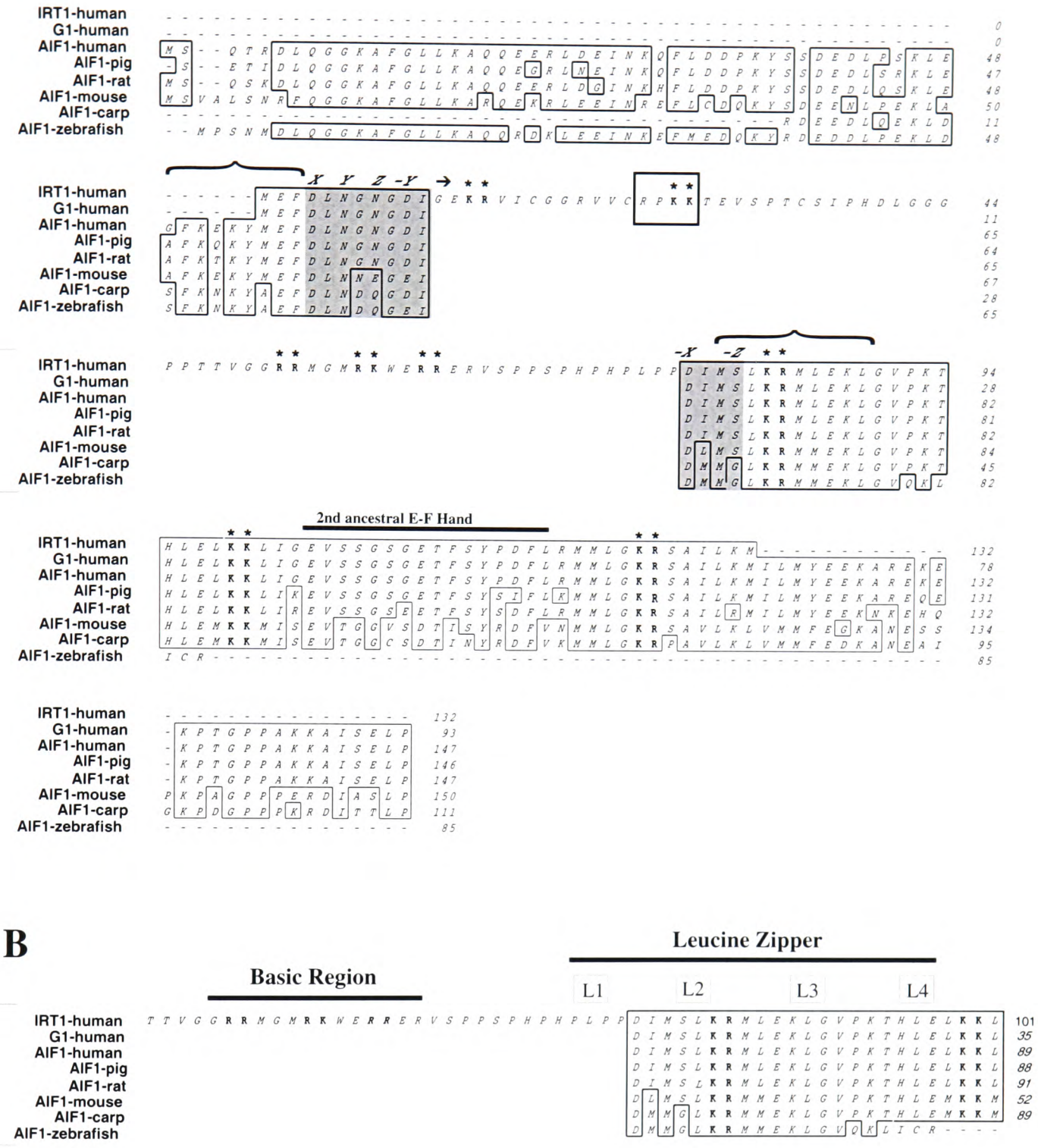


FIGURE 4.3 A multiple alignment of the AIF-1 amino acid sequences from different species and the alternatively spliced forms found in man.

A. The shaded area corresponds to the conserved loop of the first putative E-F hand calcium-binding sub-domain, with the residues involved in calcium co-ordination indicated by *X, Y, Z, -Y, -X, -Z* . The horizontal parentheses indicate the predicted positions of the conserved α-helices that border the E-F Hand loop. The predicted position of the 2nd, 'ancestral', E-F Hand is indicated above the region of sequence concerned. The paired basic residues characteristic of endoprotease cleavage sites are highlighted by asterisks '**' and the potential 4 amino acid nuclear targeting motif in IRT-1 is boxed. Both the Carp and Zebrafish AIF-1 sequences are derived from incomplete EST sequences.

B. A potential bZIP basic leucine zipper domain found between residues 51 and 90 of the IRT-1 protein product. The characteristic basic region is indicated above the relevant sequence, as is the leucine zipper region. The four leucine residues, spaced 7 amino acids apart, characteristic of the leucine zipper motif are indicated above the sequence in boxes.

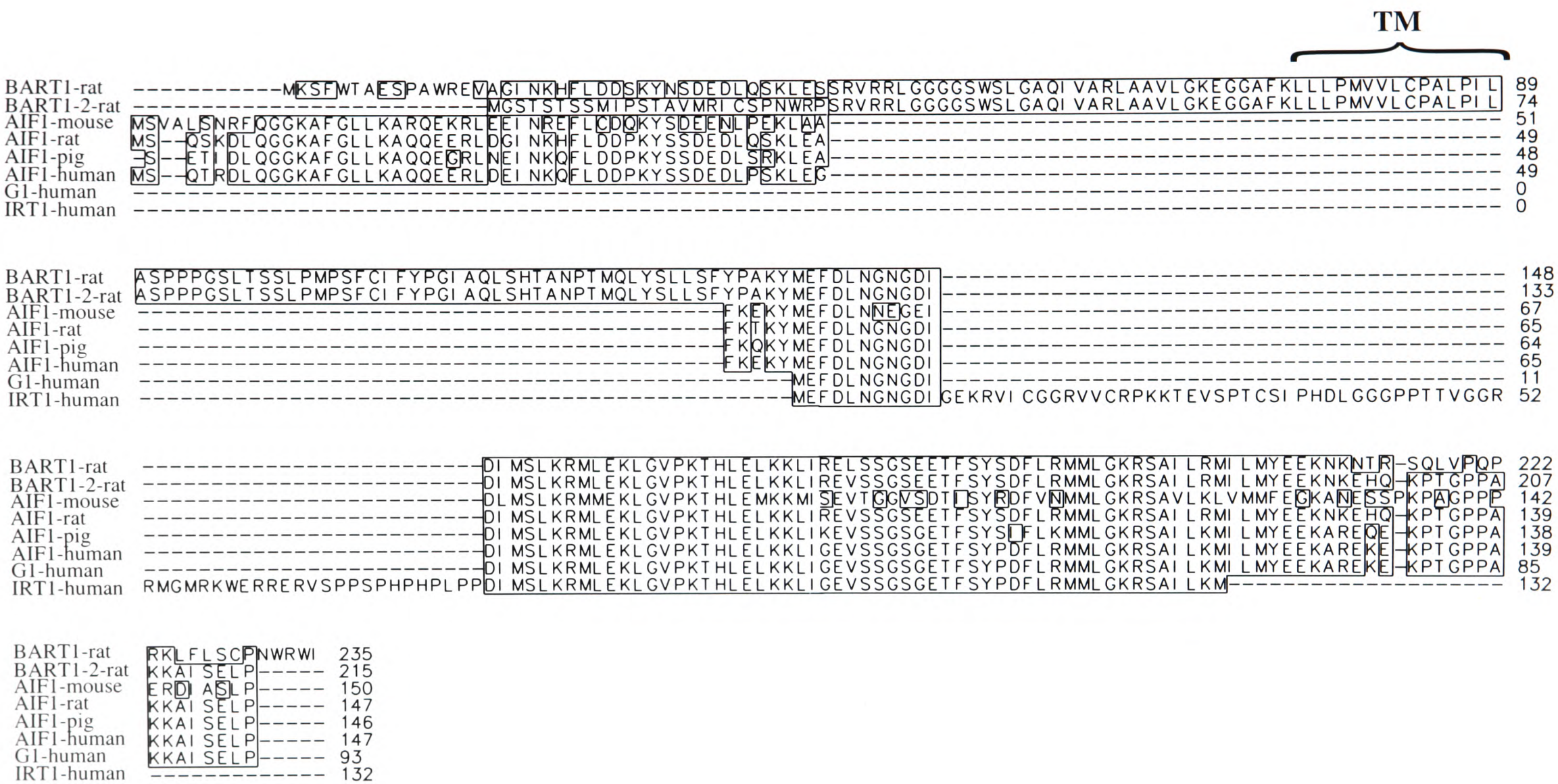


FIGURE 4.4 A multiple alignment of the AIF-1 amino acid sequence, as in Figure 4.3, with the addition of the BART-1 deduced amino acid sequence from Rat.
The amino acid sequence BART1-rat is the sequences available within the EMBL database (accession number D82069). The amino acid sequence BART1-2-rat is the deduced amino acid sequence of BART-1 after re-assessing the nucleotide sequence of D82069 and changing discrepancies found when compared to Rat AIF-1 transcripts.
The horizontal parenthesis indicates the position of a transmembrane region predicted by the TMpred server.

4.2.1.3 Analysis of the encoded amino acid sequences of the AIF-1 gene

Encoded within sequence from exons 4 and 5 of AIF-1, that is shared by all the alternative transcripts detected, is a putative EF-hand calcium-binding motif (shaded in Figure 4.3A). This motif was previously reported by Olavesen *et al.* (1991) and Utans *et al.* (1996) for G1 and AIF-1, respectively, and was also detected with the Prosite 'Profilescan' server. The splicing of the IRT-1 transcript causes an insertion of 66 amino acids into the conserved loop of the EF-hand domain. Both the PredictProtein server and the Profilescan server predicted a Leucine zipper motif (Prosite: PDOC00029) in the IRT-1 protein, between residues 75 to 96 (Figure 4.3B). This motif is preceded by a region high in basic residues (only found in IRT-1), a feature of a 'basic leucine zipper' (bZIP) DNA-binding domain (Prosite: PS00036). Also within the IRT-1 protein is a putative 4 residue nuclear targeting motif, RPKK (Figure 4.3A), predicted by the PSORT protein localisation prediction server.

The PredictProtein server identified 2 potential dibasic pro-hormone cleavage sites and one amidation site (Prosite: PS00009) in all splice variants of the AIF-1 gene and a further amidation site, plus 4 dibasic sites specific to the IRT-1 transcript (Figure 4.3A). G1, AIF-1 and IRT-1showed no evidence of leader peptides, when analysed by the GCG 'SigCleave' program, or of transmembrane helices, when submitted to the 'PredictProtein' or TMpred servers. Lastly, all the proteins were predicted to have a helix-loop-helix secondary structure by the JPRED server.

4.2.1.4 Expression profiling of the AIF-1 gene encoded transcripts by RT-PCR

Expression profiles were determined for G1, AIF-1 and IRT-1 by RT-PCR (sections 2.3.10.2 and 2.3.10.3) using transcript specific primers (Table 4.1). The results show that

G1 is expressed at high levels in the immune related cell lines Jurkat 6, Molt 4, HL60 and U937 and to a slightly lesser extent in the cell line Raji (Figure 4.2C). Although AIF-1 expression was detectable in the same cell lines as G1, it was only detected at high levels in the monocyte lineage cell lines HL60 and U937. This is consistent with the findings of Utans *et al.* (1996). No signal was detected with the IRT-1 specific, IRT1f - G1AIFr primer pair, nor was an IRT-1 specific band detected with the IRT-1 and AIF-1 specific, IRTAIF-1f - G1AIFr primer pair, whereas an AIF-1 specific band was detected (data not shown).

TABLE 4.1 Transcript specific RT-PCR primers for the AIF-1 gene.

PRIMER	SEQUENCE	TRANSCRIPT	POSITION [a]	PCR CONDITIONS
ACTINf1	5'-CTTCGCGGGC GACGATGG-3'	BETA ACTIN	+ve control	1mM $MgCl_2$, anneal 60° C, 30 cycles
ACTINr	5'-TGGTGGTGAA GCTGTAGCC-3'			
IRTAIFf	5'-GACTGCTGAAG GCCCAGC-3'	IRT-1 & AIF-1	4068-4051	
IRTf	5'-ACACTGCGGTC CCTTTCCC-3'	IRT-1	3724-3706	1mM $MgCl_2$, anneal 58° C, 30 cycles
G1f	5'-TGAAAACCCT CCAGTCAGC-3'	G1	3551-3533	
AIF1f	5'-ACCTCTACCA GCATCTGC-3'	AIF-1	4298-4281	
G1AIF1r	5'-CCATCATCCCT TTTCCCTTCA-3'	GI, AIF-1 & IRT-1	2681-2701	

[a] Positions of primers within the genomic sequence generated in this study.

4.2.2 The 1C7 gene

4.2.2.1 Database interrogation, DNA sequence analysis and cDNA clone characterisation

In addition to the published 1C7 cDNA sequence (Nalabolu *et al.* 1996), two I.M.A.G.E consortium cDNA clones, 685808 (EST AA262074) and 683968 (ESTs AA237100 and AA236886), both of B cell origin were found to align with the genomic region immediately centromeric of the LST-1 (B144) gene. Sequence analysis revealed that these

clones contained inserts of 1120 nucleotides (683963) and 721 nucleotides (685808), respectively, and both were found to align with the genomic sequence over 4 exons. Both clones extend the published 1C7 cDNA sequence (Nalabolu *et al.* 1996; de Baey *et al.* 1997) by 306 nucleotides at the 5' end and contain an additional 1C7 exon. This additional coding sequence lies 2.8kb upstream of that previously published and contains an in-frame initiation AUG codon (Figure 4.5) preceded by in-frame stop codons. Clone 683963 also extends the 1C7 sequence by 141 nucleotides at the 3' end to include a polyadenylation signal at genomic position 30678 and a polyadenylation site at genomic position 30711. This 1C7 polyadenylation signal lies only 47 nucleotides centromeric of the polyadenylation signal for LST-1, the two genes being transcribed in opposite directions. The three transcripts, derived from the 1C7 gene, have been named 1C7a (clone 683963), 1C7b (the partial 1C7 sequence already published (Nalabolu *et al.* 1996)) and 1C7c (clone 685808). These all differ in their last exon (Figure 4.6A, exons 4I, 4II and 4III). The 1C7a and 1C7b transcripts, although sharing their last 276 nucleotides, differ in the 5' splice site of their fourth and last exon, with the fourth exon of 1C7b (exon 4II) extending into intron 3 of 1C7a by 55 nucleotides. The fourth exon of 1C7c (exon 4I), on the other hand, lies entirely within intron 3 of 1C7a and 1C7b (Figure 4.5) and has a separate stop codon at genomic position 30298, a polyadenylation signal at genomic position 30319 and a polyadenylation site at position 30337.

4.2.2.2 Expression profiling of the 1C7 gene transcripts and the identification of further splice variants by RT-PCR

Transcript profiles were generated by RT-PCR using primer pairs specific for 1C7a & b (Figure 4.6B, panel I) or for 1C7c (Figure 4.6B, panel II). RT-PCR primer sequences can be found in Table 4.2. A total of nine different PCR products were consistently generated.



FIGURE 4.5 The 1C7 genomic region corresponding to co-ordinates 26401-30700 of the TN62/TN82 genomic sequence.

Exonic regions, as represented in Figure 4.6, are shown in bold. The exon number and information about the splice sites is underlined above the sequence. The 5' splice sites are indicated with a '⌊' within the sequence, and the 3' splice sites are indicated with a '⌋'. The transcriptional start and stop sites are indicated with a '|'. Alternative 5'and 3' splice sites when falling within exonic sequence are further annotated with an arrow (↓) directly above them, next to the relevant text. The initiation AUG codon is outlined in bold, and the three stop codons are outlined and indicated with an asterisk ('*') above, and numbered. Stop codon 1 is the termination codon for splice variants using exon 4I, stop codon 2 is the termination codon for splice variants using exon 4II and stop codon 3 is the termination codon for splice variants using exon 4III (Figure 4.6). The polyadenylation signals are also highlighted (AATAAA).

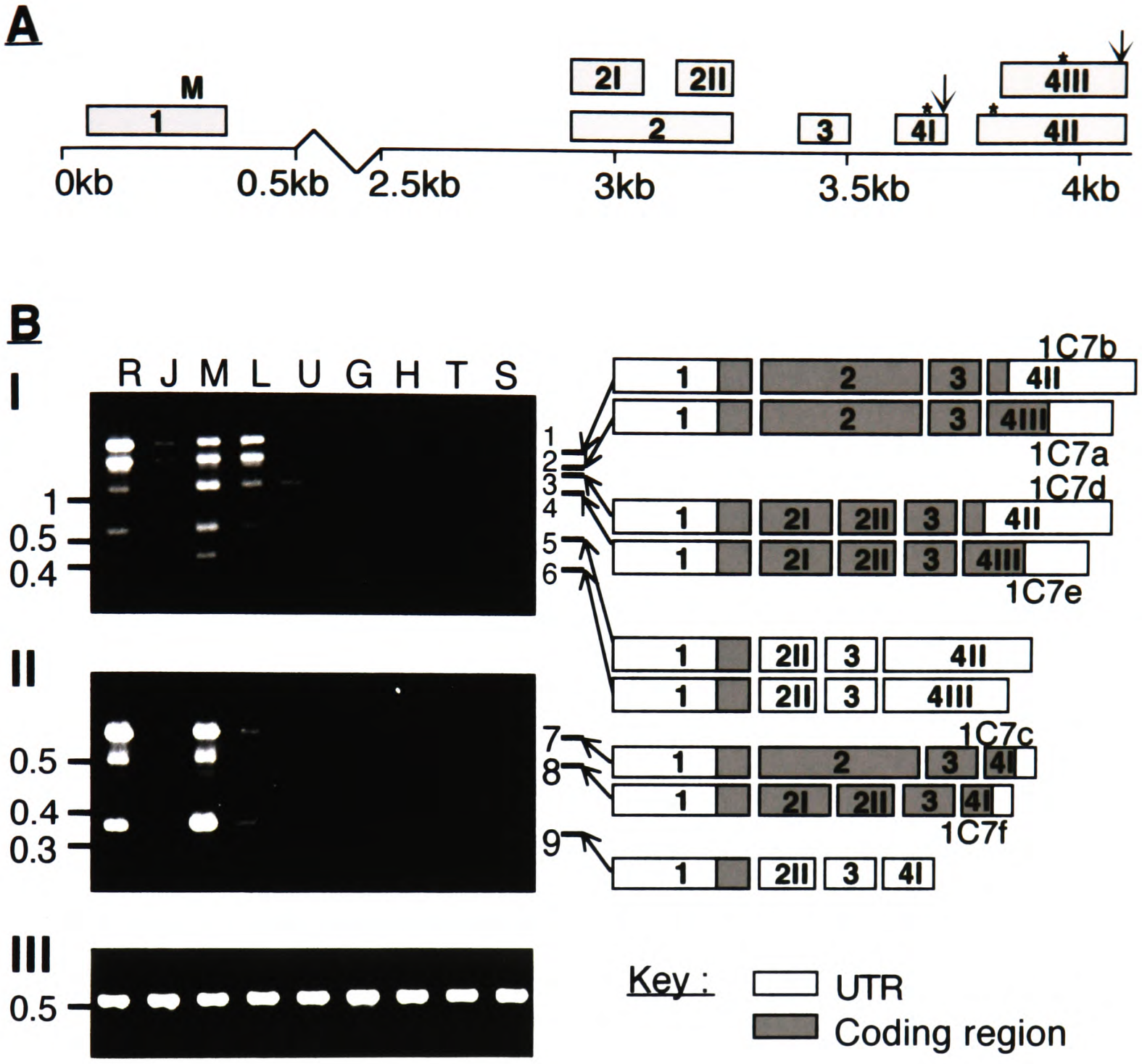


Figure 4.6 1C7 splice variants detected by RT-PCR and EST database searches.

A. The genomic organisation of exons identified within 1C7 transcripts was determined by comparison of the genomic sequence obtained here with the sequences from the EST databases and associated cDNA clones, plus the previously described partial cDNA (Nalabolu *et al.* 1996), and RT-PCR products. "M" indicates the putative initiation AUG codon. The position of the stop codons within the differentially spliced forms of exon 4 are indicated by asterisks "*" and the arrows indicate the position of the two alternative polyadenylation signals.

B. Transcript profiles of 1C7. Panel I shows the results of RT-PCR carried out using the primers 1C7289f (from exon 1) and 1C7ab3906r (from exons 4II and 4III), whilst the species in panel II were amplified using the primers 1C7289f (from exon 1) and 1C7c3688r (from exon 4I) (see Table 4.2). Panel III shows the results of RT-PCR with the control primers ACTINf1 and ACTINr1. The cell lines used were Raji (R), Jurkat 6 (J), Molt 4 (M), HL60 (L), U937 (U), HepG2 (G), HeLa (H), HT1080 (T), SW620 (S). The sizes of mol. wt. markers (in kb) are indicated on the left of each panel and the exon usage (determined by sequence analysis) of each splice variant is shown on the right. The splice variants detected, containing open reading frames of significant length have been named 1C7a to f. 1C7a represents cDNA clone 683963, 1C7b contains the partial sequence published by Nalabolu *et al.* (1996) and 1C7c represents the cDNA clone 685808.

Initial Southern blot analysis and subsequent sequence analysis, confirmed that these were all splice variants of 1C7. The exon usage of each transcript is indicated alongside the appropriate band in Figure 4.6B. Band 1 corresponds to the splice variant 1C7b, band 2 to 1C7a and band 7 to 1C7c. 1C7a and 1C7b (Figure 4.6B, panel I) show similar expression patterns, being expressed predominantly in the immune-related cell lines Raji, Molt 4 and HL60 and to a lesser extent in Jurkat 6 and U937. 1C7c consistently shows a slightly more specific expression pattern, being detected at high levels in the lymphocyte cell lines Raji and Molt 4 only (Figure 4B, panel II) with a lower level of expression detected in HL60.

No product was detected by Northern blot analysis with Raji, Jurkat 6 or HL60 cell line RNA using a PCR product corresponding to the exon 2 sequence as a DNA probe.

TABLE 4.2 Transcript specific RT-PCR primers for the 1C7 gene.

PRIMER	SEQUENCE	TRANSCRIPT	POSITION [a]	PCR CONDITIONS
ACTINf1	5'-CTTCGCGGGC GACGATGG-3'	BETA ACTIN	+ve control	1mM $MgCl_2$, anneal 60° C, 30 cycles
ACTINr	5'-TGGTGGTGAA GCTGTAGCC-3'			
1C7289f	5'-CCAACTGGGA CATCTTCCG-3'	1C7	26870-26888	1.5mM $MgCl_2$, anneal 58° C, 35 cycles
1C7c3688r	5'-GAGGACTAGG GACATCTGGG-3'	1C7c	30306-30286	
1C7ab3906r	5'-AGCAGATGTG CTGAGCTCC-3'	1C7a&b	30523-30505	

[a] Positions of primers within the genomic sequence generated in this study.

Apart from the variation in exon 4 usage described above, RT-PCR identified 1C7 transcripts where exon 2 is divided into two exons (exons 2I and 2II in Figure 4.6A). In the smaller PCR products obtained (bands 5, 6 and 9 in Figure 4.6B), only the 3' segment of exon 2 (2II) is spliced in. These transcripts encode a truncated polypeptide of only 15 amino acids, due to the incorporation of a stop codon as a result of the splicing of exon 1 to

exon 2II, and are probably non-functional. However, in the RT-PCR products 3, 4 and 8 (Figure 4.6B), both exons 2I and 2II are used, resulting in a 75 nucleotide deletion in the centre of exon 2. Despite this deletion the open reading frame is retained, resulting in a transcript 75 nucleotides shorter and an encoded polypeptide 25 amino acids shorter than in the corresponding species using the complete exon 2 (i.e. PCR products 1, 2 and 7). This variation in usage of exon 2 results in three novel splice variants: 1C7d, 1C7e and 1C7f (Figure 4.6B). The six detected transcripts of 1C7 that contain an open reading frame of significant length comprise either four or five exons, including one of three alternatively spliced last exons and either a single exon 2 or two exons from the exon 2 region. All exons are in phase 1 (Sharp 1981) at their 3' end except where exon 2 is split into two, in which case the boundary between exons 2I and 2II is in phase 0 (Figure 4.7B). The transcripts 1C7a to 1C7f encode six putative protein isoforms of 201, 178, 190, 176, 153 and 165 amino acids, respectively (Figure 4.7A).

4.2.2.3 Amino acid sequence analysis of the 1C7 gene encoded transcripts

The amino acid sequences encoded by exon 2 and by exons 2I and 2II of 1C7 were found to have significant similarity with diverse immunoglobulin (Ig) domain-containing proteins and were also predicted to contain Ig domains by the SMART server. The isoforms 1C7a, 1C7b and 1C7c are most similar to the T Cell Receptor (TcR) α-chain variable domain (accession number M2735, 33% identity and 45% similarity over 117 amino acids) while the truncated isoforms 1C7d, 1C7e and 1C7f are most similar to an IgC2 domain from Perlecan (accession number Q05793, 35% identity and 46% similarity over 77 amino acids). Structure based multiple sequence alignments of the putative Ig domain from 1C7a, b and c with other Ig variable domains and of the putative Ig domain from 1C7d, e and f with other IgC2 domains are shown in Figure 4.8. All the key residues characteristic

A

```
                                  20                                  40
MAWMLLLILIMVHPGSCALWVSQPPEIRTLEGSSAFLPCSF

                                  60                                  80
NASQGRLAIGSVTWFRDEVVPGKEVRNGTPEFRGRLAPLAS

                                 100                                 120
SRFLHDHQAELHIRDVRGHDASIYVCRVEVLGLGVGTGNGTR

                                 140                                 160
LVVEKEHPQLGAGTVLLLRAGFYAVSFLSVAVGSTVYYQG

                                 180
    CHCHMGTHCHSSDGPRGVIPEPRCP  4I (1C7 c & f)
K <
    YAKSTLSGFPQL  4II (1C7 b & d)
                                                                  200
    CLTWKGPRRQLPAVVPAPLPPPCGSSAHLLPPVPGG
                                                  4III (1C7 a & e)
```

B

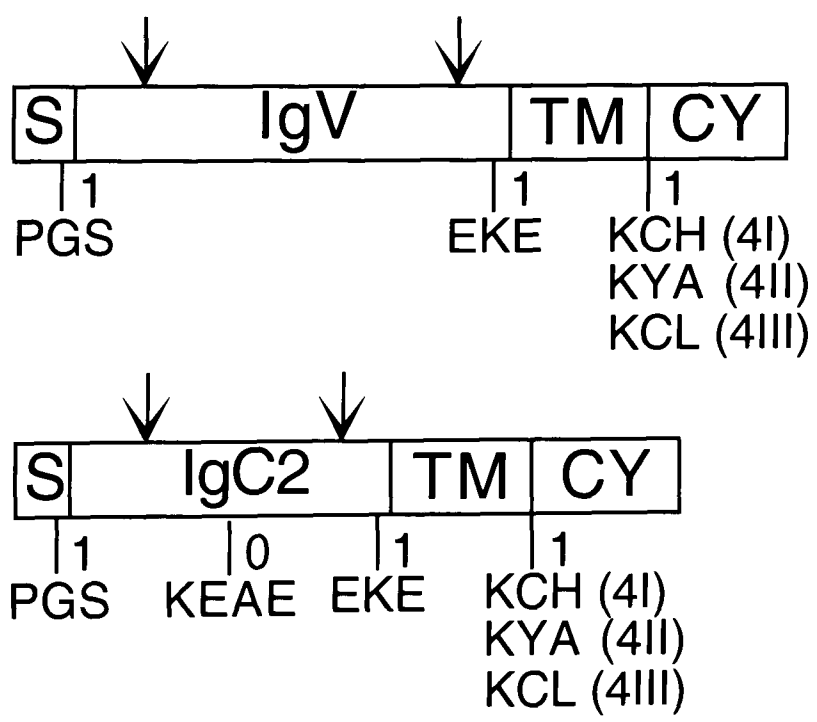


Figure 4.7 The amino acid sequence and putative structure of 1C7.
A. The amino acid sequence of the identified 1C7 splice variants. The residues spliced out of the IgC2 domain-containing isoform are underlined and the putative transmembrane region is double underlined. The two potential N-glycosylation sites are boxed and the potential SH3-domain binding motifs are shadowed, with the most likely motif underlined.
B. Domain organisation of 1C7. The positions of the predicted signal peptide (S), the Immunoglobulin V domain (found in 1C7a, b and c) and C2 domain (found in 1C7d, e and f) (IgV and IgC2, respectively), the predicted transmembrane region (TM) and the cytoplasmic domain (CY) are shown. The positions of exon boundaries are indicated by vertical lines below the Figure: the residues bordering the splice sites are shown below each line. The phase of the exon boundary is indicated to the right of each line. The positions of the two cysteine residues which form the disulphide bond are indicated by arrows.

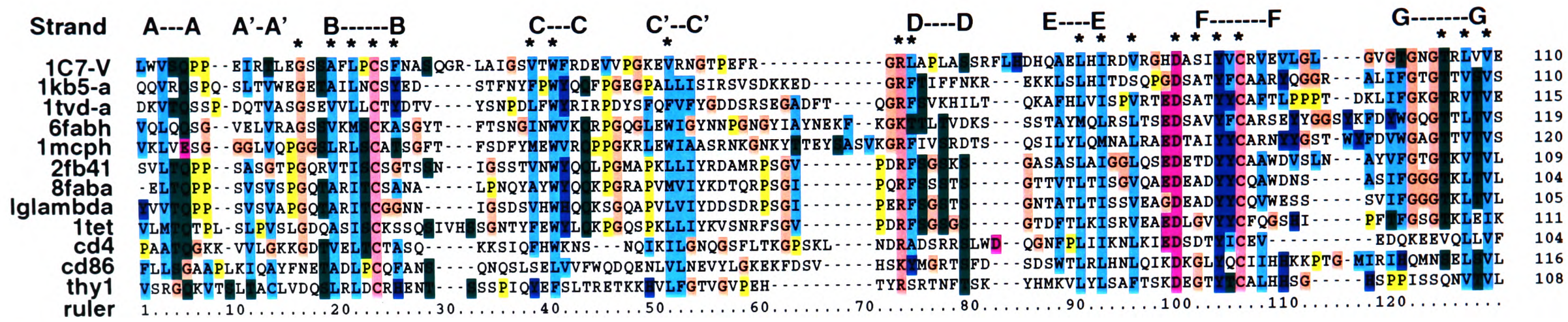


FIGURE 4.8 Structure based amino acid sequence alignments of the putative immunoglobulin domains of 1C7.
The positions of β-strands A to G are shown, and the key conserved immunoglobulin 'V frame' residues are indicated by asterisks "*" (Harpaz and Chothia, 1994). Residues that are identical, or that have similar characteristics between domains, are highlighted in the same colour, the colour scheme being the default from the GCG Clustalx program. **A**. Alignment of the predicted Ig variable domain from 1C7a, 1C7b and 1C7c with selected Ig variable domains from the protein databases. Sequence names correspond to PDB entry names except for: Iglambda, cd86 and thy1 which refer to accession numbers S24319, P42081 and P04216, respectively. **B**. Alignment of the predicted IgC2 domain from 1C7d, 1C7e and 1C7f with selected IgC2 domains from the protein databases. Sequence names refer to accession numbers: P06731 (cea), P04217 (a1bg), P16880 (mag), P13592 (ncam), P07766 (cd3e) and P09693 (cd3g). Roman numerals refer to the domain number within each protein.

of Ig domains (as outlined by Harpez and Chothia (Harpaz & Chothia 1994)) are conserved within the putative Ig domains of 1C7. Of particular interest is the result of the deletion in the central region of exon 2 in 1C7d, 1C7e and 1C7f. This changes the characteristics of the Ig domain in 1C7 from an IgV domain to an IgC2 domain, i.e. having many of the key residues of an Ig variable domain, but missing the Ig β-strand D and having a significantly shorter β-strand C' (Harpaz & Chothia 1994).

In addition to an Ig domain, the 1C7 transcripts were predicted by the SIGCLEAVE and SMART servers to encode a signal peptide spanning amino acid residues 1 to 18. Furthermore, PredictProtein and TMpred predicted a helical transmembrane region spanning residues 138 to 165 of 1C7a, 1C7b and 1C7c and residues 112 to 140 of 1C7d, 1C7e and 1C7f (Figure 4.7). Two N-linked glycosylation sites at Asn residues 42 and 121 are also present in all splice variants (Figure 4.7A). In addition, a proline-rich region encoded in exon 4III and, therefore, only present in the C-termini of 1C7a and 1C7e contains a number of potential Src homology 3 (SH3) domain binding motifs (Gout *et al.* 1993; Scaife *et al.* 1994) and a potential protein kinase C phosphorylation motif at residues 168 to 171 (Figure 4.7A).

4.2.3 An ATPase G subunit Homolog

4.2.3.1 Database interrogation, DNA sequence analysis and cDNA clone characterisation

The programme GRAIL predicted the presence of 3 exons ~1kb telomeric of the IKBL gene. The I.M.A.G.E. cDNA clone 726424 (ESTs AA401769, AA399356) and the ATCC cDNA clone 124837 (EST AA324358) were found to align with this genomic region and when sequenced were also found to span 3 exons (Figures 4.9 & 4.10). However, the clones were found to have different 3' splice sites for exon 1 and to utilise different

GAAAAAGATG GAGATGTTAA CAACGGGAGG CAGGGGAGGG GGCGGGATGG TGGAGAGAGA GAGAAAGGTA GAGAGTTAGT
TTAGAATTAA GCCCAGAATG CTCTTTTCCC AACACAGGTT GCATGATGAC ATCCTTACCT TTTCACCGCA TTCACAACCC
TTTATGCTTC CTCTGTTACC ACCAATCAAG TTCTCCTTCT CTCACCTCAG TACTCCCCCG TCTCCGCCCC TGCCTCATCC

Exon 1I & 1II

CTAGACCTTT CCGACTGGGA T|**GGCTAACCT GTTGTAAGCC CGCAGCTTTG GGCCTGGTCT CTGCTGCTCC CAGGCGGCCC**

CTTTGGTACT GCCTGAGCAA GAAGTGCTGG AGAGGAGGAC CAGTCATCAA TAGGAGGATG AGATTGGGAG AGACACTCGG

Exon 1II 3′ splice site (clone 726424)↓

TGCAGGAGGC TGAGTGAGCA GGGGAGCACT AAGACCCAGG GGTAGTGGAG GACTGCAGCA ACGAGCTGGA GGAGGAGAA⌋G

Initiation AUG codon (clone 124837)

TAAGCGGTGG GGGGTGGGAG CCATCTGGTA CTTTGACAGC ATTCAAAACA GCATCGGCCA TAACAACAGA A[ATG]GCCAGT

Exon 1I 3′ splice site (clone 124837)↓

CAGTCCCAAG GTATCCAGCA GCTTCTGCAA GCTGAGAAGC GGGCAGCTGA GAAGGTGGCA GATGCCAGAA AGA⌋GTGAGTC
TCCTCTTTCC TCCCTTAGGA GTTTGGAAAG AAAATTGGGG GTGGGGGACA GCAAACATTT TGGGAAAACC CAAGGCTGGC
GGGAAGACAG ACAGCTAGGG TCTGGAGGCT GGTTAGGAGG GAAGAAATGG ATGGATATTA GAATCTGGCA CCTGGTTGGC
TGAGAGAAGG CTGTATAACT TTCTGGAAGG GACTGACTCC TGCTATTACA TTGTGTGTGT GTGGGTCCAT CCCCACTCAC

Exon 2

Initiation AUG codon (clone 726424)

TGTCCTTTCT TCTGCCTCCA G⌊**GGAAGGCCC GGCGACTGAA GCAGGCAAAG GAGGAGGCAC AG[ATG]GAGGT GGAGCAATAC**

CGCAGAGAGC GAGAGCACGA ATTCCAGAGC AAGCAGCAGG CG⌋GTGAGTTG AGGCAGAGTC GGGTTGAGAC CCCACTGCAA
GTTGGTGGGT GCATCTAGTG AGGTGTGTAA GGGTGACTCA ACAAGAAAAT ATGGTGGCAG AGGGCTGAGG CTGAGGGGAC
CCTGGCAGGG ACCACAACAT TGGTGAAACT TTGTGATGAT ATGTAGGAGA GTCTGGGAGT TTTGAAGGCC ACATAGAGCT
TGTGGGCGGA ATGCCACAGT CTGTGTAAAG TATAACATCT ATGTGGAGTA TGATTAACAT TTGTGGTGGA GGGTAGAGTT
TTATGGTCAT GGATGGTGAG GTGGTGGGGA TATTACGGTC TGTTTTAGGA TGAAGTTGCA TGTTAGGTCT AAGGGGAAAG
GGGACTGTGT TGATCTCTTT GGTGTTGGGA TATTTCTGTG GGATGGGGGT GGTTTCTGAG AGGGCCTTTC TTCTAGGCTT
TGTTTCAGGA TCTTTCCCCT CATATGCCTG GACCCTTGTC TGTTTCTGCT TTTCCCTTTC TCTCTTCCAC CCCTCTCCCT

Exon 3

ACCCCCCAG⌊**G CCATGGGCTC CCAGGGGAAC CTGTCTGCTG AGGTGGAGCA GGCTACAAGG CGCCAGGTGC AGGGCATGCA**
GAGCTCCCAG CAGAGAAACC GAGAGCGTGT CCTGGCCCAG CTTCTTGGCA TGGTCTGCGA CGTCAGGCCC CAGGTCCACC

*

CCAACTACCG GATTTCTGCC TAGGGCCACC G[TAG]GGCCTG ACTCCTTCTG CCAGTTCCCT CCCTCAAAGA AATCCTCCAA
TCAAAATCAC CTCCCACCAT AATCCCTGTC TTCTTTCCAT CCCCTAGAAA TCCTGGGAGG CAGGATCCAA TAATTTTCCT

↓Poly A site for clone 726424

GTGACACTTA TAAATATCCT GCTCACATCT GAA|TCTCCTT GTTGTTCTTT AACCCTCACT GGGACTTTGT AAACTTCCAA
GTCATTCTCA CCTAAACCCT CTGTGAAATT TGTAATATGG GGAAGTAGGA ATGTGGAAAA CATCCTGACT TCAGTGTCTG
GCCGATGTGG GTCCCTCTCT TGACCCTGTC ACTTGCTGGC TGTGAAACCA GGACAAGCTA CTTAACTTGG TAGCCTCGAT
GTCCTCCTCT GTGAAACTGG GATGATAATA ATGCCTACCT TGTGAGGGTT GCTTCAATGA TTAGGAATCA TTCTGTAAAG
TCTAGCACAG TTCCTTGCAT GTTGTAGCAG TGATTCAGTA AGTAGCAACC CTGTGATACT ATTACCACCA CCTGCTCACT
GGTCAAAACC TACACAGCTG TTTCCTCACG TCCATCACTG GCTCTCTAAT TCCACTTGTT CATTCTGTGA CCCTAGTTAT
TTTCTGAAAA ATTGGTTCTT CTCTTTTCCC AGAGACCTTC TGATCTCCAA AAAGAGGAGA TGACTACATT TAGCCCCTCT
CTTATAATTC CAGGTAGATA ACTGCATTTT GTAGCCTCTC TTTGTTTTTC TTTTGCTGAT CTTTGTCTTT ATTAGATTTT
CCTCCTTTCC TATTTCCCCA AAGACTTATC AGATGCTCAT TGCTTTCTAA GATCTAAAAT GATACTGTGT TCCCTCATAT
GCATGCCCTT CCTTTCTATA TCCTTGACAC CTTACTTTCC CATTGTAACA ATAAAAAAAG TATCAATAAA ATAATTATTG

↓Poly A site for clone 124837

GCAAATAAA|T TGGTGAGTTG AAGCAGCCTC CTTTTGCCTC ATCATTTCTC ATTTTCAGTC ACTTTGTTTT TTTTTTTTTT
GAGATGGAGT TTTGCTCTTG TTGCCCAGGC TGGAATACAA TGGCGTGATC TCAGCTCATT GCAACCTCTG CCTCCCAGGT
TCAAGCGATT CTCCTGCCTC AGCCTCCCAA GTTGCTGGAA TTATGGGTGT GTGCCACCAC GCCTGGCTAT TTTTTGTATT
TTTAGTAGAG ATGGGGTTTC

FIGURE 4.9 The ATP6G genomic region from the genomic sequence generated in this study (co-ordinates 72500-75399).
Exonic regions, as represented in Figure 4.10, are shown in bold. The exon numbers and information about the splices sites are underlined above the sequence. The 5' splice sites are indicated with a '⌊' within the sequence, and the 3' splice sites are indicated with a '⌋'. The transcriptional start and stop sites are indicated with a '|'. Alternative 5'and 3' splice sites, are further annotated with an arrow (↓) directly above them next to the relevant text. The two alternative initiation AUG codons are outlined in bold, and the stop codon is outlined and indicated with an asterisk ('*') above. All potential poly-adenylation signals for the two identified poly-adenylation sites are highlighted (AATAAA).

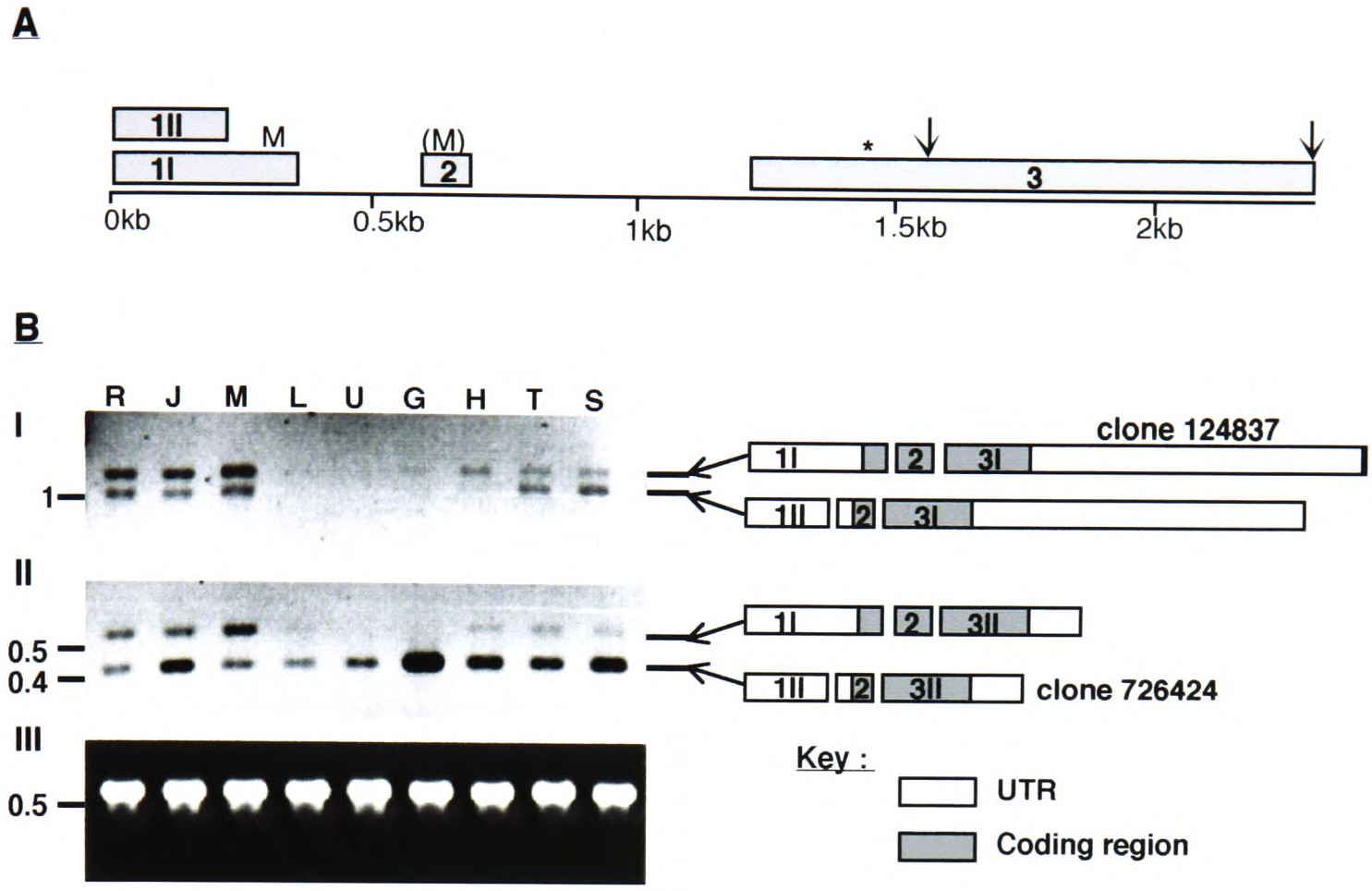


Figure 4.10 The genomic structure and expression pattern of the ATP6G gene.

A. The genomic structure of ATP6G. Exon organisation was determined by comparison of the genomic sequence obtained here with the sequences from the EST databases and associated cDNA clones and from RT-PCR products obtained in this study. The predictions from the NIX program (Figure 3.12) were also used . "M" indicates the position of the putative initiation AUG codon of transcripts using exon 1I (clone 124837). "(M)" indicates the position of the second in-frame AUG codon which is the putative initiation AUG of the truncated transcripts using exon 1II (clone 726424). The position of the stop codon is indicated by an asterisk "*" and the arrows highlight the positions of the two alternative polyadenylation sites.

B. Transcript profiles of the alternatively spliced forms of ATP6G.

Panel I shows the results of RT-PCR carried out using the primers ATP6Gf and ATP6Gr2 (specific for transcripts that contain the downstream polyadenylation signal and thus the longer 3' UTR). Panel II shows the results of RT-PCR carried out using the primers ATP6Gf and ATP6Gr1 (to amplify all ATP6G transcripts). Panel III shows the results of RT-PCR using the control primers ACTINf1 and ACTINr (Table 4.3). The cell lines used were Raji (R), Jurkat 6 (J), Molt 4 (M), HL60 (L), U937 (U), HepG2 (G), HeLa (H), HT1080 (T), SW620 (S). The positions of mol. wt. markers (in kb) are indicated on the left of each panel and the exon usage of each amplification product (determined by sequence analysis) is shown on the right.

polyadenylation sites (Figure 4.10A and B). Both clones contained transcripts with tracts of poly-As at their 3' end, clone 726424 having a run of 33 adenines and clone 124837 a run of 36 adenines. Clone 124837, which utilises the second down-stream polyadenylation site, contains a particularly long 3' untranslated region (947 nucleotides) (Figure 4.9). In addition, both polyadenylation sites observed have more than one potential polyadenylation signal. Whilst the polyadenylation site utilised in clone 124837 is preceded by two classical polyadenylation motifs (AATAAA) (see figure 4.9), this is not the case for the shorter transcript represented by clone 726424. However, the latter does contain two variants of the polyadenylation signal just upstream of the polyadenylation site: a 'AATAAT' motif 40 nucleotides upstream and a 'TATAAA' motif 20 nucleotides upstream (Figure 4.9).

Clones 726424 and 124837 contained inserts of 663 and 1628 nucleotides, encoding polypeptides of 77 and 118 amino acids, respectively. The longer of the two deduced amino acid sequences shows significant sequence similarity with the Vacuolar-ATPase G subunits of *Bos taurus* (82%), *Manduca. sexta* (75%), *Caenorhabditis elegans* (78%), *Neurospora crassa* (60%) and *Saccharomyces cerevisiae* (54%) over the entire lengths of these proteins. The novel gene described here has therefore been named ATP6G; the 6 denoting a Vacuolar-type H+ ATPase subunit (GDB nomenclature) and the G denoting a G subunit.

Figure 4.11 shows a multiple alignment of the two deduced amino acid sequences of ATP6G, together with orthologous amino acid sequences. The level of conservation between human ATP6G and the orthologous proteins in other species is particularly high over the first 50 amino acids. The truncated human protein (ATP6Galt) encoded by cDNA

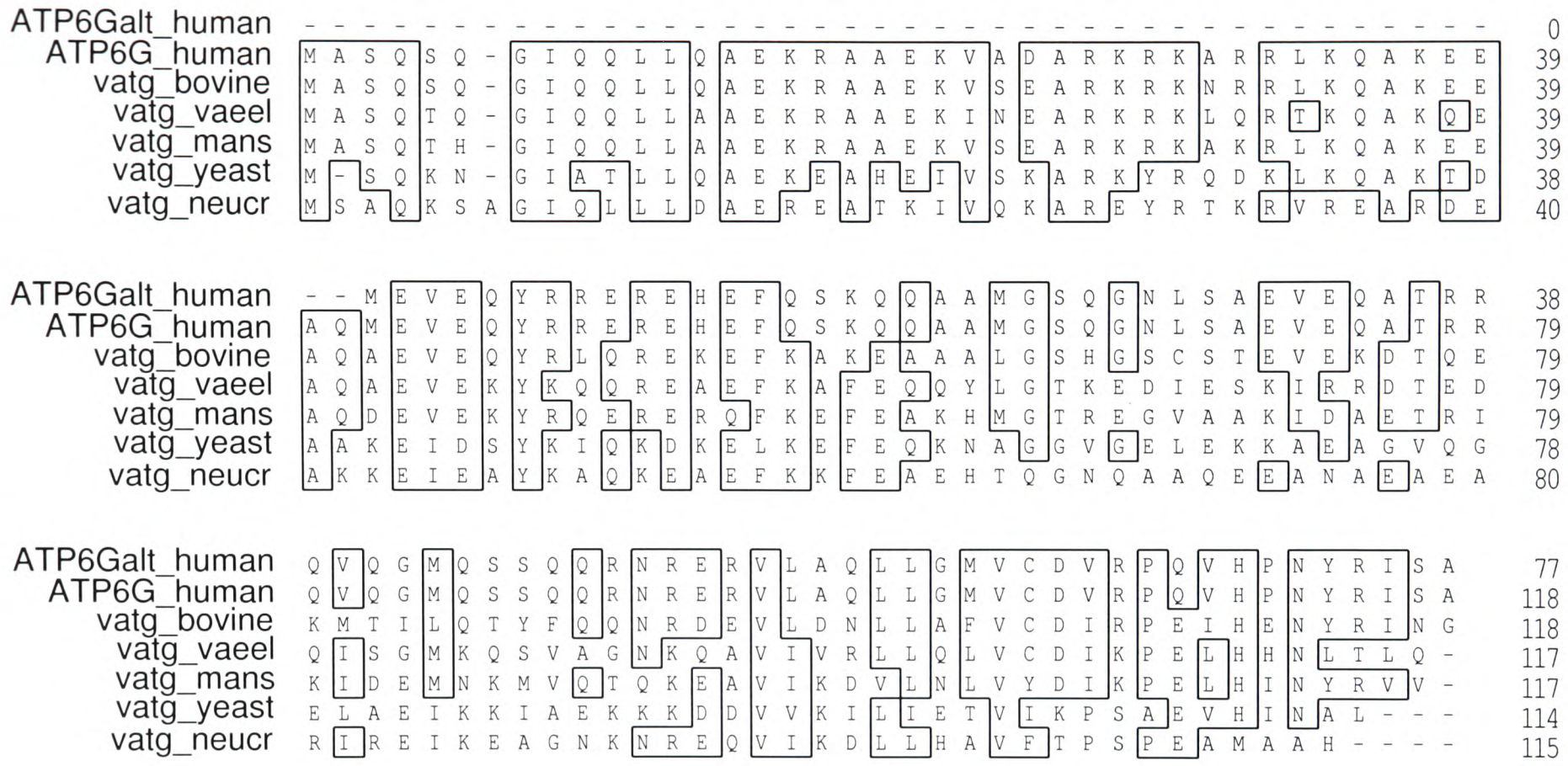


FIGURE 4.11 Multiple alignment of human ATP6G with orthologous V-ATPase G-subunits.
The top sequence, ATP6Galt_human is the truncated form of ATP6G corresponding to the smaller RT-PCR products in Figure 4.10B (panels I and II) and to the cDNA clone 726424. ATP6G_human is the full length polypeptide encoded by cDNA clone 124837 and corresponds to the larger RT-PCR products in Figure 4.10B (panels I and II). vatg_bovine corresponds to the V-ATPase G-subunit from *Bos taurus*, vatg_caeel from *Caenorhabditis elegans*, vatg_mans from *Manduca sexta*, vm10_yeast from *Saccharomyces cerevisiae* and vatg_neucr from *Neurospora crassa.*

clone 726424 lacks the first 41 of these residues since the putative initiation AUG codon encoded in clone 124837 is spliced out of clone 726424. The second initiation AUG codon, which is the putative initiation AUG codon of clone 726424 is located in exon 2 (Figures 4.9 and 4.10).

4.2.3.2 Expression profiling of the ATP6G gene encoded transcripts by RT-PCR

Expression analysis was carried out by RT-PCR using primers from the 2 alternative 3' UTRs of ATP6G (Table 4.3). Two PCR products were detected in each profile and the exon usage of these species was determined by sequence analysis (Figure 4.10B, panels I and II). The reverse primer used to generate the PCR products shown in panel I of Figure 4.10B is specific for species with the longer 3' UTR, as found in clone 124837 (Figure 4.10), while the products in panel II of Figure 4.10B were generated using a primer that does not discriminate between the alternative 3' UTRs as it is positioned within sequence common to both 3' UTRs, upstream of the first polyadenylation signal. Sequence analysis has confirmed that the upper band in both profiles corresponds to transcripts containing exon 1I (as found in clone 124837), while the lower band of both profiles corresponds to transcripts containing the alternatively spliced exon 1II (as found in clone 726424). Both variants of exon 1 splicing were detected with either the long or short 3' UTR. However, their expression varies across the 9 cell lines examined.

In Figure 4.10B panel I (amplification specific for the long 3' UTR) transcripts containing exon 1I and exon 1II show a similar pattern of expression, with the highest levels being detected in the lymphoid cell lines and little or no expression in the monocyte and macrophage-like cell lines (HL60 and U937, respectively). In contrast, in Figure 4.10B panel II (amplification of both the short and long 3' UTRs) the expression patterns of

transcripts containing exon I and exon II differ. The transcript with exon 1I (upper band) shows a similar pattern of expression to those in Figure 4.10B panel I. However, expression of the transcript with exon 1II (lower band) appears less specific, being detectable at significant levels in all the cell lines tested, the greatest level of expression being seen in the hepatocyte cell line HepG2.

TABLE 4.3 Transcript specific RT-PCR primers for the ATP6G gene.

PRIMER	SEQUENCE	TRANSCRIPT	POSITION [a]	PCR CONDITIONS
ACTINf1	5'-CTTCGCGGGC GACGATGG-3'	BETA ACTIN	+ve control	1mM $MgCl_2$, anneal 60° C, 30 cycles
ACTINr	5'-TGGTGGTGAA GCTGTAGCC-3'			
ATP6Gf	5'-AGGACCAGTC ATCAATAGGAGG-3'	ATP6Gall	72855-72876	1mM $MgCl_2$, anneal 60° C, 30 cycles
ATP6Gr1	5'-GGCAGAAGGA GTCAGGCC-3'	ATP6G all	74231-74214	
ATP6Gr2	5'-TTGGAGATCA GAAGGTCTCTGG-3'	ATP6G long	74869-74848	

[a] Positions of primers within the genomic sequence generated in this study.

4.2.3.3 Amino acid sequence analysis of the ATP6G gene encoded transcripts

The Predict Protein server predicted that ATP6G and ATP6Galt have a hydrophilic helix loop helix secondary structure with a short hydrophobic region at the C-terminus, but no transmembrane region. The Motif programme predicted a glycosylation site at Asn 68, a tyrosine kinase phosphorylation site at Tyr 77 and myristoylation sites at Gly 64, 67 and 83.

4.3 DISCUSSION

4.3.1 AIF-1, G1 and IRT-1

The human AIF-1 gene appears to encode 3 alternative transcripts, derived from differential splicing of a total of 9 exons (Figure 4.2). The distinctly different patterns of expression shown here for G1 and AIF-1, together with the fact that 3 ESTs correspond

exactly to G1, indicate that G1 and AIF-1 are genuine splice variants of the same gene, rather than G1 being an incomplete cDNA. The observation that the expression patterns of G1 and AIF-1 are almost exclusively restricted to immune related cell lines suggests that these transcripts are in some way involved in the immune or inflammatory responses. It is also clear from its splicing and specific expression pattern that IRT-1 is a third distinct splice variant of the AIF-1 gene (Figure 4.2B). IRT-1 has been shown to be expressed with a more restricted expression pattern than either G1 or AIF-1 (Autieri & Agrawal 1998) (below). However, we were not able to confirm this by RT-PCR.

It is evident from the large number of database entries aligning over this region that the AIF-1 gene is being extensively studied. It is also clear that many of these studies implicate AIF-1 in a variety of medical conditions (see below). The AIF-1 splice variant itself has been reported under a number of different names, i.e. in addition to AIF-1 it has been referred to as iba1 (ionised calcium binding adapter molecule) (Imai *et al.* 1996), MRF-1 (Microglia response factor) (accession no. AB000818) and Daintain (Chen *et al.* 1997), all of which have been isolated from rat. The AIF-1 transcript is an IFN-γ inducible molecule expressed within cells of the monocyte lineage (also found in this investigation), including CD86+ macrophages (Utans *et al.* 1996), dendritic and microglial cells (Chen *et al.* 1997). It has been associated with processes involved in chronic allograft rejection in both humans and rat (Utans *et al.* 1996). It has also been shown to both inhibit and enhance insulin secretion and has been found at high levels in macrophages isolated from pre-diabetic rat pancreatic extracts, but not in normal pancreatic extracts (Chen *et al.* 1997). Thus, an involvement of the protein product of the AIF-1 transcript in the progression of IDDM has been suggested. The IRT-1 (Interferon γ responsive transcript-1) splice variant, like the AIF-1 splice variant, is an IFN-γ-inducible factor. It was identified

by RT-PCR in vascular smooth muscle cells (VSMCs) stimulated with cytokines and is also induced by balloon angioplasty in rat carotid arteries (Autieri & Agrawal 1998). It is expressed constitutively in quiescent peripheral blood lymphocytes (PBLs), but is suppressed when these cells are stimulated to proliferate (Autieri & Agrawal 1998). Furthermore, over-expression of IRT-1 in human VSMCs results in inhibition of cell growth. On the basis of these data, Autieri & Agrawal (1998) suggested that IRT-1 is involved in negative regulation of cell growth within VSMCs and may regulate the progression of vascular proliferative diseases. The suppression of IRT-1 in proliferating cells may explain why this transcript was not detected in the RT-PCR experiments performed here, as the RNA used was from cell lines, which in contrast to cells from tissue preparations, are constitutively in a state of proliferation. Lastly, the BART-1 transcript is a novel splice variant of the AIF-1 gene found in rat which is also selectively and transiently expressed in response to vascular trauma and has been suggested to play a role in the early to middle stages of vascular restenosis (Autieri *et al.* 1996a).

A number of different features have been identified in the amino acid sequences encoded within the AIF-1 gene which may help to define the biological activities of its different protein isoforms. Firstly, a 44 amino acid segment shared by all the splice variants and homologs of AIF-1 contains a cluster of paired basic residues, KR-KK-(G)KR (Chen *et al.* 1997) (Figure 4.3). These dibasic motifs are among the cleavage motifs recognised by endoproteases and are found within many peptide hormone precursors (Bresnahan *et al.* 1990). The last potential cleavage site is preceded by a glycine residue, which is a characteristic amidation signal (Bradbury *et al.* 1982; Chen *et al.* 1997) and is also a feature of hormone precursor proteins. Interestingly, the IRT-1 splice variant contains a

further five paired basic residues (KR-KK-(G)RR-RK-RR) encoded within exon 4III, with the middle site being another possible amidation signal.

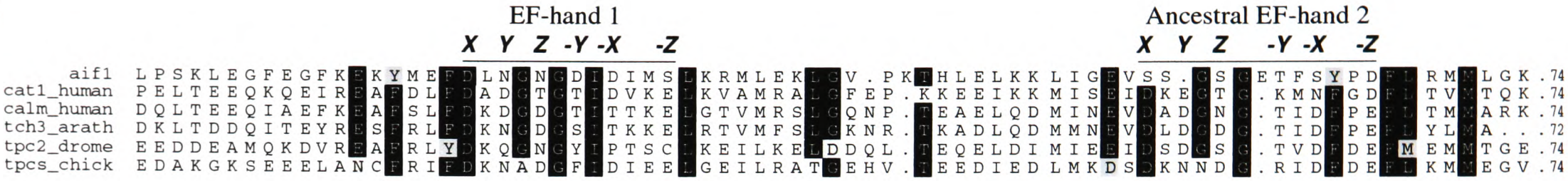
Residues 48 to 108 of the AIF-1 transcript show strong similarity to an E-F hand calcium binding domain, identified here and also discussed by Olavesen *et al.* (1991), Utans *et al.* (1995) and Imai *et al.* (1996). This type of calcium binding domain (Prosite PS00018) is characterised by a single globular domain containing a pair of highly conserved helix-loop-helix sub domains. The conserved motif itself is found within the loop and is made up of thirteen residues with the calcium co-ordination supplied by residues 1, 3, 5, 7, 9 and 12 (Tuckwell *et al.* 1992; Krebs *et al.* 1995) (Figure 4.12, below).

1	2	3	4	5	6	7	8	9	10	12	13	
X		*Y*		*Z*		*-Y*		*-X*		*-Z*		
D-x-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW]												

FIGURE 4.12 The PROSITE consensus sequence for the E-F hand loop.
The top line shows the residue numbers, the second line represents the calcium co-ordination residues and the third line shows the acceptable residues ([ ]) and unacceptable residues ({ }) at each position. x represents any residue.

In many cases only one of the sub-domains is functional (Boyhan *et al.* 1992; Tuckwell *et al.* 1992; Blanchard *et al.* 1997). In AIF-1 the second E-F hand between residues 91 and 104 is recognisable, but is almost certainly non-functional (Figure 4.3 and Figure 4.13). These degenerate EF-hand sub-domains are generally termed 'ancestral EF-hands'. Figure 4.13 shows a multiple alignment of the E-F hand domain of the protein encoded by the human AIF-1 transcript with a number of other E-F hand domains, together with the consensus secondary structure prediction generated by the JPRED server. Figure 4.13 clearly shows the predicted α-helix structure surrounding both the E-F hand loop motif, and the degenerate (ancestral) second E-F hand domain. The conserved E-F hand motif of

A



B

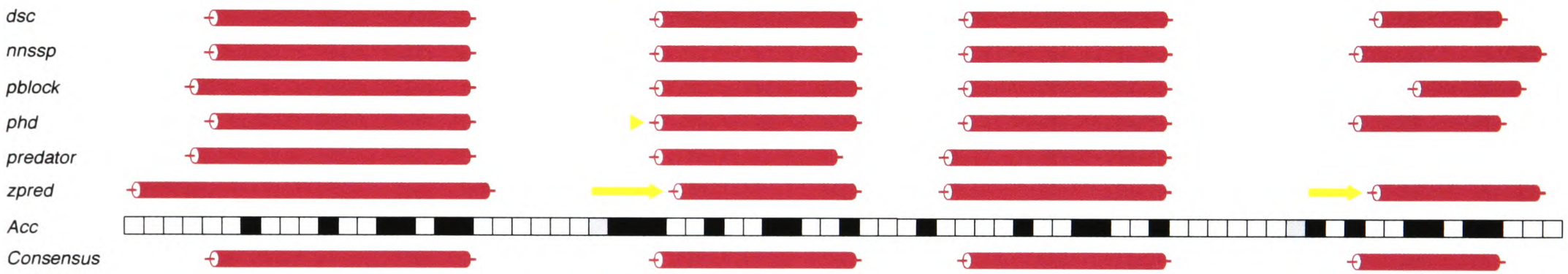


FIGURE *4.13* The putitive EF-hand calcium binding domain of AIF-1.
A. The multiple alignment of the EF-hand domain of AIF1 and various other proteins. The amino acid sequence cat1_human represents the EF-hand domain of human Caltractin (accession number P41208), calm_human represents human Calmodulin (accession number P02593), tch3_arath represents Arabidopsis Calmodulin (accession number P25017), tpc2_drome represents *Drosophila* Troponin C (accession number P47948), and tpcs_chick represents chicken Troponin C (accession number P02588). Both putitive EF-hand loops are indicated above the amino acid sequence as also shown in Figure 4.3. The 6 residues involved in calcium binding are designated *X, Y, Z, -Y, -X, -Z.* following the EF-hand consensus (Figure 4.12) **B**. The output from the JPRED-consensus secondary structure prediction server (http://circinus.ebi.ac.uk:8081/). The cylinders represent predicted α-helices and the arrows predicted β-strands*; dsc, nnssp, pblock, phd predator, zpred* are the different prediction methods used by JPRED and the consensus of these predictions is presented on the bottom line. The line denoted *'acc'* represents the accessability of the amino acids, a black box signifies a predicted buried residue and a white box an exposed residue.

AIF-1 (shaded in Figure 4.3) deviates from the consensus for this motif at its 12th (-Z) position which is almost exclusively an acidic residue, but in AIF-1 is a serine (Figure 4.13). Calcium is a 7 co-ordinate atom, therefore, both oxygen atoms from the carboxylate group of either aspartic acid or glutamic acid at this -Z position are usually utilised to complete the heptagonal co-ordination. This is not possible with serine. This does not necessarily rule out the binding of calcium by AIF-1, as there are a number of examples of functional E-F hand domains that deviate from the consensus at this 12th residue (Boyhan *et al.* 1992; Tuckwell *et al.* 1992; Blanchard *et al.* 1997). There is, however, no data so far as to whether AIF-1 exhibits calcium binding capabilities.

The question of the calcium binding abilities of AIF-1 is further complicated by the splicing of the G1 and IRT-1 transcripts. Both these variants, although still containing the conserved E-F hand loop, are truncated at their N-termini (Figure 4.3) resulting in the absence of most of the first helix of this E-F hand sub-domain, which may affect calcium binding. More significantly though, a 66 amino acid insertion within the IRT-1 splice variant falls within the conserved loop of the E-F hand (Figure 4.3). Also, if the dibasic endoprotease cleavage motifs are active the whole E-F hand domain of IRT-1 would be disrupted. If the G1 and IRT-1 transcripts are functional then it seems they may lack the calcium binding properties predicted for AIF-1.

Of additional interest in regard to IRT-1 is the identification of a potential leucine zipper motif within a predicted α-helix ($LX_6LX_6LX_6L$) (Figure 4.3B) (Landschulz *et al.* 1988), which was also reported by Autieri & Agrawal (1998). A leucine zipper is a protein domain that promotes dimer formation. The leucine residues, typically between 4 and 6, are spaced 7 amino acids apart, aligning on the same face every second turn of an α-helix

(Busch & Sassone Corsi 1990). The leucine side chains extending from the α-helices of two similar polypeptides can then inter-digitate, facilitating dimerisation (Landschulz *et al.* 1988). Interestingly, in IRT-1 the putative leucine zipper is preceded by a highly basic region (Figure 4.3B). This, with the leucine zipper is characteristic of a bZIP DNA binding domain (Busch & Sassone Corsi 1990; Pu & Struhl 1991). The leucine zipper facilitates dimerisation of two bZIP domains, putting the basic regions of the two polypeptides in close proximity to each other to form a specific DNA binding domain. These bZIP domains are common to many DNA binding regulatory proteins, and transcriptional elements, such as the Jun, Fos and C/EBPb proteins (Landschulz *et al.* 1988; Busch & Sassone Corsi 1990; Pu & Struhl 1991; Koldin *et al.* 1995; Novotny *et al.* 1998).

There is also evidence to suggest that IRT-1 may be targeted to the nucleus. PSORT identified a potential 4 residue nuclear targeting sequence within the IRT-1 specific sequence (Figure 4.3), similar to that found in the SV40 large T antigen (Kalderon et al 1984) (Autieri & Agrawal 1998). In addition to this, IRT-1 is a basic, charged molecule (pI=9.89 (Autieri & Agrawal 1998)) which is a feature of nuclear proteins. Furthermore, Autieri & Agrawal (1998) reported the localisation of IRT-1 to the perinucleus of VSMCs. The existence of a potential bZIP motif and the evidence for nuclear targeting, suggest that IRT-1 may have a DNA binding and regulatory role within the cell, perhaps acting as a transcriptional element.

The rat BART-1 transcript represents another interesting splice variant of the AIF-1 gene. The work of Autieri *et al.* (1996a) and analysis carried out here with the TMpred server both predict that BART-1 encodes an integral type 2 transmembrane protein. However, as

mentioned in the results (section 4.2.1.2) there are discrepancies between the sequence of BART-1 and other AIF-1 transcripts, suggesting that the BART-1 database entry may contain a number of errors. The encoded amino acid sequence of this transcript should, therefore, be regarded with caution until the sequence is confirmed. BART-1 is the most divergent of all the AIF-1 gene transcripts. On alignment of the BART-1 amino acid sequence with the human genomic sequence (Figure 4.14) the BART-1 transcript specific sequence, i.e. the sequence between AIF-1 exons 3 and 4 is not conserved within the human AIF-1 intron 3 sequence, whereas the AIF-1 exonic sequence is. Therefore, sequence encoding a membrane spanning region may be unique to rat. Alternatively, as the sequence goes into AIF-1 intronic sequence it may be that the BART-1 transcript is a pre-mRNA of AIF-1.

Work undertaken so far suggests that the polypeptides encoded by the different AIF-1 transcripts may be involved in the inflammatory response, in the regulation of cell proliferation and in a number of disease processes including diabetes, organ transplant rejection and vascular restenosis. The high degree of sequence conservation at the amino acid level between the AIF-1 proteins in species as divergent as human and fish (Figure 4.3) also suggests that AIF-1 plays a functionally important role in the cell, although expression patterns suggest that this may be different in different cell types. The function(s) of the AIF-1 gene products remain to be determined experimentally. In the meantime the sequence of these proteins present a confusing collection of clues. There is clearly a strong similarity to an EF-hand calcium binding domain; the probability of the EF-hand like pattern occurring by chance is remote. Yet there is also evidence to suggest that the IRT-1 transcript encodes a DNA-binding bZIP protein that is targeted to the nucleus. In the IRT-1 splice variant it is most unlikely that the EF-hand motif is functional

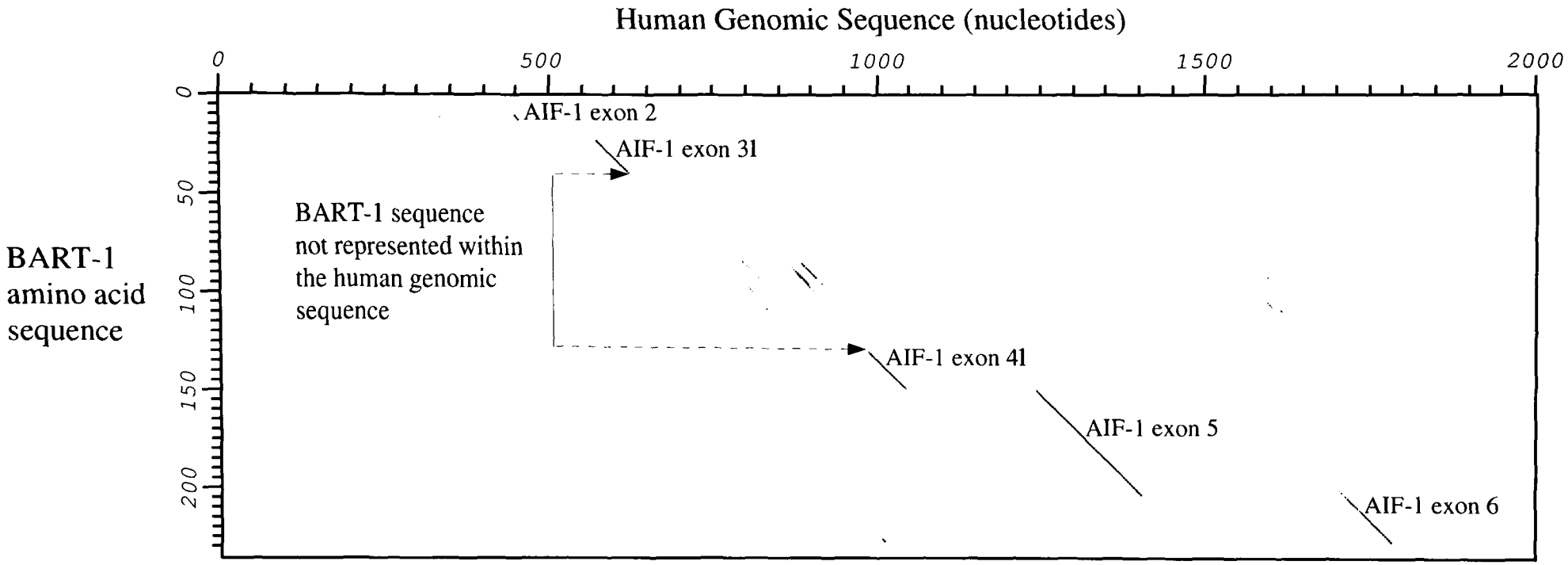


FIGURE 4.14 The dot-matrix alignment of the rat BART-1 transcript amino acid sequence against the human AIF-1 genomic region DNA sequence. The novel rat AIF-1 transcript BART-1 contains coding sequence unique to this transcript. The dot-mastrix alignment also demonstrates that this sequence is unique to the rat genome (compared to the human genome) and found between human AIF-1 exons 3I and 4I. If the complete cDNA sequence was represented within the human genomic sequence then the only significant breaks in the diagonal line would be along the horizontal axis, representing introns in the genomic sequence. In this case though, there is also a gap along the vertical axis of approximately 90 amino acids.This represents the rat BART-1 amino acid sequence that does not have orthologous couterpart within the human genome.
The horizontal axis relates to the human AIF-1 genomic sequence (nucleotides) and the vertical axis relates to the rat BART-1 sequence (amino acids).

as the EF-hand motif is completely disrupted by a 66 amino acid insertion. Conversely, in the AIF-1 transcript there is no bZIP motif and only three out of the four leucines from within the potential leucine zipper motif are present. Furthermore, if the endoprotease and amidation signals are active, both the EF-hand and bZIP motifs of the different AIF-1 genes transcripts would be disrupted. However, the rapid degradation of a regulatory molecule, as AIF-1 may turn out to be, by endoproteases may be advantageous. It seems possible that AIF-1 may be a multifunctional gene with each of its different transcripts being involved in distinct effector mechanisms.

Taken together, these observations highlight AIF-1 as a very interesting gene and a possible candidate disease susceptibility locus warranting further study. Future work will no doubt shed light on the significance of the different splice variants and the functions of their protein products.

4.3.2 1C7, A novel member of the immunoglobulin super-family

All members of the Ig superfamily share a common structure, comprising two β-sheets with a core consisting of β-strands A, B and E in one sheet and G, F and C in the other (Williams & Barclay 1988). This superfamily can be subdivided into the V, C1, C2 and I Ig sets which are distinguished on the basis of other features (Williams & Barclay 1988). Harpaz and Chothia (1994) defined a group of 46 residues occupying 20 key sites that form the characteristic folds of the IgV set, which they called the " V frame". Alignment of the 1C7a, 1C7b and 1C7c Ig-like domain with other IgV domains (Figure 4.8A) highlights the conservation of all the key "V frame" residues in these 1C7 isoforms. In addition, the Ig-like domain is encoded by an exon with a phase 1 splice junction at either

end (Figure 4.7B), another feature common to Ig domains. This provides convincing evidence that 1C7 is a new member of the Ig superfamily.

Of particular interest is the identification of the transcripts 1C7d, 1C7e and 1C7f (Figure 4.6B). In these transcripts the Ig-like domain is split over 2 exons (exons 2I and 2II) and the central segment of exon 2, covering 75 nucleotides, is spliced out. The splicing of exon 2I to exon 2II in these transcripts, encodes a truncated form of the Ig domain (25 amino acids shorter over its central region) with features characteristic of the IgC2-set rather than IgV-set. The IgC2-set is characterised by the retention of the core β-strands found in all Ig domains and the conserved E-F loop characteristic of the IgV-set, but lacks the central region of the V-set domain containing β-strand D and most of β-strand C' (Williams & Barclay 1988) (Figure 4.8B). As mentioned above, one of the features of Ig domains is that the majority are encoded by a single exon, which is not the case here for the 1C7 IgC2 domain. However, there are exceptions to this rule and examples of Ig domains that are divided over 2 exons are found in molecules including CD4, Po, PolyIgR, NCAM (Williams & Barclay 1988) and the third domains of the fibroblast growth factor receptors (FGFRs) I, II and III (Bateman & Chothia 1995). In the case of the FGFRs, the IgIII domain is the product of the splicing of exon IIIa with either exon IIIb or exon IIIc, the alternative products having different affinities for fibroblast growth factors (Bateman & Chothia 1995). Significantly, in nearly all cases where an intron exists within an Ig domain the splice junction is in a phase other than phase 1, as is the case with 1C7d, 1C7e and 1C7f (Figure 4.7B). This would preclude the functional splicing of just half the domain within the same gene, as is the case for bands 5, 6 and 9 in Figure 4.6B (panels I and II). In Figure 4.6B panel II, the truncated, and almost certainly non-functional, lower band (band 9) is expressed at higher levels than the C2-like domain encoding 1C7f (band

8). The detection of a transcript at the RNA level does not necessarily indicate the expression of a functional protein. However, if both the alternative Ig-like domain-containing forms of 1C7 do turn out to be functional, then, this would appear to be the first identified example of a gene capable of Ig-set switching. Furthermore, the existence of the splicing pattern observed here raises interesting questions regarding the evolution of the immunoglobulin superfamily (IgSF) (see section 4.3.2.1, below).

The mature 1C7 polypeptide is predicted to be a type-I integral membrane protein with no cysteine residues N-terminal of its membrane spanning region, other than the two within its single Ig-like domain. This suggests that the 1C7 protein exists as a monomeric membrane bound molecule (Figure 4.15). Other Ig superfamily members with a similar structure that exist as monomers include THY-1, CD83, CD7 and CD79a (Barclay *et al.* 1997). As mentioned above in relation to the Ig domain, 1C7 also exhibits a complex pattern of alternative splicing (Figure 4.6). However, transcript bands 5, 6 and 9 in Figure 4.6B are unlikely to encode a functional protein and transcripts 1C7a, 1C7b, 1C7c and 1C7f were the only transcripts consistently detected at high levels. Interestingly, exon 4III in transcripts 1C7a and 1C7e encodes a number of potential SH3 domain binding motifs (Figure 4.7A). The SH3 domain binding motif is a proline rich region that is thought to interact with SH3 domain-containing proteins during tyrosine kinase receptor activation (Scaife *et al.* 1994) as part of a signal transduction cascade. Although attempts to identify a consensus for this motif have revealed a degree of variation, an overall consensus of: 'XpΦPpXP' within a proline rich region emerges, where 'P' represents a conserved proline, 'p' a residue tending to be a proline, 'X' a non-conserved residue and 'Φ' a hydrophobic residue (Yu *et al.* 1994). The sequence encoded by exon 4III, highlighted in Figure 4.7A, fits this consensus as well as many other experimentally confirmed SH3

domain binding motifs. Thus, at least this variant of exon 4 contains features suggestive that it is associated with the expression of a functional protein. Added to this, Pende *et al.* (1999) have shown the product of exon 4I to be expressed at the protein level (see below).

The expression patterns of the 1C7 transcripts further suggest that this gene encodes protein(s) involved in the immune response (Figure 4.6B). The highest levels of 1C7 transcripts were consistently detected in the cell line Raji (B cell), with slightly lower levels in the cell line Molt4 (T cell). At lower levels again, transcripts were also detected in the cell line HL60 (Monocyte) (Figure 4.6B). The fact that no signal was detected by Northern blot analysis suggests that all transcripts are expressed at relatively low levels in the cell lines used, as Northern blot analysis is not as sensitive as RT-PCR and would, therefore, not detect low levels of expression.

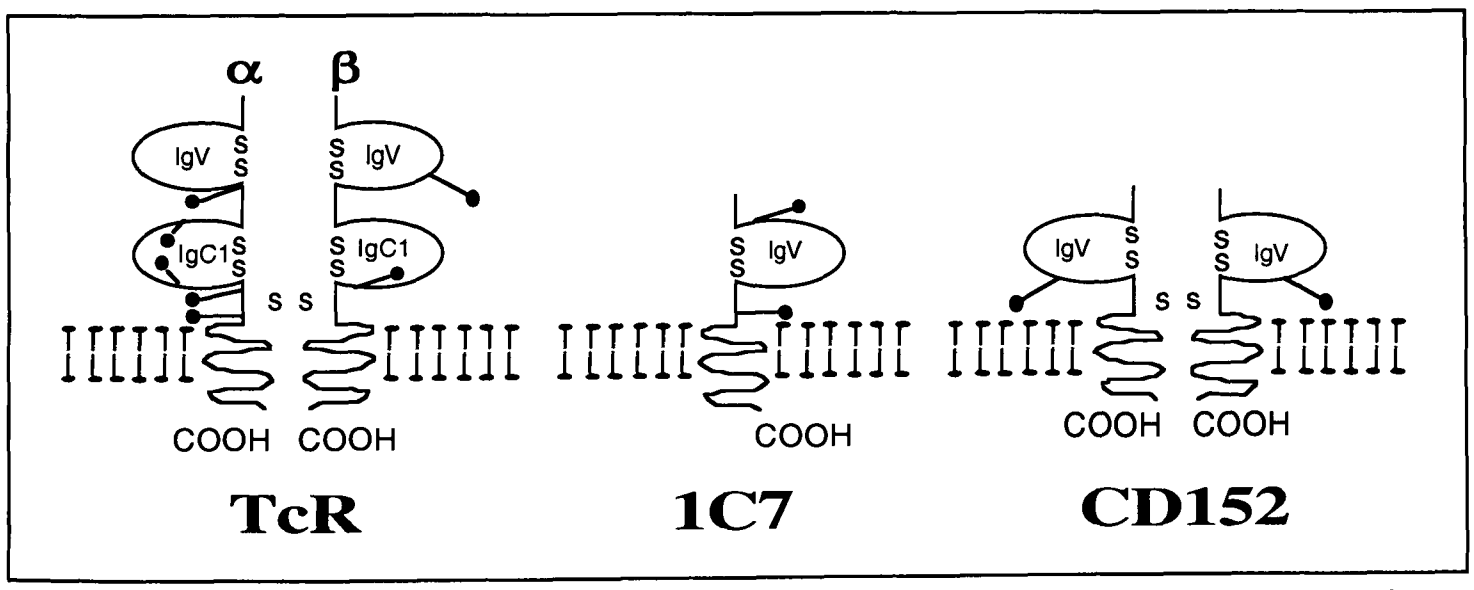


FIGURE 4.15 The putative structure of 1C7 with the structures of the cell surface molecules CD152 (CTLA-4) and the T Cell receptor (TcR) for comparison. The variable domains are represented by IgV, the constant domains of TcR by IgC1 and the disulphide bonds by S S. The diagrams of CD152 and TcR were adapted from those in Barclay *et al.* (1997).

Recent work by Pende *et al.* (1999) (as referred to in the above text) identified a novel natural killer (NK) cell triggering receptor called NKp30. This receptor has been identified

as the product of the 1C7 gene discussed here and, more specifically, it represents the product of the 1C7c splice variant. NKp30 is one of three NK triggering receptors (the other two being NKp46 and NKp44) which have been shown to co-operate in the induction of NK-mediated cytotoxicity against the majority of target cells. NKp30 represents the major triggering receptor for a number of tumours, including the human melanoma, MEL15392 (Pende *et al.* 1999). It has also been shown that NKp30 associates with CD3ξ, which becomes tyrosine phosphorylated upon sodium pervanadate treatment of NK cells. As reported here and in Neville & Campbell (1999), Pende *et al.* (1999) describe NKp30 as a single IgV domain-containing integral membrane protein. NKp30 was detected by Northern blot analysis in spleen and NK cells but was not detected in lymphoid cell lines (i.e. the cell lines in which we detected 1C7 transcripts with RT-PCR) and was only detected on the surface of subsets of NK cells (Pende *et al.* 1999). This is consistant with our lack of success in identifying 1C7 transcripts in lymphoid cell lines by Northern blot analysis and Pende *et al.* (1999) go on to suggest that NKp30 is only expressed at significant levels in NK cells. This last statement, however, is not entirely supported by the data from the original EST database search in which a number of clones of B-cell origin containing the 1C7 transcript were identified. Also, in the paper by Pende *et al.* (1999) there was no mention of alternative splicing. Therefore, there is no data on the expression of the IgC2 domain or on the alternative C-termini, including the product of exon 4III which contains the potential SH3 domain binding motifs. It may well be that the expression of 1C7 on the surface of other cell types and the expression of the different spliced forms are restricted to specific conditions.

Of further interest is the lack of an open reading frame corresponding to 1C7 and the loss of the consensus polyadenylation signal in the Murine genomic sequence (Accession

number AF109719) (see section 3.2.2.6). At least two mouse strains have been looked at so far and in both cases 1C7 appears to be a pseudogene (M. Hollyoake, personal communication). A rat EST (AA894045) has been identified that does retain the conserved polyadenylation signal. This EST is incomplete, however, so no assessment can be made as to whether it contains a complete open reading-frame corresponding to 1C7. Interestingly, on the basis of the sequence available the rat transcript appears to be an ortholog of 1C7c and the most well conserved sequence between man and rat is around the third hypervariable region of the Ig domain, between β-strands F and G. The conservation of this hypervariable region may well be significant in terms of the ligand binding specificity of 1C7. Lastly, the identification of a non-functional gene in mouse and possibly a functional gene in rat will be a useful resource when considering animal models to further characterise the function of this gene.

The identification of 1C7 as a putative novel member of the Ig superfamily, its immune related expression pattern and the findings by Pende *et al.* (1999) highlights this as a potentially important candidate disease susceptibility gene. Clearly, further work is needed to confirm the expression of the various isoforms of 1C7 at the protein level and to resolve their tertiary structures as proof of their Ig superfamily membership. Also, although Zoo blot analysis has identified orthologous sequences in monkey, rat, mouse, dog, cow and rabbit (Pende *et al.* 1999) 1C7 may not be functional in all these species.

4.3.2.1 Speculations on the evolution of the IgSF from the 1C7 Ig domain splicing

As mentioned above, the IgSF is divided into four structural sets: the Ig V-, C1-, C2- and I-sets (Williams & Barclay 1988; Harpaz & Chothia 1994). The first divisions to be identified were between the V-set (variable domains) and the C-set (constant domain) (Hill

et al. 1966). However, despite the differences between these two groups, it was noticed that sequence and structural similarities indicated that they had a common ancestry (Edelman 1970; Poljak *et al.* 1973; Edmundson *et al.* 1976). Further analysis defined the IgC2-set, which includes domains that share some V-set sequence features but are the same size as the C-set domains (Williams & Barclay 1988). The original C-set was subsequently renamed the C1-set. Most recently the pool of sequences that originally fell into the C2 set were further divided between the C2-set and the I-set (Intermediate) (Harpaz & Chothia 1994) with the I-set representing a structural intermediate between the C1-set and the V-set.

There has long been speculation regarding the evolution of the IgSF, but due to the low level of conservation between Ig domains the task of building an accurate phylogenetic tree has proven extremely difficult (Singer & Doolittle 1966; Bourgois 1975; Edmundson *et al.* 1976; Williams & Barclay 1988; Bateman 1997; Smith & Xue 1997). Originally, it was believed that the Ig domain progenitor was a gene encoding a single V-like domain, much like THY-1 (Williams & Barclay 1988), as at that time the only examples of single domain non-associated molecules were of the V-set. More recent analysis, however, suggests that the Ig domain originated from either a C2-like (Smith & Xue 1997) or I-like (Bateman 1997) ancestral domain as they both share features with the V-set and C-set. What is particularly interesting though, in relation to 1C7, is the hypothesis that the Ig domain may have evolved from two 'half-domains' (Bourgois 1975; Murray *et al.* 1995). Both Bourgois (1975) and Murray *et al.* (1995) suggest that a precursor molecule comprising 3 or 4 β-strands may have been duplicated, or that two similar structures became associated and intercalated, resulting in the two β-sheet structure now recognised as the Ig domain. Bourgois (1975) adopted a sequence based approach, demonstrating that

intra-domain sequence conservation exists around the cysteine residues within V-set domains, suggesting a duplication event involving an ancestral cysteine containing molecule. Murray *et al.* (1995) on the other hand, base their hypothesis on observations of the folding of the CD2 membrane glycoprotein and its ability to form a metastable dimeric form *in vitro*. Despite the lack of the characteristic cysteine residues which normally form the stabilising disulphide bond between the two β-sheets, two CD2 Ig domains were found to intercalate with each other to form two Ig-like domains with β-strand pairs (A/D, B/E, C/F and C'/G) and with a frequency greater than that due to chance (15%) (Murray *et al.* 1995). The splicing pattern of 1C7 is interesting in that each of the exons 2I and 2II contain a half-domain, exon 2I encoding β-strands A, B and C and exon 2II encoding β-strands E, F and G. It is the intronic region between these two exons that encodes the V-set specific sequence. It is conceivable that during the evolution of the IgSF two such exons came together within the same genomic segment in much the same way as is seen in 1C7. Then, increasing ambiguity within the splice site sequences may have caused miss-splicing between the two exons, resulting in the inclusion of the intronic sequence, encoding the additional amino acid sequence. This splicing then became advantageous to the organism, marking the transition between an ancestral C2-like domain and a V-like domain. Then, over time, the loss of splice site recognition sequence altogether would have resulted in a single exon encoding a V-like domain. This is clearly only speculative and in no way relates to the origins of the 1C7 gene itself, but it does present an interesting model to explain the process by which the IgV-sets may have evolved from a IgC2-set and how this may relate to the half-domain hypothesis.

4.3.3 A Putative Human Homolog of the Tobacco Hornworm V-ATPase G Subunit

Analysis of the ~7kb gap between IKBL and BAT1 has led to the identification of the new gene, ATP6G: a homolog of the novel G-subunit of a *Manduca sexta* (Tobacco Horn Worm) V-ATPase. The Vacuolar-type ATPases are H+ translocating ATPases found in endomembranes and various plasma membranes of eukaryotic cells (Harvey 1992) and are involved in a broad range of functions including bone re-absorption; glycosylation in the Golgi; degradation of cellular debris in lysosomes; and the processing of endocytosed receptor-ligand complexes (Crider *et al.* 1997).

The V-ATPases are heteromultimeric enzymes consisting of V_1 peripheral complexes and V_0 membrane complexes. Together these two complexes contain up to 6 constitutive V-ATPase subunits, which have been identified at the molecular level (for full summary see Harvey 1992). The G subunit is one of two known peripheral components of the V1 catalytic ATPase complex involved in the catalysis of ATP hydrolysis (Lepier *et al.* 1996; Crider *et al.* 1997), the second peripheral subunit being the H subunit. The G subunit has so far been characterised in Tobacco Horn Worm, rat, chicken and cow. Both the G and H subunits have been shown to participate in ATPase activity, but rather than acting together it seems that either subunit alone can initiate ATP hydrolysis (Crider *et al.* 1997). Furthermore, the two subunits appear to be expressed in different tissue types (Crider *et al.* 1997). Thus, it is likely that they confer the same activity, but in different locations. In this investigation, the absence of RT-PCR products corresponding to the longer form of the ATP6G protein in the monocyte and macrophage cell lines (Figure 4.10B panel I) may suggest that in these cell lines the H rather than the G subunit is constitutively expressed.

The significance of the alternate polyadenylation sites and signals in ATP6G is not clear. For both polyadenylation sites the second potential polyadenylation signal is the most probable as active polyadenylation signals are generally found between 10 and 30 nucleotides up-stream of the polyadenylation site (Wahle & Keller 1992). This would preclude the first potential signal in both cases (Figure 4.9). The existence of two alternate polyadenylation sites presents another question. Of the two sites, only the down-stream site is in close proximity to a conserved AATAAA polyadenylation signal motif and, therefore, should be the most favourable. However, clone 726424 does contain a significant tract of adenines at its 3' end, suggesting that it has been cleaved and polyadenylated. The most likely signal motif is the TATAAA motif 20 nucleotides upstream of the cleavage site for this transcript. Variants of the polyadenylation signal have been observed (Wahle & Keller 1992) and point mutation analyses of the polyadenylation signal sequence (Wilusz *et al.* 1989; Sheets *et al.* 1990) have shown that the most common variation found, ATTAAA, works with between 60% and 80% efficiency when compared to AATAAA (Sheets *et al.* 1990). The sequence found here has a much lower efficiency, however, of approximately 17% (Sheets *et al.* 1990). How these two signals would compete is not clear. Maybe the reduced efficiency of a weak signal is offset by the negative effect of the extended 3' untranslated region of the longer transcript. The site of 3' cleavage and polyadenylation is also directed by elements down-stream of the cleavage site. There is no distinct motif for such down-stream elements, although G-, U-, or GU- rich sequences have been implicated (Wilusz *et al.* 1989; Wahle & Keller 1992). In addition to this, work with the Hepatitis B viruses (Hepadnaviruses) has revealed the existence of multiple upstream enhancer elements (Russnak & Ganem 1990). A number of hepadnaviruses were identified which carried the variant polyadenylation signal seen in ATP6G, but showed no marked reduction in polyadenylation efficiency. However,

the deletion of upstream elements effectively abolished proper polyadenylation site usage (Russnak & Ganem 1990). It may be that the upstream polyadenylation site of ATP6G, although having a less conserved polyadenylation signal, is associated with additional upstream or down-stream elements. Of course, it cannot be discounted that the poly-A tail of this cDNA clone is just an artifact of the cloning process and that this is in-fact an incomplete cDNA. This would not explain the alternate splicing of exon 1 of the cDNA clone in question, however.

A further point in relation to the 2 variants of the ATP6G found here is that the level of amino acid sequence conservation between the G and H subunits and between the G subunits of different species is particularly high over the first 50 amino acids (see Figure 4.11) suggesting that this region is an important functional domain. This would preclude the truncated isoforms of human ATP6G from being functional.

The function of the peripheral V-ATPase subunits is not yet completely understood, although they have been shown to be involved in initiation of ATP hydrolysis (Crider *et al.* 1997). In yeast the 13-kDa homologous G-subunit (Vma10) co-purifies with the V_0 complex (Superkova *et al.* 1995) and not the V_1 complex as was found by Lepier *et al.* (1996) for the *M. sexta* G-subunit. This may suggest that the G-subunit associates with both the V_0 and V_1 complexes (Lepier *et al.* 1996). Therefore, this subunit may actually link the two complexes: the conserved, highly charged and hydrophilic N-terminus associating with the V_1 complex and the less hydrophilic C-terminus either associating with the Vo complex or anchoring it to the cell membrane. The covalent addition of fatty acids such as myristic acid may aid in anchoring the subunit to the membrane. A myristoylation site at Gly 64 is conserved in all the homologs, except Vma10, Gly 67 is

conserved between *B. taurus* and the human protein, and Gly 83 is conserved between the human and *C. elegans* proteins. Therefore, if a function of the G-subunit is to link the V1 and Vo subunits together then the truncated ATP6Galt product may act as a competitive inhibitor of ATP6G.

Of particular interest in view of the location of ATP6G in the MHC class III region is the suggested role of V-ATPases in the inflammatory and immune responses. Swallow *et al.* (1990) demonstrated the importance of plasmalemmal V-ATPases in pH homeostasis in macrophages, particularly under conditions of extracellular acidification that may occur during the inflammatory response. Following on from this, Brisseau *et al.* (1996) have shown plasmalemmal V-ATPase activity to be modulated by IL-1 in a dose-dependent fashion. They went on to suggest this as a process by which V-ATPases are modulated in order to provide a pH regulatory mechanism during the inflammatory response, thus countering intracellular acid accumulation and resultant cellular dysfunction (Brisseau *et al.* 1996).

Further studies are needed to determine whether the mRNA levels of the V-ATPase G-subunit are transcriptionally up-regulated in IL-1 treated cells and whether any response correlates with the dose response activity observed for the V-ATPase in macrophages. It also needs to be determined whether the truncated isoform of ATP6G is functional and whether expression of this isoform might inhibit the activity of the full length subunit.

Chapter 5

CHAPTER 5: INVESTIGATION OF POLYMORPHISMS WITHIN THE AIF-1, 1C7 AND ATP6G GENOMIC REGIONS AND WITHIN THE BAT1 PROMOTER REGION.

5.1 INTRODUCTION

A primary goal of genome research is the understanding of the genetic basis of disease. There are a number of scientific approaches available to aid this understanding and, among these, the genomic screening and linkage analysis of affected families is one of the most widely used. Linkage analysis has defined the aetiologies of many monogenic diseases (Riordan *et al.* 1989; Feder *et al.* 1996) and, to a lesser extent, has contributed to the understanding of a number of complex multifactorial diseases (Davies *et al.* 1994; Schroeder *et al.* 1998). However, this method has its limitations. Large numbers of affected kindreds are usually required and are not always available. Furthermore, the ability to detect multi-gene effects and minor gene interactions is limited. In addition to this, some areas of the genome are in linkage disequilibrium restricting the fine mapping of disease loci in such regions, as is the case with the MHC region (see section 1.1.4). A complementary approach to genome screening is the analysis of variation within individual 'candidate genes' hypothesised to be involved in disease pathogenesis (Mullighan *et al.* 1999) and/or which fall within a genomic region bordered by markers associated with a disease phenotype. It is this second approach that is the subject of this chapter.

As discussed in Chapter 1, there is evidence to suggest the existence of a number of disease susceptibility loci within the MHC class III region (section 1.1.4.1). Of particular importance to this work, and the reason for choosing Cosmids TN62 and TN82 for sequence analysis (section 1.2), is the possibility of disease susceptibility loci within the

telomeric end of the class III region (section1.1.4.2). There is evidence for the existence of loci which may contribute to the progression of: IDDM (Moghaddam *et al.* 1997), MG (Degli Esposti *et al.* 1992b) and, most convincingly, CVID (Schroeder *et al.* 1998). Chapters 3 and 4 of this Thesis describe the identification and characterisation of coding regions within the TN62/TN82 genomic region with the aim of defining new disease candidate genes. Following this, an important aim was to identify polymorphisms within such candidate genes. To this end, the three genes analysed and discussed in Chapter 4 were selected for mutation analysis: AIF-1, due to its suggested involvement in a number of disease processes (Utans *et al.* 1996; Chen *et al.* 1997; Autieri & Agrawal 1998) including IDDM (Chen *et al.* 1997); 1C7, as its leukocyte specific expression pattern and its identification as an immunoglobulin domain-containing cell surface molecule makes it a strong candidate gene for CVID; and ATP6G, a novel gene identified in this study. In addition to this, a 15 nucleotide deletion 5' of the BAT1 coding region, identified by comparison of the TN62/TN82 genomic sequence with the EMBL database sequence AC004181, warranted confirmation and further characterisation.

5.2 MUTATION DETECTION WITHIN THE AIF-1, 1C7 AND ATP6G GENOMIC REGIONS AND CHARACTERISATION OF THE BAT1 5' DELETION.

5.2.1 Methodology

The AIF-1, 1C7 and ATP6G genomic regions, and the BAT1 promoter region, were screened for mutations. This was achieved by direct sequencing of PCR products generated from genomic DNA, as described in section 2.3.14. The principal of a DNA sequencing based approach to mutation analysis involves the screening of sequence trace files by eye for heterozygous mutations which are discernible as two overlapping peaks of approximately half the height of neighbouring peaks. This approach was adopted as it is very accurate and informative and utilised techniques already familiar to the author.

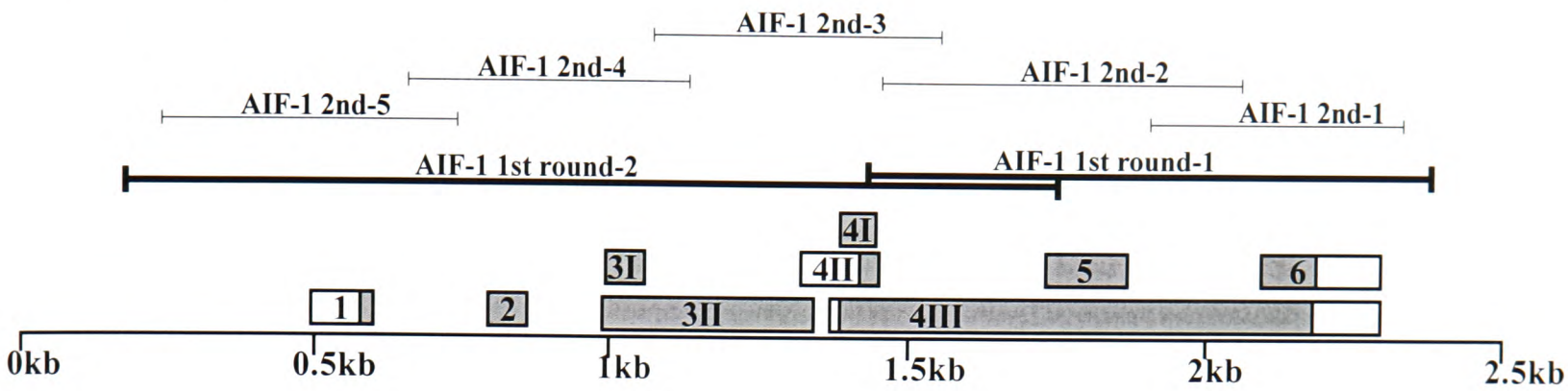
Reliable sequence based mutation detection necessitates the generation of sequence trace files with uniform peak heights and low background so that confident identification of heterozygous mutations can be achieved. This has become more feasible in recent years due to advances in the dye-terminator sequencing chemistry. Useful sequence data was previously only attainable with the dye-primer chemistry which is a less robust chemistry and required much more stringent optimisation and purification of the PCR product making it a time consuming technique. The chemistry used in this investigation was the ThermoSequenase dye-terminator cycle sequencing pre-mix kit, version 2.0 (Amersham). It was found that after a simple dilution step PCR products could be sequenced directly using this chemistry, generating good quality sequence data and thus avoiding time consuming purification steps.

To achieve the best quality template possible a nested PCR method was used. This involves two successive rounds of PCR, the second round primers falling within the first round PCR product, thus increasing the specificity and concentration of the second round

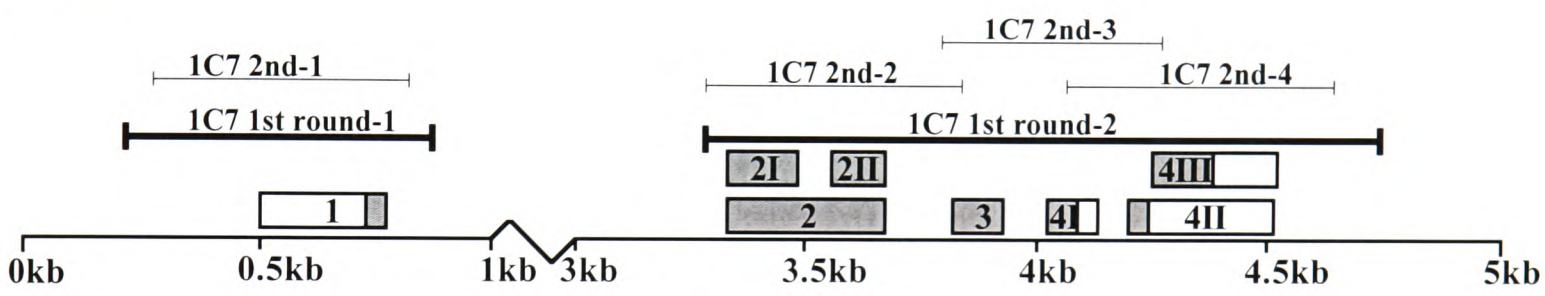
PCR product. As described in section 2.3.9.5, the first round PCR product was diluted before use in the second round amplification reaction in order to minimise interference of first round primers in the second round reaction. Also, the concentrations of second round primers and dNTPs were kept to a minimum so that PCR reagents were not carried over into the sequencing reaction. In addition to this, primers for the second round of PCR were M13-tailed so that all PCR products could be sequenced using the standard M13-21 or M13-reverse primer, thus aiding automation of the process. Details of the PCR products generated are represented in Figure 5.1 (except for the BAT1 promoter region which involved a single first round and second round reaction and thus did not warrant a diagram) and in Table 5.1. The sequences of the primers listed in Table 5.1 are recorded in Appendix IIIA. A total of 2327 nucleotides was screened from the AIF-1 genomic region, as well as 2349 nucleotides from 1C7, 2207 nucleotides from ATP6G and 559 nucleotides from BAT1.

The investigation described in this section represents a pilot screen on a minimal number of genomic templates from a control cohort (i.e. a population of healthy individuals) the aim being to identify polymorphisms within the general population. The most commonly used definition of a polymorphism is that its most common allele should have a population frequency of less than 95% (Ott 1991). This means that a rare allele, with a frequency of 5%, can be detected by sequencing a minimum of 20 chromosomes. Therefore, genomic DNA derived from 12 HLA heterozygous healthy controls (kindly donated by K. I. Welsh, Nuffield Department of Surgery, Churchill Hospital, Oxford) was selected. This equates to a total of 24 chromosomes and thus meets the minimal requirements outlined by Ott (1991). The genotypes of the control DNAs screened are given in Table 5.2.

A. AIF-1



B. 1C7



C. ATP6G

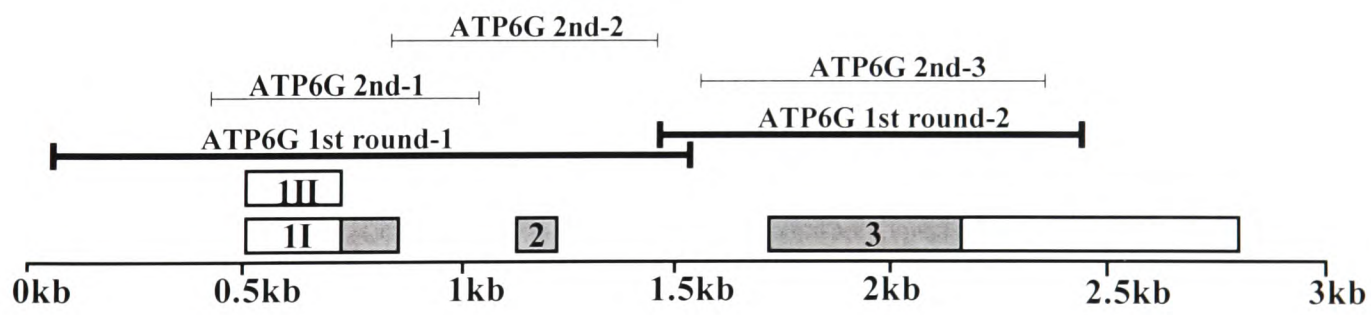


FIGURE 5.1 The location of PCR products used for mutation analysis across the genomic sequences of AIF-1, 1C7 and ATP6G.
The positions of the exons within the three genomic regions are shown as boxes with the coding regions shaded. The extents of the first round PCR products are indicated with solid bold lines and the second round PCR products are depicted by fine lines. The PCR primer names and product sizes are given in Tables 5.1, the PCR primer sequences are recorded in Appendix IIIA and the PCR conditions are described in section 2.3.9.5.

TABLE 5.1 PCR templates generated for mutation analysis of the AIF-1, 1C7 and ATP6G genomic regions and the BAT1 promoter region.

Gene	1st round Primers[ae]	PCR product[b]	2nd round Primers[ce]	M13 tailed PCR product[d]	2nd round product coordinates[f]
AIF-1	AIF388f-AIF1605r	AIF-1 1st round-1 (1217bp)	AIFm13f1-AIFm13r1	AIF-1 2nd-1(552bp)	2463-3015
			AIFm13f2-AIFm13r2	AIF-1 2nd-2 (593bp)	2876-3469
	AIF1201f2-AIF2715r	AIF-1 1st round-2 (1514bp)	AIFm13f3-AIFm13r3	AIF-1 2nd-3 (509bp)	3345-3854
			AIFm13f4-AIFm13r4	AIF-1 2nd-4 (498bp)	3710-4208
			AIFm13f5-AIFm13r5	AIF-1 2nd-5 (547bp)	4052-4599
1C7	1C7211f1-1C7933r1	1C7 1st round-1 (781bp)	1C7m13f1-1C7m13r1	1C7 2nd-1 (635bp)	26413-27048
	1C7359f2-1C71926r2	1C7 1st round-2 (1568bp)	1C7m13f2-1C7m13r2	1C7 2nd-2 (609bp)	29381-29990
			1C7m13f3-1C7m13r3	1C7 2nd-3 (644bp)	29820-30464
			1C7m13f4-1C7m13r4	1C7 2nd-4 (610bp)	30231-30841
ATP6G	ATP130f-ATP1472r2	ATP6G 1st round-1 (1342bp)	ATPm13f1-ATPm13r1	ATP6G 2nd-1 (615bp)	72641-73256
			ATPm13f2-ATPm13r2	ATP6G 2nd-2 (609bp)	73084-73693
	ATP1374f2-ATP2337r	ATP6G 1st round-2 (963bp)	ATPm13f3-ATPm13r3	ATP6G 2nd-3 (639bp)	73798-74437
BAT1	BAT1-249f-BAT1-768r	BAT1 1st round (559bp)	BAT1m13f1-BAT1m13r1	BAT1-2nd (463bp)	78744-79207

[a] Primers used in the first round PCR reactions to amplify from genomic DNA.
[b] Product generated by first round PCR and used as template in second round.
[c] Primers used to generate M13-tailed products in the second round PCR reactions.
[d] M13-tailed sequencing templates generated by second round PCR.
[e] Primer sequences can be found in Appendix IIIA.
[f] Coordinates relate to the TN62/TN82 sequence generated in this study (Appendix I).

TABLE 5.2 The HLA haplotypes for the 12 heterozygous control DNAs used in the mutation analysis and the homozygous database entries that overlap with the genomic regions analysed.

DNA [a]	HLA-DQB1	HLA-DRB1	HLA-B	HLA-C	HLA-A
B491	*02	*15	*0702/4/5/6	*0702	*03
	*0602	*0701	*1401	*0802	*11
B492/E398	*05	*0101/2/4	*0801	*0701	*01
	*0302	*04	*18	*0701	*02
B493/E389	*05	*1001	*0702/4/5/6	*0702	*03
	*05		*5201	*1202	*03
B494	*02	*0301/4	*0801	*0701	*01
	*0301/4	*1101/4/6/10/12	*4002/4/6	*0202	*24
B495/E382	*04	*0701	*35	*04	*02
	*0303	*08	*5701	*0602	*24
B638	*05	*0101	*4402	*0102	*0201/4/6
	*0301/4	*0401/9/13/14/6	*5601	*0304	*0201/4/6
B952	*0401/2	*0801/5-8/11/16	*1801-5	*0602	*0301
	*0301/4	*1101/4/6/10/12	*3701	*0701	*24
B958	*05	*0101	*4402	*0501	*0101
	*0301/4	*0401/9/13/14/16	*5701	*0602	*0201/4/6
C016	*05	*1501	*0801	*0701	*0201/4/6
	*0602/10/11	*1404	*5201	*1202	*2901
C022	*0201/2	*1501	*4403	*0701	*0201/4/6
	*0602/10/11	*0701/3	*5801	*1601	*1101/3/4
D301	*02	*0101	*0801	*0701	*0101
	*05	*0301/4	*0801	*0701	*0101
E393	*0201/2	*0301/4	*0801	*0501	*0101
	*0302	*0401/9/13/14/16	*4403	*0701	*02
HSY14768	-	*02	*07	-	*02
U00921	-	*07	-	-	*03
AB000876	-	*04	*62	-	*02
Z15025	-	*04	-	-	-
Z15027	-	*04	-	-	-
AC004181	-	-	-	-	-
AF129756	-	-	-	-	-

- = not determined

[a]: The DNAs B491 to E393 are those donated by K. I. Welsh for mutation analysis. The DNA HSY14768 is the database entry generated in this Thesis from cosmids TN62 and TN82 and the DNAs U00921, AB000876, Z15025 Z15027, AC004181 and AF129756 are the EMBL database entries that overlap with the TN62/TN82 genomic sequence. There is no genotype data available for database entries AC004181 and AF129759.

5.2.2 Results

A total of 228 M13 tailed PCR templates were sequenced to screen the chosen regions for polymorphisms: 98 templates for AIF-1, 83 for 1C7, 56 for ATP6G and 16 for the BAT1 promoter region. Initially all templates were sequenced in the forward orientation with the M13-21 sequencing primer. Following this, the second round PCR products AIF-1 2nd-1, AIF-1 2nd-2, AIF-1 2nd-5, 1C7 2nd-2, 1C7 2nd-3 and ATP6G 2nd-2 (see Figure 5.1) had to be sequenced in the reverse orientation to resolve sequencing problems such as stops and compressions. Heterozygous mutations were easily detected on the resulting sequence trace files as two overlapping peaks of similar height. Figure 5.2 shows a selection of the different transition and transversion polymorphisms detected in this study. In all cases, where a possible polymorphism was detected the nested PCR and sequencing reactions were repeated to ensure that sequence variations were not due to errors introduced during PCR. No such artifacts were observed. To simplify the detection of sequence variation, all trace files were processed by the 'PREGAP' Staden suite of software then entered into the GAP v4.0-beta4 sequence alignment software (section 2.3.13.2). Within the GAP4 graphical interface all the trace files could be viewed in alignment with each other and heterozygous mutations, indicated by an 'N' on the text line of the trace file (see Figure 5.2), were detected by the GAP4 software as a discrepancy in the alignment. Corresponding trace files from the original cosmid sequencing project were also entered into the alignment to act as a consensus sequence so that any homozygous mutations could be detected.

A total of 28 polymorphisms were detected over the four genomic regions examined, which represented a total of 7442 nucleotides of genomic sequence. This corresponds to an average of 3.8 polymorphisms per kilobase. In addition to this, 17 further differences

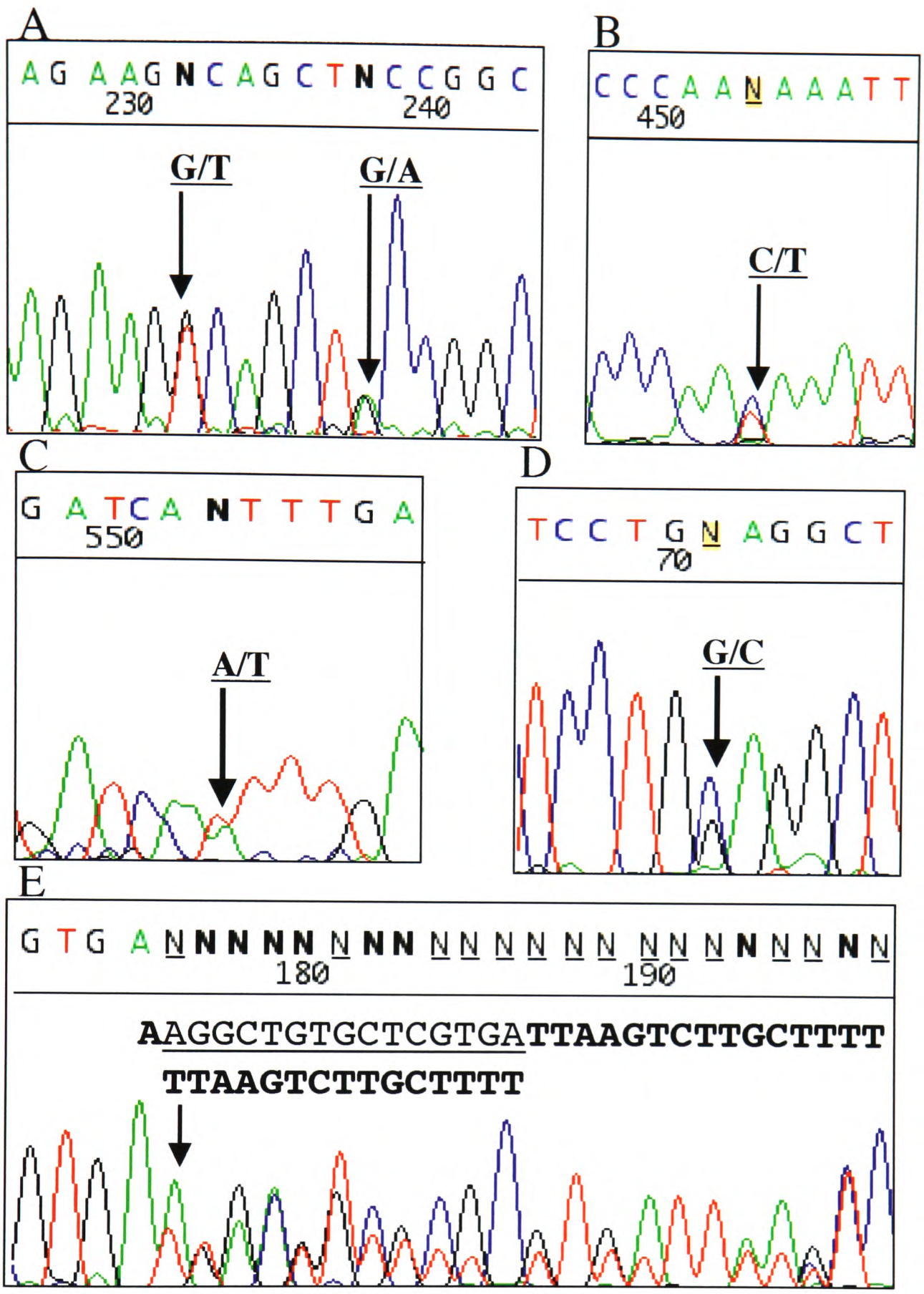


FIGURE 5.2 Examples of polymorphisms identified by sequence analysis.

Heterozygous mutations were identified on ABI sequence trace files as two overlapping peaks. Polymorphisms are marked by arrows with the base change indicated above. Examples of polymorphisms identified in this study include the G/T transversion and the G/A transition at SNPs 1C7*3571 and 1C7*3577, respectively (**A**), the C/T transition at SNP 1C7*3790 (**B**), the A/T transversion at SNP 1C7*3843 (**C**) and the G/C transversion at SNP 1C7-412 (**D**) (all found within the 1C7 genomic region).

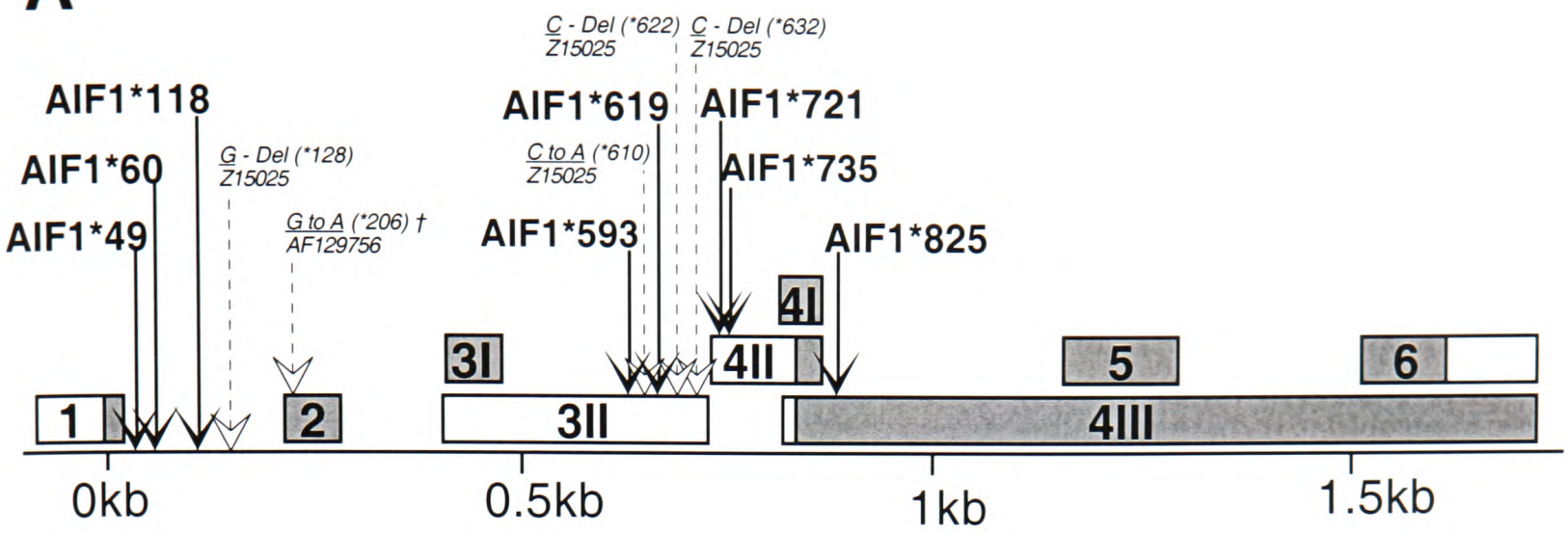
The 15 nucleotide deletion found within the BAT-1 promoter region (BAT1-180/-165) is clearly identifiable as the overlap of two different sequences at the point of the deletion, indicated with an arrow (**E**). In box **E**, the sequence common to both chromosomes is written in bold and the sequence deleted in one of the chromosomes is underlined. The two overlapping traces can clearly be seen after the deletion point .

were identified between genomic sequences deposited in the EMBL database and the four regions investigated here. It is not possible to state whether these are haplotype specific variations or errors in the sequence data as they were not confirmed experimentally in this study.

5.2.2.1 Polymorphisms within the AIF-1 genomic region

The results of the mutation analysis within the AIF-1 genomic region are presented in Figure 5.3. 8 single nucleotide polymorphisms (SNPs) were identified within the 2327 nucleotides screened from the 12 control DNAs and these were all confirmed by repeated analysis. This corresponds to an average of 3.4 SNPs per kilobase of DNA. Of these, three mutations are intronic (AIF1*49, AIF1*60 and AIF1*118), two fall within the IRT-1 transcript 5' untranslated region (AIF1*593 and AIF1*619) and two lie within the G1 transcript 5' untranslated region (AIF1*712 and AIF1*735). Only AIF1*825 lies within a coding region. This C to T transition falls within the coding region of the IRT-1 splice variant of the AIF-1 gene in exon 4III. This polymorphism was found in three out of the 24 chromosomes tested and results in an amino acid change from arginine to tryptophan at residue 15 of IRT-1. This was identified as a possible di-basic endoprotease cleavage site in the IRT-1 protein (see section 4.3.1 and Figure 4.3). Figure 5.3 also shows a further 5 differences between the TN62/TN82 sequence and other genomic EMBL database entries covering the AIF-1 gene. Of these, a G to A transition at position 206 (i.e. 206 nucleotides from the start codon) falls within the coding region of the AIF-1 transcript (in exon 2) and would result in an amino acid change from a glycine to an arginine at residue 14 of the AIF-1 protein.

A



B

Polymorphism name	Base change	Frequency of rare allele in control chromosomes	Control DNAs and database entries in which the rare allele is found #	Comment
AIF1*49	G to A	12 A in 24 (50%)	B492, B494, B495(Hom), B638(Hom), B952(Hom), B958(Hom), C016, E393, *Z15025(Hom)*.	Intronic
AIF1*60	G to A	1 A in 24 (4.2%)	C022	Intronic
AIF1*118	G to A	1 A in 24 (4.2%)	B495	Intronic
AIF1*593	C to T	1 T in 24 (4.2%)	C022	In the 5' untranslated region of the IRT-1 transcript.
AIF1*619	C to T	1 T in 24 (4.2%)	C022	In the 5' untranslated region of the IRT-1 transcript.
AIF1*721	T to G	10 G in 24 (42%)	B492, B494, B495(Hom), B638(Hom), B958(Hom), C016, E393,*Z15025(Hom)*.	In the 5' untranslated region of the G1 transcript
AIF1*735	C to G	2 G in 24 (8.4%)	B638, B958	In the 5' untranslated region of the G1 transcript.
AIF1*825	C to T	3 T in 24 (12.5%)	B492, C022, D301	An amino acid change within the IRT-1 protein only: Arg-15-Trp.

FIGURE 5.3 Polymorphisms, and database differences, identified within a 2.3kb segment of genomic DNA covering the AIF-1 region.

A. The genomic structure of AIF-1. The boxes represent exons with the shaded boxes corresponding to the coding regions and unshaded boxes to the untranslated regions. The positions of polymorphisms identified experimentally by mutation analysis are indicated by bold text and solid arrows. The gene name is indicated directly before the asterisk'*' and the genomic location of the polymorphism (in relation to the start codon of the AIF-1 transcript in exon 1) is indicated directly after the asterisk'*'. Differences identified by comparison with genomic sequences deposited within the databases, but not confirmed experimentally here, are indicated by italic text and broken arrows. The nature of the differences is underlined and its position, in relation to the start codon, is indicated in brackets.

† - This transition seen in database entry AF129756, if real, causes an amino acid change, Gly - 14 - Arg, within a highly conserved region of the protein.

B. Additional information about the polymorphisms identified by mutation analysis.

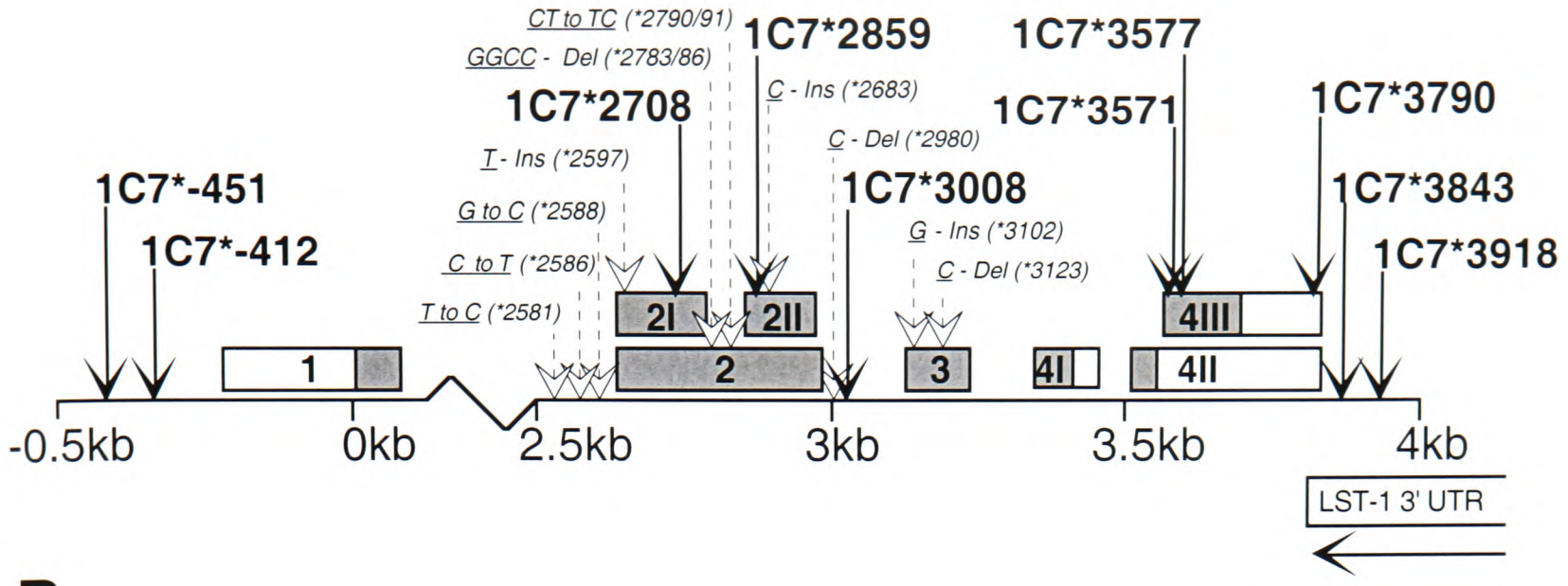
- (Hom) = Homozygous

5.2.2.2 Polymorphisms within the 1C7 genomic region

The results of the mutation analysis within the 1C7 genomic region are presented in Figure 5.4. Intron 1 was not included in this analysis due to its large size. Within the 2349 nucleotides that were screened from the 12 control DNAs, 9 SNPs were identified and confirmed by repeating the analysis i.e. 3.8 SNPs per kilobase of DNA on average. Of these, 1C7*-451 and 1C7*-412 fall within the promoter region of 1C7, 1C7*3008 falls within intron 2 and 1C7*2708 and 1C7*3577 fall within the coding region of 1C7 but do not cause amino acid changes. The latter of the two silent mutations represents a previously published *Pvu* II polymorphism (de Baey *et al.* 1995). SNP 1C7*3571 on the other hand, a G to T transversion in exons 4II and 4III, falls within the coding region of transcripts 1C7a and 1C7e and causes an amino acid change at residue 174 of 1C7a from arginine to serine. This position is within a candidate SH3 domain binding motif (see section 4.3.2 and Figure 4.7). SNP 1C7*3790 although not falling within a coding region, disrupts the polyadenylation signal of 1C7a, 1C7b and 1C7e, changing it from 'AATAAA' to AACAAA. In addition, SNP 1C7*3843 falls within the polyadenylation signal of the LST-1 gene changing it from AATAAA to AATTAA. The LST-1 gene overlaps with the 3' untranslated region of 1C7, being transcribed in the opposite direction.

Also included within Figure 5.4 is SNP 1C7*2859 which, whilst not identified in this study, was identified in a previous study using similar methods (R. E. March, personal communication). It was identified in IDDM patient DNAs on a DR4/DR1 background with 3 out of 32 DNAs being positive for the rare allele (i.e. 9.3%). 1C7*2859 causes an amino acid change from alanine to serine at residue 103 in 1C7a, b and c and residue 78 in 1C7d, e and f. This amino acid lies within the immunoglobulin domain of 1C7 between β strands E and F. In addition to the above SNPs, a further 10 differences were identified

A



B

Polymorphism name	Base change	Frequency of rare allele in control chromosomes	Control DNAs and database entries in which the rare allele is found #	Comment
1C7*-451	G to A	1 A in 24 (4.2%)	B638	1C7 promoter region.
1C7*-412	G to C	4 C in 24 (17%)	B493, B952(Hom), C016, *AF129756*	1C7 promoter region.
1C7*2708	C to T	1 T in 24 (4.2%)	B494	Within coding region but silent.
1C7*2859	G to T	Not determined	Identified by R. E. March (Personal communication)	Amino acid change in loop between ß strands E & F in Ig domain: 1C7a, b, c: Ala-103-Ser, 1C7d, e, f: Ala-78-Ser.
1C7*3008	G to A	2 A in 24 (8.4%)	B491, C022	Intronic.
1C7*3571	G to T	2 T in 24 (8.4%)	B491, C022	Amino acid change within candidate SH3 domain binding motif: 1C7a Arg-174-Ser.
1C7*3577	G to A	2 A in 24 (8.4%)	B491, C022	Silent. A published *Pvu* II polymorphism (de Baey, *et al.* 1995)
1C7*3790	T to C	2 C in 24 (8.4%)	B492, B638, *U00921*	Disrupts the Poly-A signal of 1C7 a, b, and e, changing it from AATAAA to AACAAA
1C7*3843	T to A	2 A in 24 (8.4%)	B495, B958	Falls within the poly-A signal of LST1, causing a change from AATAAA to AATTAA.
1C7*3918	A to T	3 T in 24 (12.5%)	B491, B494, B495, C022	LST1 3' untranslated region

FIGURE 5.4 Polymorphisms and database differences identified within a 2.3kb segment of genomic DNA covering the 1C7 region.

A. The genomic structure of 1C7. The boxes represent exons with shaded boxes corresponding to coding regions and unshaded boxes to untranslated regions. The positions of polymorphisms identified experimentally by mutation analysis are indicated by bold text and solid arrows. The gene name is indicated directly before the asterisk'*' and the genomic location of the polymorphism (in relation to the start codon) is indicated directly after the asterisk'*'. Differences identified with the database entry U00921, but not confirmed experimentally in this study, are shown in italic text with broken arrows, the nature of the difference is underlined and its position (in relation to the start codon) is indicated in brackets. The 3' UTR of the LST-1 gene is indicated below the scale with the direction of transcription shown with an arrow.

B. Additional information about the polymorphisms identified by mutation analysis.

- (Hom) = Homozygous

between the TN62/TN82 sequence and the EMBL database entry U00921 (Figure 5.4). None of these were confirmed experimentally.

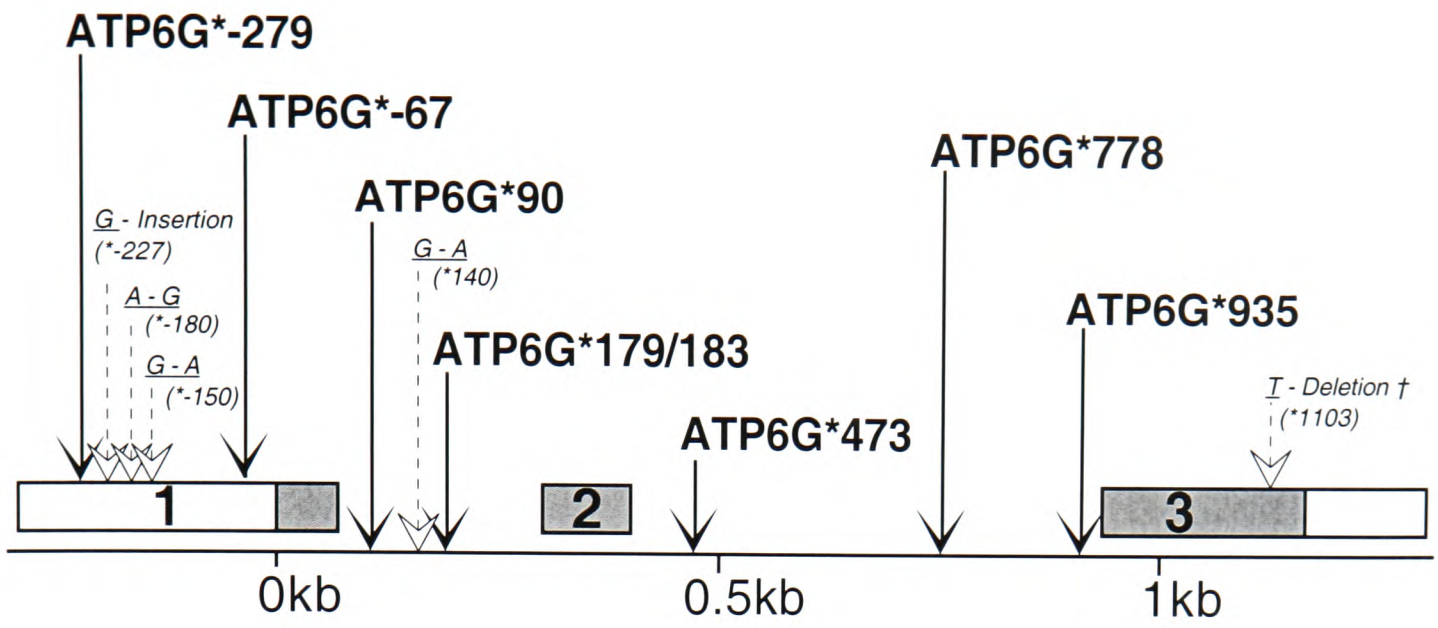
5.2.2.3 Polymorphisms within the ATP6G genomic region

The results of the mutation analysis within the ATP6G genomic region are presented in Figure 5.5. With the 2207 nucleotides screened, 6 SNPs and 1 deletion were identified and these were confirmed by repeating the analysis. This corresponds to an average of 3.2 polymorphisms per kilobase of DNA. None of the polymorphisms detected lie within the ATP6G coding region. ATP6G*-279 and ATP6G*-67 lie in the 5' untranslated region, whilst the other mutations, including the 4 nucleotide deletion (ATP6G*179/183) all occur in intronic sequence. In addition to the above SNPs, a further 5 differences were identified between the TN62/TN82 sequence and the EMBL database entry Z15027 (Figure 5.5). Among these, a T deletion in exon 3, if real, causes a frame shift in the open reading frame of ATP6G resulting in a change in the last 16 amino acids and an extension of the polypeptide for an additional 27 residues until a stop codon is encountered.

5.2.2.4 Polymorphisms within the BAT1 promoter region

Screening the 559 nucleotides 5' of the BAT1 exon 1 sequence in the 12 control DNAs confirmed the existence of the 15 nucleotide deletion first identified by comparison of the sequence generated in this study (HSY14768) with EMBL database entry AC004181. One of the 12 DNAs tested (C022) was heterozygous for this deletion, i.e. it was identified in 1 out of 24 chromosomes. In addition to the deletion, two SNPs were also identified in this genomic region (Figure 5.6): a G/C transversion and a T/C transition.

A



B

Polymorphism name	Base change	Frequency of rare allele in control chromosomes	Control DNAs and database entries in which the rare allele is found #	Comment
ATP6G*-279	T to C	1 C in 24 (4.2%)	B491	5' untranslated region
ATP6G*-67	C to T	1 T in 24 (4.2%)	C022	5' untranslated region
ATP6G*90	T to C	5 C in 24 (21%)	B492, B493, D301(Hom) E393	Intronic
ATP6G*179 /183	ACAG insertion	3 in 24 (12.5%)	B491, B493, C016, *Y14768(Hom)*	Intronic
ATP6G*473	A to T	3 T in 24 (12.5%)	B491, B493, C016, *Y14768(Hom)*	Intronic
ATP6G*778	G to T	1 T in 24 (4.2%)	B494, *Z15027(Hom)*, *AB00876(Hom)*	Intronic
ATP6G*935	C to T	3 T in 24 (12.5%)	B492, B638, C022, *AC004181(Hom)*	Intronic

FIGURE 5.5 Polymorphisms and database differences identified within a 2.2kb segment of genomic DNA covering the ATP6G region.

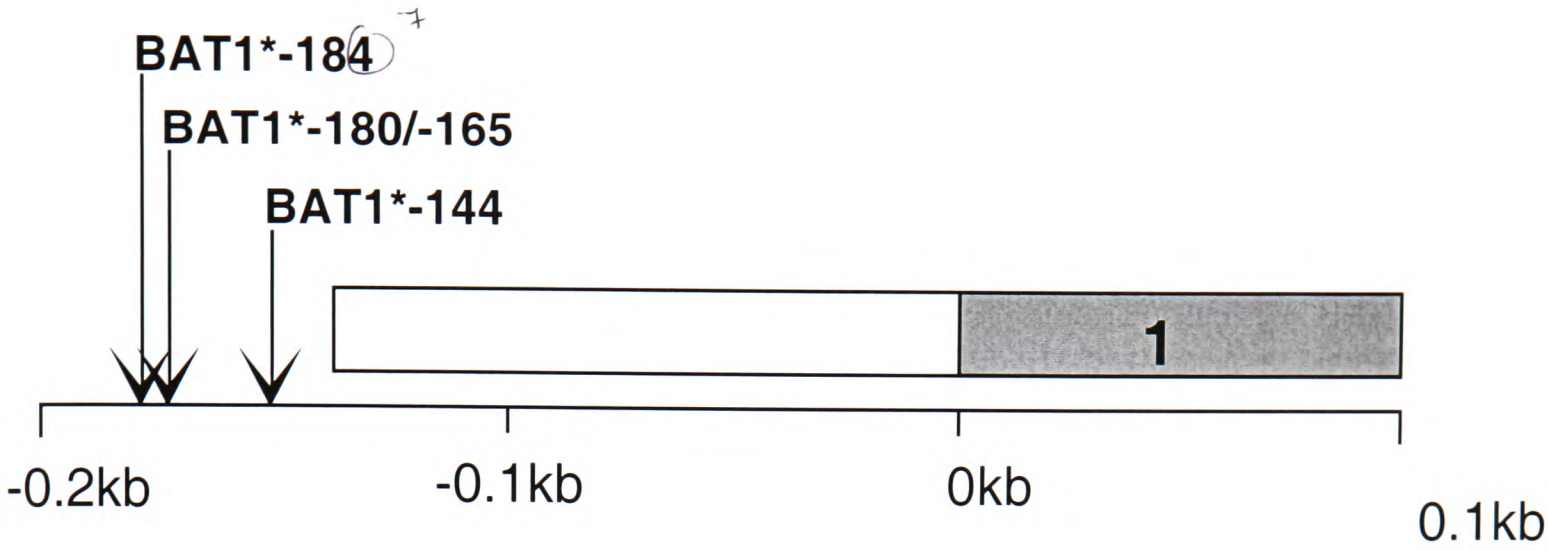
A. The genomic structure of ATP6G. The boxes represent exons with the shaded boxes corresponding to coding regions and unshaded boxes to untranslated regions. The positions of polymorphisms identified experimentally by mutation analysis are indicated by bold text and solid arrows. The gene name is indicated directly before the asterisk'*' and the genomic location of the polymorphism (in relation to the start codon) is indicated directly after the asterisk'*'. Differences identified with the database entry Z15027, but not confirmed experimentally in this investigation, are indicated by italic text and broken arrows. The nature of the difference is underlined and its position, in relation to the start codon, is indicated in brackets.

† - This deletion seen in database entry Z15027, if real, causes a frame shift. This results in a change in the last 16 amino acids of ATP6G and an extension of the polypeptide for an additional 27 residues until a stop codon is encountered.

B. Additional information about the polymorphisms identified by mutation analysis.

- (Hom) = Homozygous

A



B

Polymorphism name	Base change	Frequency of rare allele in control chromosomes	Control DNAs and database entries in which the rare allele is found #	Comment
BAT1*-187	G to C	1 C in 24 (4.2%)	C022, *AC004181(Hom)*	BAT1 promoter region.
BAT1*-180 /-165	AGGCTGT-GCTCGTGA Deletion	1 in 24 (4.2%)	C022, *AC004181(Hom)*	BAT1 promoter region.
BAT1*-144	T to C	1 C in 24 (4.2%)	B494, *HSY14768*	BAT1 promoter region.

FIGURE 5.6 Polymorphisms and database differences identified within a 0.55kb segment of genomic DNA covering the BAT1promoter region.
A. The genomic location of polymorphisms identified within the BAT1 promoter region. The boxed region represents the 5' end of BAT1 exon 1, the shaded box representing the coding region and the unshaded box the 5' untranslated region. The positions of polymorphisms identified experimentally by mutation analysis are indicated by bold text and solid arrows. The gene name is indicated directly before the asterisk'*' and the genomic location of the polymorphism in relation to the start codon is indicated directly after the asterisk'*'.
B. Additional information about the polymorphisms identified by mutation analysis.
- (Hom) = Homozygous

5.3 THE SCREENING OF POLYMORPHISMS IDENTIFIED IN THIS STUDY FOR ASSOCIATIONS WITH COMMON VARIABLE IMMUNODEFICIENCY AND IgA DEFICIENCY.

Common variable immunodeficiency (CVID) and IgA deficiency (IgAD) are the two most common primary immunodeficiency conditions, characterised by low serum immunoglobulin levels and increased susceptibility to recurrent infections (Mullighan *et al.* 1999). Although there are clinical differences between the two conditions it has been suggested that they represent opposite poles of an immunodeficiency spectrum (Schroeder *et al.* 1998) and that they share common genetic defects. As a result many association studies have combined the two patient groups (Schaffer *et al.* 1989; Volanakis *et al.* 1992; Schroeder *et al.* 1998). The causative agents, however, are still unknown as the identification of associated alleles has proven extremely difficult. This is partly due to the highly variable age of onset of disease and to the diverse clinical manifestations of CVID and IgAD. For example, whilst immunodeficiency is present in all patients, complications such as lymphoproliferative disease, autoimmune disease and granulomatous disease are frequent, but not universal (Mullighan *et al.* 1999). This suggests that multiple loci may contribute to the clinical phenotypes observed.

Genetic studies investigating CVID/IgAD have all suggested the presence of susceptibility loci within the human MHC region. Most patients carry either complete copies of or extended fragments of two MHC haplotypes, designated haplotype 1 (HLA-DQB1 0201 – HLA-DR3 – C4B-Sf – C4A-0 – G11-15, Bf-0.4 – C2-a – HSP-7.5 – TNF-5 – HLA-B8 – HLA-A1) and haplotype 2 (HLA-DQB1 0201 – HLA-DR7 – C4B-S – C4A-L – G11-4.5, Bf-0.6 – C2-b – HSP-9 – TNF-9 – HLA-B44 – HLA-A29) (Volanakis *et al.* 1992). In addition to this, the haplotype DR15-DQ6 appears to be protective (Olerup *et al.* 1990). One possible susceptibility factor is the polymorphic codon 57 of DQB1 and it has been

suggested that an alanine or valine substitution at this position conveys susceptibility to CVID/IgAD, whilst an aspartic acid residue at this position is protective (Olerup *et al.* 1990). However, this has not been confirmed in later studies (Mullighan *et al.* 1999). The recent study by Mullighan *et al.* (1999) identified a significant association between CVID and an allele of IL-1α, the first association found outside the MHC region. Furthermore, after segregation of the clinical phenotypes, associations were detected between alleles of the vitamin D receptor and IL-6 and the abnormalities characteristic of more severe disease as well as between TNF and IL-10 alleles (in an interactive manner) and the granulomatous form of CVID.

As mentioned in section 5.1 (above) and in Chapter 1 of this Thesis, there is strong evidence that a CVID susceptibility gene is present within the AIF-1-HLA-A portion of haplotype 1 (Schroeder *et al.* 1998). This segment of DNA includes the TN62/TN82 region analysed here. The recent findings of Mullighan *et al.* (1999) suggest an association with the TNF +488A allele (D'Alfonso 1996) (as mentioned above) which may explain, in part, the AIF-1-HLA-A association. However, the TNF association was detected after segregation of patients on the basis of disease phenotype. No significant association was detected when the CVID cohort was examined as a whole. Therefore, there may well be additional susceptibility loci in the AIF-1-HLA-A region. It will, therefore, be of great interest to investigate the polymorphisms described in section 5.2 in relation to CVID/IgAD. However, the polymorphisms described here were identified using a sample of only 24 chromosomes. Meaningful statistical analysis will require the screening of a much larger number of chromosomes, from both healthy (control) and affected individuals. To this end a collaboration has been set up with Dr Ken Welsh (Nuffield Department of Surgery, Churchill Hospital, Oxford). In the study reported above, Mullighan and co-

workers from Ken Welsh's group collected cohorts of 163 CVID patients and 101 Caucasoid controls. A PCR-SSP technique, capable of genotyping multiple polymorphisms under identical conditions was used to screen 23 polymorphisms in 13 immunoregulatory genes from these individuals (Mullighan *et al.* 1999). These resources were made available to this investigation.

5.3.1 Resources and Methodology

5.3.1.1 DNA resources

The DNAs used were those described by Mullighan *et al.* (1999). These include a total of 163 Caucasoid CVID patients, recruited from English and Welsh specialist Clinical Immunology centres, together with 101 kidney donors as controls. The controls represent cadaveric organ donors from around the United Kingdom genotyped at the Oxford Tissue Typing Laboratory (Churchill Hospital, Oxford) for transplantation purposes. This population has been extensively characterised and utilised for a variety of HLA studies and has the expected HLA allelic frequencies for UK Caucasoids (Bunce *et al.* 1995; Bunce *et al.* 1996).

5.3.1.2 Polymorphisms screened

1C7 was considered a good candidate susceptibility gene for CVID/IgAD and initial screening has primarily focused on this gene. All 9 of the 1C7 SNPs identified here (section 5.2.2.2) together with an SNP identified by Dr R.E. March were analysed. In addition to this, the AIF-1 polymorphism AIF1*825 and the difference with EMBL database entry AC004181 at AIF-1 position 206 (section 5.2.2.1), now given the name AIF1*206, together with the BAT1 deletion BAT1*-180/-165 were considered to be of possible importance (see section 5.4) and were included in the study. Although SNP

AIF1*206 had not been verified experimentally (section 5.2.2.1) it was hoped that analysis of a larger number of samples would confirm its existence.

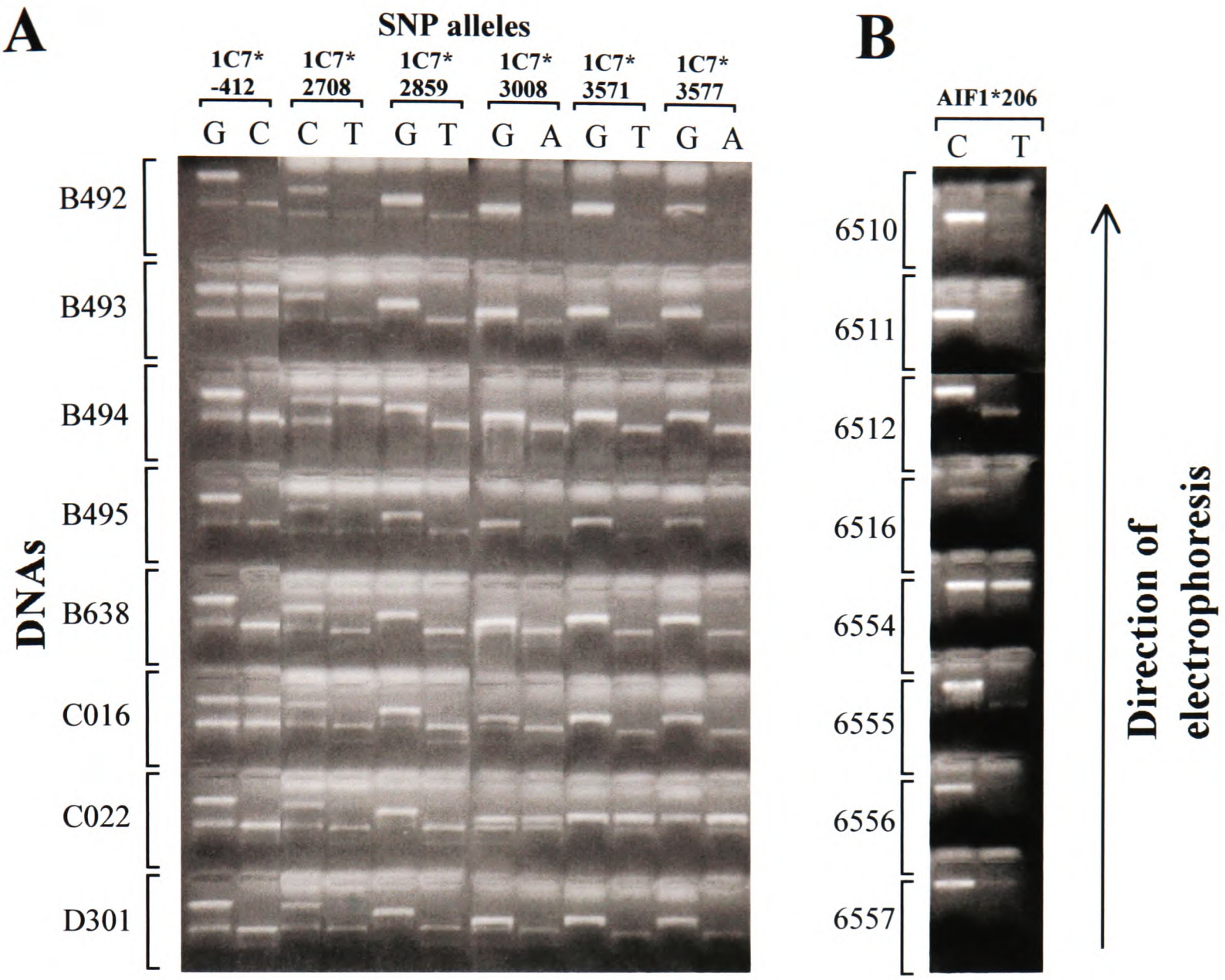
5.3.1.3 Methodology

A unified, sequence-specific primer, polymerase chain reaction (PCR-SSP) method to simultaneously genotype multiple polymorphisms under identical conditions has been developed at the Tissue Typing Centre, Nuffield Department of Surgery, Churchill Hospital, Oxford (Bunce *et al.* 1995; Bunce *et al.* 1996; Mullighan *et al.* 1999). This method is described in section 2.3.15 and is currently being used to genotype the polymorphisms selected in section 5.3.1.2. PCR-SSP is an allele specific PCR assay in which two pairs of primers are included in each reaction. One primer pair amplifies a control DNA fragment to test for PCR failure, while the second pair of primers amplifies an allele specific product from the gene of interest. For a dimorphic locus, two PCR reactions are required for each DNA sample genotyped, one for each set of allele specific primers. Using this method, homozygous and heterozygous genotypes can be determined with confidence, since a control band of 796 bp, amplified from an invariant region of the human DRB1 gene, will be present for all successful reactions, regardless of genotype. With the methodology adopted, all PCR buffer and cycling conditions remain identical. Therefore, a large number of polymorphisms can be successfully and reliably genotyped simultaneously, substantially reducing the time needed to genotype a large number of polymorphisms against a significant cohort of DNAs compared to previous methods. This approach has been used very successfully for HLA class I and II genotyping as well as non-HLA disease association studies (Bunce *et al.* 1995; Mullighan *et al.* 1999).

5.3.2 Results

The assay development stage of this investigation is still in progress. Initial primers (see Appendix IIIB) were designed based on criteria set out in section 2.3.15 and assays (Table 5.3) were tested using the 12 control DNAs described in section 5.2 to identify those conditions that worked reliably and those that needed further modification of the primer concentration or sequence. At present 7 out of the 12 assays are working (indicated in Table 5.3 with a ✓ in the far right column) and modification to the 5 remaining assays is in progress. An example of the results obtained with the 6 successful 1C7 assays, against 8 out of the 12 control DNAs (section 5.2) is shown in Figure 5.7A and an example of the results obtained with the single working AIF-1 assay, against a selection of CVID patient DNAs is shown in Figure 5.7B.

The screening of the 163 CVID DNAs and the 100 control DNAs for the 7 SNPs for which assays have been established is currently in progress. From the analysis carried out to date, the AIF1*206 putative SNP has been confirmed (Figure 5.7B) but appears to be rare, at least in the CVID DNAs (a frequency of about 1.5% in 320 CVID chromosomes). Also, in relation to the G/T mutation identified by R.E. March (personal communication) in 1C7 (SNP 1C7*2859) only the G allele has been seen in the CVID cohort. However, 1C7*2859 has not yet been genotyped in the control DNAs. It should be noted though, that until a chromosome with the T allele is identified, there is always a possibility that the variant allele assay does not work, as no DNA known to carry the T allele was available to test the initial assay for false negative results. This is also an issue with regard to assaying for individuals homozygous for rare alleles. None of the test DNAs used for any of the SNP assays so far working were homozygous for the rare allele, therefore, it cannot be said



C

	1C7* -412	1C7* 2708	1C7* 2859	1C7* 3008	1C7* 3571	1C7* 3577		AIF1* 206
B492	G	C	G	G	G	G	6510	C
B493	G/C	C	G	G	G	G	6511	C
B494	G	C/T	G	G	G	G	6512	C
B495	G	C	G	G	G	G	6516	C
B638	G	C	G	G	G	G	6554	C/T
C016	G/C	C	G	G	G	G	6555	C
C022	G	C	G	G/A	G/T	G/A	6556	C
D301	G	C	G	G	G	G	6557	C

FIGURE 5.7 PCR-SSP results for 1C7 and AIF-1.

A & B. Test gels. Gel lanes are grouped in pairs, representing the two possible alleles of each SNP. The first lane of each pair is the common allele and the second is the rare allele. The nucleotide present in each allele is written above the lanes. The DNAs screened for the six 1C7 SNPs (A) and single AIF-1 SNP (B) are indicated on the left of each panel. In each gel lane, the lower band, when visible, is the 796 bp DRB1 control band, whereas the upper band, when present, is the allele specific band.

C. Allele scoring table. A tabular representation of the results from Figure 5.7A.

TABLE 5.3 PCR-SSP genotyping assay primer mixes

Loci [a]	Allele primer No. [b]	Consensus primer No. [b]	Product size	Present status [c]
1C7*-421(G)	2796	2795	187	✓
1C7*-421(C)	2797	2795	187	✓
1C7*2708(C)	2873	2872	215	✓
1C7*2708(T)	2874	2872	215	✓
1C7*2859(G)	2875	2872	366	✓
1C7*2859(T)	2876	2872	366	✓
1C7*3008(G)	2877	2882	522	✓
1C7*3008(A)	2878	2882	522	✓
1C7*3571(G)	2883	2881	417	✓
1C7*3571(T)	2884	2881	417	✓
1C7*3577(G)	2885	2881	422	✓
1C7*3577(A)	2886	2881	422	✓
1C7*3790(T)	2887	2881	641	✗
1C7*3790(C)	2888	2881	641	✗
1C7*3843(T)	2889	2895	367	✗
1C7*3843(A)	2890	2895	367	✗
1C7*3918(A)	2893	2895	293	✗
1C7*3918(T)	2894	2895	293	✗
BAT1*-180/-165 (ins)	2935	2934	233	✗
BAT1*-180/-165 (del)	2936	2934	221	✗
AIF1*825 (C)	2938	2937	269	✗
AIF1*825 (T)	2939	2937	269	✗
AIF1*206 (G)	2941	2940	237	✓
AIF1*206 (A)	2942	2940	237	✓

[a]: Loci refer to polymorphisms identified in section 5.2. The allele specific base is shown in brackets except for BAT1*-180/165 where (ins) represents the allele with the 15 base insertion and (del) represents the allele with the 15 base deletion.

[b]: Primer numbers refer to primers listed in Appendix IIIB.

[c]: ✓ signifies that the assay works satisfactory with the primer pair shown and ✗ signifies that the assay failed with that primer pair.

Note. The optimal final concentration of all primer pairs that worked was 0.01 μg/μl.

with certainty that false positive results are not being observed in the common allele assays. Work is ongoing to confirm that all the assays are working reliably.

5.4 DISCUSSION

5.4.1 The significance of sequence variations in 1C7, AIF-1 and BAT1 and the ongoing genotyping effort

The polymorphisms identified in this study and their possible significance in relation to the expression and function of the relevant gene products are discussed in this section. The majority of polymorphisms identified, however, do not fall within coding regions or promoter regions and it seems unlikely that they could have any effect on the function of the genes in question. Such polymorphisms may be useful as markers for closely linked loci that were not detected in this study or lie outside the regions studied here. There is good evidence for susceptibility loci for CVID/IgAD and possibly also for IDDM or MG within, or in close proximity to the AIF-1/BAT1 region. Polymorphic information from this region may aid the identification of these loci.

At this point no useful statistical analysis can be carried out on the data described in section 5.2 as the sample size is far too small for any linkage or association study. Therefore, the PCR-SSP genotyping of CVID patients and controls was undertaken. The PCR-SSP method developed by the Tissue Typing Centre with whom we are collaborating in this effort, has been found in initial trials (section 5.3.2) to be very versatile and robust. The genotyping of the CVID and control DNAs should be completed within the next few months and the data generated should bring us closer to explaining the disease association of CVID with the telomeric end of the MHC class III region.

5.4.2 Variation within the AIF-1, 1C7, ATP6G and BAT1 genes

5.4.2.1 The AIF-1 Gene

A full explanation of the alternatively spliced forms and products of the AIF-1 gene is given in section 4.3.1. Two SNPs have been identified within the coding region of different AIF-1 transcripts (Figure 5.3). AIF1*825 lies in the coding region of the IRT-1 transcript in exon 4III and AIF1*206 (a difference with database sequence AC004181 which has been confirmed by PCR-SSP (section 5.3.2 and Figure 5.7)) lies in exon 2. Both these SNPs cause potentially significant amino acid substitutions.

AIF1*206 is a G to A transitional polymorphism that causes a glycine to arginine substitution at residue 14 within the protein encoded by the AIF-1 transcript. Glycine-14 falls within a block of 12 amino acids near the N-terminus of the AIF-1 protein which have remained completely conserved despite the evolutionary distance between zebrafish and man (Figure 4.3). Such conservation would suggest that this domain has significant functional constraints that would not tolerate variation. The substitution of glycine (a small non-polar residue) by arginine (a positively charged basic residue) could, therefore, significantly affect the function of this domain. The preliminary genotyping data indicate that the A allele of AIF1*206 is very rare (a frequency of ~1.5%) in the CVID DNAs. The significance of this cannot be assessed until the frequency of this SNP is known in the control DNAs, which are still to be assayed.

The second SNP that causes a codon change in the AIF-1 gene, AIF1*825, is a C to T transitional polymorphism that causes an arginine to tryptophan substitution at residue 15 in the product of the IRT-1 transcript. As with AIF1*206 this is a significant change, this time from a basic residue to a non-polar residue. In section 4.3.1 a number of potential di-

basic endoprotease cleavage sites were identified and discussed. Arg-15 lies within the first of these di-basic sites in IRT-1 and its replacement by Trp in the AIF1*825 variant allele (Lys-Arg to Lys-Trp) disrupts this potential cleavage site. Further work is needed to assess whether the di-basic site around Arg-15 is utilised and whether the above substitution is of functional significance.

Among the other 7 SNPs identified in AIF-1 by mutation analysis two fall within the IRT-1 5' untranslated region and two within the G1 5' untranslated region, whilst there are three successive G/A SNPs within the intron 1 of AIF-1 (Figure 5.3). The significance of these mutations is unclear.

5.4.2.2 The 1C7 gene

Nine 1C7 SNPs have been identified in this study (Figure 5.4). An additional SNP was identified by Dr R.E. March (1C7*2859) in a study of IDDM patients. A further ten differences have been identified with EMBL database entry U00921 some of which may turn out to be haplotype specific. This makes the 1C7 gene region the most polymorphic of the four regions screened. Of the ten SNPs identified by mutation analysis, 1C7*-451 and 1C7*-412 fall within the promoter region and, therefore, could possibly influence transcription. SNPs 1C7*2708 and 1C7*3577 fall within the coding region of 1C7, but are silent and, therefore, can be discounted as potential disease causing mutations. However, they have been included within the genotyping investigation as they may be useful as markers. The same is true for 1C7*3008 which falls within intron 2.

SNPs 1C7*2859 and 1C7*3571 both cause amino acid substitutions. 1C7*2859 was identified in a study using a method similar to that described here in a screen for IDDM

susceptibility genes (R.E March, personal communication). It causes an alanine to serine substitution within the immunoglobulin domain of 1C7 between β strands E and F and could possibly effect ligand binding. This mutation was, therefore, considered a strong candidate for influencing disease susceptibility. However, no variant alleles of this SNP have been observed in the CVID DNAs tested here. The relevance of this fact cannot be commented on until the screening of the control DNAs has been completed. SNP 1C7*3571 falls within the coding region of exon 4III causing an amino acid substitution from arginine (a charged basic residue) to serine (an uncharged polar residue). This substitution is in close proximity to a potential SH3 domain binding motif (section 4.3.2) on the cytoplasmic tail of 1C7 transcripts utilising exon 4III (section 4.2.2). Considerable work is still needed to assess the functional significance of the potential SH3 domain binding motifs identified in this study and whether SNP 1C7*3571 affects their activity. However, since SH3 domain binding motifs are involved in signal transduction, SNP 1C7*3571 is also considered a candidate disease susceptibility loci.

Both 1C7*3790 and 1C7*3843 disrupt polyadenylation signals. 1C7*3790 causes a AATAAA to AACAAA change within the polyadenylation signal of 1C7 transcripts utilising exons 4III or 4II. 1C7*3843, on the other hand, causes a AATAAA to AATTAA change within the polyadenylation signal of LST-1 (section 3.3.3) which is encoded on the reverse strand and overlaps the 1C7 gene (Figure 5.4). Deviations from the conserved polyadenylation signal (AATAAA) have been observed *in vivo* (Wahle & Keller 1997) although it seems likely that they have a reduced efficiency (Sheets *et al.* 1990). However, investigations into the effects of such variations *in vitro*, have demonstrated that the mutations identified here for 1C7 and LST-1 would effectively prevent the cleavage and polyadenylation of transcripts (Wickens & Stephenson 1984; Sheets *et al.* 1990).

Significantly, the T to C transition seen in the 1C7 polyadenylation signal has also been observed in the polyadenylation signal of a β-globin gene cloned from an individual with β-thalassemia and was found to be the cause of this condition (Orkin *et al.* 1985). In this example the efficiency of polyadenylation was reduced but processed transcripts were detected *in vivo,* as a second signal, 900 nucleotides downstream, was utilised. In the case of 1C7 no other polyadenylation motif has been detected in more than 5kb of sequence downstream. Therefore, in this case, unlike in the β-thalassemia patient, it is unlikely that a homozygous variant individual could generate processed transcripts of 1C7 containing exons 4II and 4III. The significance of this is unclear as the only 1C7 transcripts shown to be expressed at the protein level are those containing exon 4I (section 4.3.2). These utilise a different polyadenylation signal and thus would not be effected by the 1C7*3790 SNP.

1C7*3790 and 1C7*3843, at first glance, are among the most important SNPs identified in this study and work is ongoing to genotype them. At the time of writing this Thesis there was no genotyping data available so no comment can be made on their significance. Unfortunately, it is unlikely that it will be possible to assay the LST-1 SNP (1C7*3843) with PCR-SSP as it falls within a T and A repeat region. However, SNP 1C7*3843 can be detected using denaturing HPLC (WAVE) (G. Ribas-Despuig, personal communication). Therefore, this technology could be applied to typing this SNP.

5.4.2.3 The ATP6G gene

Seven SNPs were identified within the ATP6G gene. None of these lie within coding regions and although two SNPs fall within the 5' untranslated region (Figure 5.5) it is unclear how these would effect expression or gene function. Comparison with EMBL database entry Z15027 revealed a number of differences with the TN62/TN82 genomic

sequence (Figure 5.5) including a nucleotide deletion (T) within exon 3. This results in a change in the last 16 amino acids of ATP6G and an extension of the polypeptide for an additional 27 residues until a stop codon is encountered. Further analysis is needed to assess whether this deletion is real or whether it is the result of a sequencing error.

5.4.2.3 The BAT1 promoter region

The 15 nucleotide deletion within the BAT1 promoter region (accession number Z37166) was first identified as a difference with genomic EMBL database entry AC004181. As it falls just 30 nucleotides upstream of the putative transcriptional start site of the BAT1 database entry (Z37166), and is thus within the promoter region, this deletion was considered to be significant. Mutation analysis has confirmed that BAT1*-180/165 is a real polymorphism and, therefore, should be considered as a potential disease causing mutation due to its position within the BAT1 promoter region. The BAT1*-180/165 insertion/deletion has been included in the genotyping investigation outlined in section 5.3, but no genotyping data is yet available concerning this polymorphism.

The BAT1*-180/165 polymorphism also has implications in relation to the findings described in Chapter 3 regarding BAT1 splicing. Section 3.3.4 describes the identification of a novel splice variant of the BAT1 gene, which contains an additional GC-rich exon, 5' of exon 1 of the BAT1 database entry (Z37166). We hypothesised that if both the splice variants are functional and utilise different promoters, then the novel splice variant would be translated less efficiently than the splice variant represented by the database entry due to its GC-rich 5' untranslated region causing extensive secondary structure (see section 3.3.4 for details). The BAT1*-180/165 polymorphism falls within the first intron of this novel splice variant and would, therefore, be quite unlikely to have any effect on this transcript.

If then, the BAT1*-180/165 deletion does knock out, or reduce the efficiency of, the promoter for the splice variant representing the BAT1 database entry, only the GC encumbered transcript would be transcribed in individuals homozygous for the deletion. Thus, according to our hypothesis, the product of the BAT1 gene would still be generated but less efficiently.

5.4.3 DNA variation and the Human Genome Project

A total of 7.4kb of genomic DNA has been screened by sequence analysis of 24 chromosomes and this has led to the identification of 28 polymorphisms. These can be subdivided into 26 single nucleotide polymorphisms (SNPs) and 2 insertions/deletions (indels). The overall polymorphism frequency observed is ~3.8 per kb. It is likely though that this figure is an underestimate and that the screening of 24 chromosomes was insufficient to detect all polymorphisms in this region. Screening of the PCR products tested here with denaturing HPLC against 24 cell lines has revealed additional SNPs (G. Ribas-Despuig, personal communication) and comparison with database sequences has identified candidate SNPs (Figures 5.3 to 5.6) that, at least in the case of the SNP 1C7*206, have been proven to be real (section 5.3.2). Even so, this is still higher than previous estimates of between 1-2 SNPs per Kb for the genome as a whole. However, as the Human Genome Project nears its completion date it is becoming clear that this is a significant underestimate. DNA polymorphism discovery resources are being assembled to develop high throughput mutation detection technologies (Collins *et al.* 1998) and SNP databases are growing. As of 10th January 2000, the database dbSNP (WWW.ncbi.nih.gov/SNP/) contained 26249 SNPs. Based on this, SNPs which represent the commonest form of variation within the genome, are thought to occur once every 100 to 300 nucleotides. This revised figure puts the sequence variation observed in this study at

around the average for the genome as a whole. The occurrence of SNPs is, however, not uniform. For example, the frequency of variation in the sequence flanking the DRB1 locus in the MHC class II region averages about 5% over 220 kb and reaches as high as 10% locally, whereas in the neighbouring chromosome band 6p23 the region encompassing the *SCA1* gene has a variation frequency of only ~0.12% over 138kb (Horton *et al.* 1998).

The generation of SNP databases will clearly have great benefits. DNA sequence variants are used in all aspects of genetic investigation including evolutionary and population studies, the analysis, diagnosis and identification of genetic diseases, forensics and the study of human phenotypic variation. The completion of the Human Genome Project will mark a major achievement in genetic study, but of equal importance will be the development of high throughput SNP detection and scoring technologies. The sequence based approach used in this investigation has proven to be a very informative and robust approach and suited the purposes of this small scale pilot screen. However, as a method of large scale mutation detection, direct sequencing is too labour intensive and expensive to meet the demands of the post-Human Genome Project era. Other systems such as oligonucleotides arrays (DNA chips) or conformational change detection technologies such as single-strand conformation polymorphism (SSCP) analysis or denaturing high performance liquid chromatography (DHPLC) will be more appropriate in the future.

Chapter 6

CHAPTER 6: CONCLUSIONS

6.1 THE COMPLEXITY OF TRANSCRIBED REGIONS IDENTIFIED.

This Thesis has described the detailed characterisation of an 82kb segment of DNA at the telomeric end of the MHC class III region. One striking feature of this region is the apparent complexity of the transcripts identified. The genes AIF-1, 1C7, LST-1, ATP6G and BAT1 all appear to give rise to multiple spliced species at the RNA level. In the cases of LST-1 and 1C7, at least, a large number of different variants have been detected, each encoding a distinct protein product. In addition, AIF-1 and BAT1 (and possibly LST-1) appear to utilise more than one promoter region whilst 1C7 and ATP6G show evidence of utilising more than one polyadenylation signal. In fact, nearly every category of transcriptional variation currently known can be found in this relatively small segment of DNA. Therefore, at first glance this region would appear to be exceptionally complex. However, analysis of the human genome is still in its infancy. As the EST databases continue to grow and analysis software improves it will eventually be possible to apply a similar detailed analysis to that carried out here to the whole genome. It may then become apparent that the level of complexity identified for this region is not unusual. It is also important to note that although it is tempting to envisage the transcriptional and translational machinery working in a faultless manner, this is most likely not the case. There is as yet no proof that all the 1C7 and LST-1 transcripts are functional. It seems likely that the transcriptional machinery is far from perfect, generating miss-spliced and incompletely spliced transcripts at high enough levels to be detected by RT-PCR and to be incorporated in EST databases. A great deal of work is still needed to assess the true importance of these multiple transcripts. Also, whilst the EST databases are an invaluable resource it is not yet know what percentage of the sequences entered into them represents

genomic contamination and incomplete/miss-spliced transcripts. These resources should, therefore, be used with caution.

Whilst noting the pitfalls and restrictions of database analysis, the wealth of data presented here further supports previous statements that the MHC class III region is one of the most complex regions of the human genome so far analysed. In addition, it remains one of the most gene dense regions, with on average one gene per 14.3kb of DNA. Indeed, the gene density of the 82kb region discussed here is now one gene per 8kb of DNA.

6.2 THE FUNCTION OF ENCODED GENE PRODUCTS

The MHC class I and class II regions have long been associated with the immune system (Chapter 1). The connection of the MHC class III region with immunity on the other hand has, in the past, been questioned. However, it is becoming increasingly apparent from information already available on genes such as LTB, TNF and LTA and from the results described here, that many of the genes located within the class III region are involved in the immune and/or inflammatory responses. There is good evidence that six out of the nine genes identified in the TN62/TN82 region encode proteins with an immune-related function: LTB/TNF/LTA are known cytokines; AIF-1 is expressed by immune-related cells and has been implicated in allograft rejection; 1C7 is also expressed in immune-related cells and most recently one of its splice variants has been identified as an NK cell triggering receptor; and LST-1 exhibits a leukocyte-restricted expression pattern and is stimulated by IFN-γ. In addition to this, there is a tentative involvement of the V-ATPase G-subunit in macrophage acidity regulation. Of the remaining two genes, IKBL and BAT1, little is known, but it cannot be discounted that these, also, may play a role in immunity.

6.3 FUTURE WORK

The study presented here has generated a large amount of data. However, it has been reiterated throughout this Thesis that although transcripts have been shown to exist at the RNA level we cannot assume they all encode proteins with functional relevance. A considerable amount of work is still needed to fully characterise the transcripts identified in this study. For example, LST-1 has not yet been identified at the protein level. The significance of the large number of alternatively spliced LST-1 transcripts is, therefore, not known.

A large number of questions are still unanswered regarding the functions of the different 1C7 transcripts. We have identified a novel splicing of exon 2 of this gene that results in an IgV domain or an IgC2 domain in the encoded protein depending on the splice pattern. Further work is needed to identify whether both forms exist at the protein level *in vivo* and how this may affect function. Determination of the structure of 1C7 Ig domain (s) will aid this and will assist the identification of ligand binding sites. The splice variant identified here as 1C7c is now known to be the NK cell triggering receptor NKp30 (Pende *et al.* 1999). Therefore, this splice variant, at least, is known to exist at the protein level. However, no data is available on any of the other splice variants and further work is needed in this area. In addition, it appears that 1C7 is not a functional gene in mouse. This presents us with an ideal animal model with which to study this gene, especially as preliminary data suggests that it may be functional in rat. These findings need to be confirmed, however, before functional investigations in these organisms can proceed.

Lastly, a number of polymorphisms have been identified in the AIF-1, 1C7, and BAT1 genes and work is ongoing to genotype these polymorphisms in control DNAs and DNA

samples from CVID patients. The completion of this study will add valuable information to the hunt for disease susceptibility genes. This genotyping should also then be applied to DNA from patients with other medical conditions associated with this region of the MHC, such as IDDM. However, as the function of these genes is elucidated the search for candidate disease susceptibility genes will hopefully be narrowed down. In the interim, further mutation analysis, such as that reported in Chapter 5 will add valuable information.

6.4 FINAL COMMENTS

The association of the MHC, including the class III region, with the immune system and with diseases such as IDDM, RA, MG, HIV, Narcolepsy and CVID/IgAD has made this segment of DNA the focus of considerable efforts in terms of sequencing projects, both on a megabase scale and on a more modest kilobase scale such as this. However, it is because of the modest scale of this project that such a large amount of data has been generated and, although the region sequenced here represents a fraction of a percent of the entire Human Genome Project, it is an important region and, therefore, has warranted close scrutiny. The work presented in this Thesis is a good example of the amount of data that can be generated from database analysis alone and although the telomeric end of the MHC may turn out to be a particularly gene dense and complex region, it is somewhat daunting to extrapolate the amount of work still remaining to be done, and the amount of data that could potentially be generated, in the analysis of the whole genome. What is even more imposing is that much of the data presented here is only a beginning that still needs to be backed up by a significant amount of experimental investigations.

Appendices

APPENDIX I

The TN62/TN62 EMBL database entry Y14768

```
ID HSY14768 standard; DNA; HUM; 81800 BP.
XX
AC Y14768;
XX
SV Y14768.1
XX
NI e1334696
XX
DT 25-OCT-1998 (Rel. 57, Created)
DT 25-OCT-1998 (Rel. 57, Last updated, Version 1)
XX
DE Homo sapiens DNA, cosmid clones TN62 and TN82
XX
KW 1C7 gene; AIF-1 gene; Allograft Inflammatory Factor-1; BAT1
gene;
KW CpG island; I-kappa B family; IKBL gene;
KW leukocyte specific transcript-1; LINE/L1; LINE/L2; LST-1 gene;
LTB gene;
KW lymphotoxin-alpha; lymphotoxin-beta; microsatellite; repeat
region;
KW SINE/Alu; SINE/MIR; TNF gene; tumor necrosis factor; V-ATPase;
KW vacuolar ATPase G subunit.
XX
OS Homo sapiens (human)
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
Eutheria;
OC Primates; Catarrhini; Hominidae; Homo.
XX
RN [2]
RP 1-81800
RA Neville M.J.;
RT ;
RL Submitted (05-SEP-1997) to the EMBL/GenBank/DDBJ databases.
RL M.J. Neville, University of Oxford, MRC Immunochemistry Unit,
Department
RL of Biochemistry, South Parks Road, Oxford OX1 3QU, UK
XX
RN [3]
RX MEDLINE; 93272029.
RA Iris F., Bougueleret L., Prieur S., Caterina D., Primas G.,
Perrot V.,
RA Jurka J., Rodriguez-Tome P., Claverie J., Cohen D., Dausset
J.;
RT "Dense Alu clustering and a potential new member of the
NFkappaB family
RT within a 90 kilobase HLA class III segment";
RL Nat. Genet. 3:137-145(1993).
XX
RN [4]
RA Utans U., Quist W.C., McManus B.M., Wilson J.E., Arceci R.J.,
RA Wallace A.F., Russell M.E.;
RT "Allograft Inflammatory Factory-1. A cytokine-responsive
macrophage
RT molecule expressed in transplanted human hearts";
RL Transplantation 61:1387-1392(1996).
XX
RN [5]
RX MEDLINE; 96006565.
RA Holzinger I., de Baey A., Messer G., Kick G., Zwierzina H.,
Weiss E.H.;
RT "Cloning and genomic characterization of LST1: a new gene in
the human
RT TNF region";
RL Immunogenetics 42:315-322(1995).
XX
RN [6]
RX MEDLINE; 93208881.
RA Browning J.L., Ngam-ek A., Lawton P., DeMarinis J., Tizard R.,
RA Chow E.P., Hession C., O'Brine-Greco B., Foley S.F., Ware
C.F.;
RT "Lymphotoxin-beta: A new member of the TNF family that forms a
RT heteromeric complex with lymphotoxin on the cell surface";
RL Cell 72:847-856(1993).
XX
RN [7]
RX MEDLINE; 86016093.
RA Nedwin G.E., Naylor S.L., Sakaguchi A.Y., Smith D., Jarrett-
Nedwin J.,
RA Pennica D., Goeddel D.V., Gray P.W.;
RT "Human lymphotoxin and tumor necrosis factor genes: structure,
homology
RT and chromosomal localization";
RL Nucleic Acids Res. 13:6361-6373(1985).
XX
RN [8]
RX MEDLINE; 91086846.
RA Messer G., Spengler U., Jung M.C., Honold G., Bloemer K., Pape
G.R.,
RA Riethmueller G., Weiss E.H.;
RT "Polymorphic Structure of the Tumor Necrosis Factor (TNF)
Locus: An NcoI
RT Polymorphism in the First Intron of the Human TNF-beta Gene
Correlates
RT with A Variant Amino Acid in Position 26 and a Reduced Level
of TNF-beta
RT Production";
RL J. Exp. Med. 173:209-219(1991).
XX
RN [9]
RX MEDLINE; 91139175.
RA Abraham L.J., Du D.C., Zahedi K., Dawkins R.L., Whitehead
A.S.;
RT "Haplotypic polymorphisms of the TNFB gene";
RL Immunogenetics 33:50-53(1991).
XX
RN [10]
RX MEDLINE; 94362679.
RA Albertella M.R., Campbell D.R.;
RT "Characterization of a novel gene in the human major
histocompatibility
RT complex that encodes a potential new member of the I kappa B
family of
RT proteins";
RL Hum. Mol. Genet. 3:793-799(1994).
XX
RN [11]
RX MEDLINE; 95324911.
RA Peelman L., Chardon P., Nunes M., Renard C., Geffrotin C.,
Vaiman M.,
RA Van Zeveren A., Coppieters W., Van de Weghe A., Bouquet Y.,
Choy W.,
RA Strominger J., Spies T.;
RT "The BAT1 Gene in the MHC Encodes an Evolutionarily Conserved
Putative
RT Nuclear RNA Helicase of the D-E-A-D Family";
RL Genomics 26:210-218(1995).
XX
RN [12]
RA Neville M.J., Campbell R.D.;
RT "Alternative splicing of the LST-1 gene located in the major
RT histocompatibility complex on human chromosome 6";
RL DNA Seq. 8:155-160(1997).
XX
RN [13]
RA de Baey A., Fellerhoff B., Maier S., Martinozzi S., Weidle U.,
RA Weiss E.H.;
RT "Complex expression patern of the TNF region gene LST1 through
RT differential regulation, initiation, and alternative
splicing.";
RL Genomics 45:591-600(1997).
XX
RN [14]
RA Shiina T., Tamiya G., Oka A., Yamagata T., Yamagata N.,
Kikkawa E.,
RA Goto K., Mizuki N., Watanabe K., Fukuzumi Y., Taguchi S.,
Sugawara C.,
RA Ono A., Chen L., Yamazaki M., Tashiro H., Ando S., Ikemura T.,
RA Kimura M., Inoko H.;
RT "Nucleotide sequencing analysis of the 146-kilobase segment
around the
RT IkBL and MICA genes at the centromeric end of the HLA class I
region";
RL Genomics 47:372-382(1998).
XX
RN [15]
RA Neville M.J., Milner C.M., Campbell R.D.;
RT "A new member of the immunoglobulin superfamily and a V-ATPase
G subunit
RT are amongst the predicted products of novel genes close to the
TNF locus
RT in the human MHC";
RL Unpublished.
XX
DR SPTREMBL; O14930; O14930.
DR SPTREMBL; O14931; O14931.
DR SPTREMBL; O14932; O14932.
DR SPTREMBL; O43904; O43904.
DR SPTREMBL; O95667; O95667.
DR SPTREMBL; O95668; O95668.
DR SPTREMBL; O95669; O95669.
DR SPTREMBL; O95670; O95670.
XX
CC Related publications:
CC 1) Olavesen M.G., Proc. 11th Intl. Histocompatibility workshop
&
CC conference 2,190-193 (1991)
CC 2) Nalabolu S.R., Genomics 31,215-222 (1996)
XX
FH Key Location/Qualifiers
```

```
FH
FT source 1..81800
FT /chromosome="6"
FT /db_xref="taxon:9606"
FT /haplotype="HLA A2,B7,DR2,C2c,BFs,C4A 3,C4B QO"
FT /organism="Homo sapiens"
FT /lab_host="E.coli (NM554)"
FT /cell_line="EBV transformed lymphoblastoid cell line"
FT /cell_type="lymphoblastoid"
FT /clone_lib="MboI cut genomic DNA cloned into PDVCOS cosmid
FT vector"
FT /clone="TN62"
FT /clone="TN82"
FT /sub_clone="M13mp18"
FT /map="p21.3"
FT repeat_region 10..227
FT /note="AluSg"
FT /rpt_family="SINE/Alu"
FT repeat_region 243..538
FT /note="AluSp"
FT /rpt_family="SINE/Alu"
FT repeat_region 617..914
FT /note="AluSc"
FT /rpt_family="SINE/Alu"
FT repeat_region 938..1064
FT /note="AluSq/x"
FT /rpt_family="SINE/Alu"
FT repeat_region 1354..1459
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT repeat_region complement(1485..1786)
FT /note="AluSq"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(1871..2164)
FT /note="AluSp"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(2228..2302)
FT /note="MIR"
FT /rpt_family="SINE/MIR"
FT exon complement(2583..3509)
FT /note="4III"
FT /number=4
FT /gene="AIF1"
FT exon complement(2583..2788)
FT /number=6
FT /gene="AIF1"
FT polyA_signal complement(2604..2609)
FT /gene="AIF1"
FT exon complement(3098..3260)
FT /number=5
FT /gene="AIF1"
FT exon complement(3459..3592)
FT /note="4II"
FT /number=4
FT /gene="AIF1"
FT exon complement(3459..3500)
FT /note="4I"
FT /number=4
FT /gene="AIF1"
FT CDS join(complement(3459..3492),complement(3098..3260),
FT complement(2704..2788))
FT /citation=[15]
FT /db_xref="PID:e1334697"
FT /gene="AIF1"
FT /product="G1, putative splice variant of allograft
FT inflamatory factor-1"
FT /protein_id="CAA75060.1"
FT /translation="MEFDLNGNGDIDIMSLKRMLEKLGVPKTHLELKKLIGEVSSGSGE
FT TFSYPDFLRMMLGKRSAILKMILMYEEKAREKEKPTGPPAKKAISELP"
FT mRNA join(complement(3459..3592),complement(3098..3260),
FT complement(2583..2788))
FT /citation=[15]
FT /gene="AIF1"
FT /product="G1, putative splice variant of allograft
FT inflamatory factor-1"
FT exon complement(3561..3934)
FT /note="3II"
FT /number=3
FT /gene="AIF1"
FT exon complement(3868..3934)
FT /note="3I"
FT /number=3
FT /gene="AIF1"
FT exon complement(4022..4083)
FT /number=2
FT /gene="AIF1"
FT mRNA join(complement(4022..4083),complement(3561..3934),
FT complement(2583..3509))
FT /note="interferon gamma responsive transcript"
FT /note="U95213"
FT /gene="AIF1"
FT /product="IRT-1"
FT CDS complement(3094..3492)
FT /db_xref="PID:e1334698"
FT /db_xref="SPTREMBL:O43904"
FT /note="interferon gamma responsive transcript"
FT /note="U95213"
FT /gene="AIF1"
FT /product="IRT-1"
FT /protein_id="CAA75061.1"
FT /translation="MEFDLNGNGDIGEKRVICGGRVVCRPKKTEVSPTCSIPHDLGGGP
FT PTTVGGRRMGMRKWERRERVSPPSPHPHPLPPDIMSLKRMLEKLGVPKTHLELKKLIGE
FT VSSGSGETFSYPDFLRMMLGKRSAILKM"
FT mRNA join(complement(4250..4348),complement(4022..4083),
FT complement(3868..3934),complement(3459..3500),
FT complement(3098..3260),complement(2583..2788))
FT /citation=[4]
FT /note="U19713"
FT /gene="AIF1"
FT /product="allograft inflamatory factor-1"
FT CDS join(complement(4250..4274),complement(4022..4083),
FT complement(3868..3934),complement(3459..3500),
FT complement(3098..3260),complement(2704..2788))
FT /citation=[4]
FT /db_xref="PID:e1334699"
FT /note="U19713"
FT /gene="AIF1"
FT /product="allograft inflamatory factor-1"
FT /protein_id="CAA75062.1"
FT /translation="MSQTRDLQGGKAFGLLKAQQEERLDEINKQFLDDPKYSSDEDLPS
FT KLEGFKEKYMEFDLNGNGDIDIMSLKRMLEKLGVPKTHLELKKLIGEVSSGSGETFSYP
FT DFLRMMLGKRSAILKMILMYEEKAREKEKPTGPPAKKAISELP"
FT exon complement(4250..4348)
FT /number=1
FT /gene="AIF1"
FT repeat_region complement(5465..5759)
FT /note="AluSc"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(5879..6024)
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT repeat_region 6037..6336
FT /note="AluSg"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(6430..6729)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(6730..7006)
FT /note="AluSg"
FT /rpt_family="SINE/Alu"
FT repeat_region 7038..7204
FT /note="FRAM"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(7210..7262)
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT repeat_region complement(7263..7563)
FT /note="AluSg"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(7879..8173)
FT /note="AluSc"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(8878..9176)
FT /note="AluY"
FT /rpt_family="SINE/Alu"
FT repeat_region 9181..9922
FT /note="MER21B"
FT /rpt_family="Other/MER21_gro"
FT repeat_region 9918..9965
FT /note="simple repeat (CA)n"
FT /rpt_family="microsatellite 82-1"
FT repeat_region complement(9984..10265)
FT /note="AluJb"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(10274..10568)
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(10583..10881)
FT /note="AluY"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(10882..11175)
FT /note="AluY"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(11214..11393)
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT repeat_region 11530..11809
FT /note="AluSp"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(12174..12333)
FT /note="AluJb"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(12339..12642)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(12657..12871)
FT /note="AluSg/x"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(12927..13055)
FT /note="AluYb8"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(13059..13358)
FT /note="AluSc"
FT /rpt_family="SINE/Alu"
```

```
FT repeat_region complement(13636..13927)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(13963..14015)
FT /note="MER80"
FT /rpt_family="DNA/Other"
FT repeat_region complement(14118..14414)
FT /note="AluY"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(14515..14799)
FT /note="AluSg"
FT /rpt_family="SINE/Alu"
FT repeat_region 15103..15341
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region 15374..15415
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT repeat_region 15470..15759
FT /note="MER35"
FT /rpt_family="Other"
FT repeat_region complement(15781..16164)
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT repeat_region 16198..16324
FT /note="FLAM_A"
FT /rpt_family="SINE/Alu"
FT repeat_region 16360..16661
FT /note="AluSp"
FT /rpt_family="SINE/Alu"
FT repeat_region 16923..17193
FT /note="L1MB8"
FT /rpt_family="LINE/L1"
FT repeat_region 17194..17321
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(17322..17628)
FT /note="AluY"
FT /rpt_family="SINE/Alu"
FT repeat_region 17629..17807
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region 17809..18022
FT /note="L1ME3"
FT /rpt_family="LINE/L1"
FT repeat_region 18033..18334
FT /note="AluY"
FT /rpt_family="SINE/Alu"
FT repeat_region 18397..18691
FT /note="AluSq"
FT /rpt_family="SINE/Alu"
FT repeat_region 19097..19399
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region 19517..19812
FT /note="AluSg"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(20293..20592)
FT /note="AluSq"
FT /rpt_family="SINE/Alu"
FT repeat_region 21026..21335
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region 21541..21839
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region 21870..22170
FT /note="AluSg"
FT /rpt_family="SINE/Alu"
FT repeat_region 22187..22421
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region 22573..22771
FT /note="TIGGER1"
FT /rpt_family="DNA/MER2_type"
FT repeat_region complement(22772..22932)
FT /note="AluJo/FRAM"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(22933..23224)
FT /note="AluSp"
FT /rpt_family="SINE/Alu"
FT repeat_region 23257..23547
FT /note="AluJb"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(23551..23706)
FT /note="FAM"
FT /rpt_family="SINE/Alu"
FT repeat_region 23707..23763
FT /note="TIGGER1"
FT /rpt_family="DNA/MER2_type"
FT repeat_region 23764..24064
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region 24068..24128
FT /note="TIGGER1"
FT /rpt_family="DNA/MER2_type"
FT repeat_region complement(24143..24372)
FT /note="AluJb"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(24373..24673)
FT /note="AluY"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(24674..24760)
FT /note="AluS"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(24808..25103)
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region 25104..25248
FT /note="TIGGER1"
FT /rpt_family="DNA/MER2_type"
FT repeat_region complement(25250..25550)
FT /note="AluSp"
FT /rpt_family="SINE/Alu"
FT repeat_region 25553..25659
FT /note="TIGGER1"
FT /rpt_family="DNA/MER2_type"
FT repeat_region 25672..25940
FT /note="AluJb"
FT /rpt_family="SINE/Alu"
FT repeat_region 25939..26214
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region 26194..26285
FT /note="simple repeat (CA)n"
FT /rpt_family="microsatellite 82-3"
FT CDS join(26891..26933,29486..29830,29979..30086,30436..30545)
FT /db_xref="PID:e1334700"
FT /db_xref="SPTREMBL:O14931"
FT /gene="1C7"
FT /product="1C7a"
FT /protein_id="CAA75063.1"
FT /translation="MAWMLLLILIMVHPGSCALWVSQPPEIRTLEGSSAFLPCSFNASQ
FT GRLAIGSVTWFRDEVVPGKEVRNGTPEFRGRLAPLASSRFLHDHQAELHIRDVRGHDAS
FT IYVCRVEVLGLGVGTGNGTRLVVEKEHPQLGAGTVLLLRAGFYAVSFLSVAVGSTVYYQ
FT GKCLTWKGPRRQLPAVVPAPLPPPCGSSAHLLPPVPGG"
FT CDS join(26891..26933,29486..29830,29979..30086,30381..30418)
FT /db_xref="PID:e1334701"
FT /db_xref="SPTREMBL:O14930"
FT /gene="1C7"
FT /product="1C7b"
FT /protein_id="CAA75064.1"
FT /translation="MAWMLLLILIMVHPGSCALWVSQPPEIRTLEGSSAFLPCSFNASQ
FT GRLAIGSVTWFRDEVVPGKEVRNGTPEFRGRLAPLASSRFLHDHQAELHIRDVRGHDAS
FT IYVCRVEVLGLGVGTGNGTRLVVEKEHPQLGAGTVLLLRAGFYAVSFLSVAVGSTVYYQ
FT GKYAKSTLSGFPQL"
FT CDS join(26891..26933,29486..29830,29979..30086,30224..30300)
FT /db_xref="PID:e1334702"
FT /db_xref="SPTREMBL:O14932"
FT /gene="1C7"
FT /product="1C7c"
FT /protein_id="CAA75065.1"
FT /translation="MAWMLLLILIMVHPGSCALWVSQPPEIRTLEGSSAFLPCSFNASQ
FT GRLAIGSVTWFRDEVVPGKEVRNGTPEFRGRLAPLASSRFLHDHQAELHIRDVRGHDAS
FT IYVCRVEVLGLGVGTGNGTRLVVEKEHPQLGAGTVLLLRAGFYAVSFLSVAVGSTVYYQ
FT GKCHCHMGTHCHSSDGPRGVIPEPRCP"
FT exon 26642..26933
FT /number=1
FT /gene="1C7"
FT CDS join(26891..26933,29486..29637,29713..29830,29979..30086,
FT 30436..30545)
FT /db_xref="PID:e1334703"
FT /db_xref="SPTREMBL:O95667"
FT /gene="1C7"
FT /product="1C7e"
FT /protein_id="CAA75066.1"
FT /translation="MAWMLLLILIMVHPGSCALWVSQPPEIRTLEGSSAFLPCSFNASQ
FT GRLAIGSVTWFRDEVVPGKEAELHIRDVRGHDASIYVCRVEVLGLGVGTGNGTRLVVEK
FT EHPQLGAGTVLLLRAGFYAVSFLSVAVGSTVYYQGKCLTWKGPRRQLPAVVPAPLPPPC
FT GSSAHLLPPVPGG"
FT CDS join(26891..26933,29486..29637,29713..29830,29979..30086,
FT 30381..30418)
FT /db_xref="PID:e1334704"
FT /db_xref="SPTREMBL:O95668"
FT /gene="1C7"
FT /product="1C7d"
FT /protein_id="CAA75067.1"
FT /translation="MAWMLLLILIMVHPGSCALWVSQPPEIRTLEGSSAFLPCSFNASQ
FT GRLAIGSVTWFRDEVVPGKEAELHIRDVRGHDASIYVCRVEVLGLGVGTGNGTRLVVEK
FT EHPQLGAGTVLLLRAGFYAVSFLSVAVGSTVYYQGKYAKSTLSGFPQL"
FT CDS join(26891..26933,29486..29637,29713..29830,29979..30086,
FT 30224..30300)
FT /db_xref="PID:e1334705"
FT /db_xref="SPTREMBL:O95669"
FT /gene="1C7"
FT /product="1C7f"
FT /protein_id="CAA75068.1"
FT /translation="MAWMLLLILIMVHPGSCALWVSQPPEIRTLEGSSAFLPCSFNASQ
FT GRLAIGSVTWFRDEVVPGKEAELHIRDVRGHDASIYVCRVEVLGLGVGTGNGTRLVVEK
FT EHPQLGAGTVLLLRAGFYAVSFLSVAVGSTVYYQGKCHCHMGTHCHSSDGPRGVIPEPRCP"
FT repeat_region complement(27183..27348)
FT /note="AluSc"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(27349..27644)
```

```
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(27645..27775)
FT /note="AluSq/x"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(27790..28078)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 28036
FT /replace="c"
FT /citation=[5]
FT conflict 28062
FT /replace="c"
FT /citation=[5]
FT conflict 28064..28065
FT /replace="a"
FT /citation=[5]
FT repeat_region 28161..28293
FT /note="MER5A"
FT /rpt_family="DNA/MER1_type"
FT conflict 28468^28469
FT /replace="g"
FT /citation=[5]
FT repeat_region 28479..28774
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 28518
FT /replace=""
FT /citation=[5]
FT conflict 28576^28577
FT /replace="g"
FT /citation=[5]
FT conflict 28825
FT /replace="c"
FT /citation=[5]
FT conflict 28864^28865
FT /replace="a"
FT /citation=[5]
FT repeat_region complement(28892..29292)
FT /note="AluSg"
FT /rpt_family="SINE/Alu"
FT conflict 29275..29276
FT /replace=""
FT /citation=[5]
FT conflict 29279
FT /replace="g"
FT /citation=[5]
FT conflict 29358
FT /replace=""
FT /citation=[5]
FT conflict 29471
FT /replace="c"
FT /citation=[5]
FT conflict 29476
FT /replace="t"
FT /citation=[5]
FT conflict 29478
FT /replace="c"
FT /citation=[5]
FT exon 29486..29830
FT /number=2
FT /gene="1C7"
FT exon 29486..29637
FT /note="2I"
FT /number=2
FT /gene="1C7"
FT conflict 29488^29489
FT /replace="t"
FT /citation=[5]
FT conflict 29673..29676
FT /replace=""
FT /citation=[5]
FT conflict 29680..29681
FT /replace="tc"
FT /citation=[5]
FT exon 29713..29830
FT /note="2II"
FT /number=2
FT /gene="1C7"
FT conflict 29753^29754
FT /replace="g"
FT /citation=[5]
FT conflict 29870
FT /replace=""
FT /citation=[5]
FT exon 29979..30086
FT /number=3
FT /gene="1C7"
FT conflict 29993^29994
FT /replace="g"
FT /citation=[5]
FT conflict 30013
FT /replace=""
FT /citation=[5]
FT exon 30224..30337
FT /note="4I"
FT /number=4
FT /gene="1C7"
FT polyA_signal 30319..30324
FT /note="1C7c and 1C7f only"
FT /gene="1C7"
FT exon 30381..30711
FT /note="4II"
FT /number=4
FT /gene="1C7"
FT exon 30436..30711
FT /note="4III"
FT /number=4
FT /gene="1C7"
FT polyA_signal 30678..30683
FT /gene="1C7"
FT polyA_signal 30678..30678
FT /note="1C7a, 1C7b, 1C7d and 1C7e only"
FT /gene="1C7"
FT conflict 30680
FT /replace="c"
FT /citation=[5]
FT polyA_site 30711
FT /gene="1C7"
FT polyA_signal complement(30731..30736)
FT /gene="LST1"
FT conflict 30889
FT /replace=""
FT /citation=[5]
FT repeat_region complement(31131..31242)
FT /note="simple repeat (GA)n"
FT /rpt_family="microsatellite TNFd"
FT conflict 31132
FT /replace=""
FT /citation=[5]
FT repeat_region complement(31231..31360)
FT /note="FLAM_A"
FT /rpt_family="SINE/Alu"
FT conflict 31258
FT /replace=""
FT /citation=[5]
FT conflict 31329
FT /replace=""
FT /citation=[5]
FT repeat_region complement(31361..31551)
FT /note="simple repeat (GA)n"
FT /rpt_family="microsatellite TNFe"
FT conflict 31719..31721
FT /replace=""
FT /citation=[5]
FT mRNA join(complement(32294..32495),complement(31879..31971),
FT complement(31646..31668),complement(30703..31049))
FT /note="leukocyte specific transcript-1"
FT /note="splice variant LST1-4 in citation [12]"
FT /gene="LST1"
FT repeat_region 32396..32425
FT /note="MIR"
FT /rpt_family="SINE/MIR"
FT conflict 32480
FT /replace=""
FT /citation=[5]
FT mRNA join(complement(32496..32580),complement(32294..32412),
FT complement(31879..31971),complement(31646..31668),
FT complement(30703..31094))
FT /note="leukocyte specific transcript-1"
FT /note="splice variant LST1-5 in citation [12]"
FT /note="splice variant LST1/G in citation [13]"
FT /gene="LST1"
FT mRNA join(complement(32496..32580),complement(32294..32412),
FT complement(31646..31668),complement(30703..31094))
FT /note="leukocyte specific transcript-1"
FT /note="splice variant LST1-6 in citation [12]"
FT /note="splice variant LST1/H in citation [13]"
FT /gene="LST1"
FT mRNA join(complement(32805..32913),complement(32294..32412),
FT complement(31879..31971),complement(30703..31094))
FT /citation=[12]
FT /citation=[5]
FT /note="leukocyte specific transcript-1"
FT /note="splice variant LST1-db"
FT /note="U00921"
FT /gene="LST1"
FT conflict 33059
FT /replace=""
FT /citation=[5]
FT conflict 33113
FT /replace=""
FT /citation=[5]
FT mRNA join(complement(33297..33410),complement(32294..32412),
FT complement(30703..31094))
FT /note="leukocyte specific transcript-1"
FT /note="splice variant LST1-3 in citation [12]"
FT /note="splice variant LST1/D in citation [13]"
FT /gene="LST1"
FT mRNA join(complement(33297..33410),complement(32294..32412),
FT complement(31879..31971),complement(31646..31668),
FT complement(30703..31094))
FT /note="leukocyte specific transcript-1"
FT /note="splice variant LST1/A in citation [13]"
```

```
FT /gene="LST1"
FT mRNA join(complement(32805..32913),complement(32294..32412),
FT complement(31879..31971),complement(30703..31115))
FT /note="leukocyte specific transcript-1"
FT /note="splice variant LST1/C in citation [13]"
FT /gene="LST1"
FT mRNA join(complement(32673..32764),complement(32294..32412),
FT complement(30703..31094))
FT /note="leukocyte specific transcript-1"
FT /note="splice variant LST1/E in citation [13]"
FT /gene="LST1"
FT mRNA join(complement(33297..33410),complement(32294..32412),
FT complement(31646..31668),complement(30703..31094))
FT /note="leukocyte specific transcript-1"
FT /note="splice variant LST1/F in citation [13]"
FT /gene="LST1"
FT exon complement(33297..33410)
FT /gene="LST1"
FT exon complement(32805..32913)
FT /gene="LST1"
FT exon complement(32673..32764)
FT /gene="LST1"
FT exon complement(32496..32580)
FT /gene="LST1"
FT exon complement(32294..32412)
FT /gene="LST1"
FT exon complement(32294..32495)
FT /gene="LST1"
FT exon complement(31879..31971)
FT /gene="LST1"
FT exon complement(31646..31668)
FT /gene="LST1"
FT exon complement(30703..31115)
FT /gene="LST1"
FT exon complement(30703..31094)
FT /gene="LST1"
FT exon complement(30703..31049)
FT /gene="LST1"
FT conflict 33300
FT /replace="c"
FT /citation=[5]
FT conflict 33308
FT /replace=""
FT /citation=[5]
FT repeat_region complement(34337..34622)
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(34824..35119)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(35133..35434)
FT /note="AluSq"
FT /rpt_family="SINE/Alu"
FT repeat_region 35469..35564
FT /note="MIR"
FT /rpt_family="SINE/MIR"
FT repeat_region 35594..35886
FT /note="AluJb"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(36055..36357)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region 36868..36944
FT /note="MIR"
FT /rpt_family="SINE/MIR"
FT exon 37185..37356
FT /number=1
FT /gene="LTB"
FT CDS join(37195..37356,37753..37798,37982..38053,38449..38903)
FT /db_xref="PID:e1334706"
FT /gene="LTB"
FT /product="lymphotoxin-beta"
FT /protein_id="CAA75069.1"
FT /translation="MGALGLEGRGGRLQGRGSLLLAVAGATSLVTLLLAVPITVLAVLA
FT LVPQDQGGLVTETADPGAQAQQGLGFQKLPEEEPETDLSPGLPAAHLIGAPLKGQGLGW
FT ETTKEQAFLTSGTQFSDAEGLALPQDGLYYLYCLVGYRGRAPPGGGDPQGRSVTLRSSL
FT YRAGGAYGPGTPELLLEGAETVTPVLDPARRQGYGPLWYTSVGFGGLVQLRRGERVYVN
FT ISHPDMVDFARGKTFFGAVMVG"
FT exon 37753..37798
FT /number=2
FT /gene="LTB"
FT exon 37982..38053
FT /number=3
FT /gene="LTB"
FT exon 38449..39054
FT /number=4
FT /gene="LTB"
FT polyA_signal 39020..39025
FT /gene="LTB"
FT conflict 39964^39965
FT /note="Z15026"
FT /replace="t"
FT /citation=[3]
FT conflict 39973^39974
FT /note="Z15026"
FT /replace="t"
FT /citation=[3]
FT conflict 39978
FT /note="Z15026"
FT /replace="t"
FT /citation=[3]
FT conflict 40022..40023
FT /note="Z15026"
FT /replace="gt"
FT /citation=[3]
FT repeat_region complement(40032..40323)
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(40486..40622)
FT /note="FLAM_C"
FT /rpt_family="SINE/Alu"
FT exon complement(41286..42496)
FT /number=4
FT /gene="TNF"
FT polyA_signal complement(41295..41300)
FT /gene="TNF"
FT conflict 41341
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT conflict 41355^41356
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT conflict 41414
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT conflict 41487^41488
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT exon complement(42798..42846)
FT /number=3
FT /gene="TNF"
FT conflict 42963^42964
FT /note="Z15026"
FT /replace="t"
FT /citation=[3]
FT conflict 42985^42986
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT exon complement(43037..43078)
FT /number=2
FT /gene="TNF"
FT conflict 43217
FT /note="X02159"
FT /replace=""
FT /citation=[7]
FT repeat_region complement(43220..43390)
FT /note="MER35"
FT /rpt_family="Other"
FT repeat_region complement(43437..43485)
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT conflict 43562
FT /note="X02159"
FT /replace="t"
FT /citation=[7]
FT conflict 43576^43577
FT /note="X02159"
FT /replace="t"
FT /citation=[7]
FT conflict 43637^43638
FT /note="X02159"
FT /replace="t"
FT /citation=[7]
FT exon complement(43685..44051)
FT /number=1
FT /gene="TNF"
FT CDS join(complement(43685..43870),complement(43037..43078),
FT complement(42798..42846),complement(42075..42496))
FT /db_xref="PID:e1334707"
FT /gene="TNF"
FT /product="tumor necrosis facor"
FT /protein_id="CAA75070.1"
FT /translation="MSTESMIRDVELAEEALPKKTGGPQGSRRCLFLSLFSFLIVAGAT
FT TLFCLLHFGVIGPQREEFPRDLSLISPLAQAGSSSRTPSDKPVAHVVANPQAEGQLQWL
FT NRRANALLANGVELRDNQLVVPSEGLYLIYSQVLFKGQGCPSTHVLLTHTISRIAVSYQ
FT TKVNLLSAIKSPCQRETPEGAEAKPWYEPIYLGGVFQLEKGDRLSAEINRPDYLDFAES
FT GQVYFGIIAL"
FT conflict 43898
FT /note="X02159"
FT /replace="c"
FT /citation=[7]
FT TATA_signal  complement(44073..44076)
FT /gene="TNF"
FT conflict 44294
FT /note="U42625"
FT /replace="t"
FT /citation=[7]
FT CAAT_signal  complement(44376..44379)
FT /gene="TNF"
```

```
FT conflict 44491^44492
FT /note="Z15026 and X02159"
FT /replace="g"
FT /citation=[3]
FT /citation=[7]
FT conflict 44507
FT /note="X02159"
FT /replace=""
FT /citation=[7]
FT conflict 44554
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT conflict 44555
FT /note="X02159"
FT /replace=""
FT /citation=[7]
FT conflict 44618
FT /note="X02159"
FT /replace=""
FT /citation=[7]
FT conflict 44618
FT /note="Z15026 and X59352"
FT /replace="c"
FT /citation=[3]
FT /citation=[8]
FT conflict 44627
FT /note="X02159"
FT /replace=""
FT /citation=[7]
FT conflict 44638
FT /note="X02159"
FT /replace=""
FT /citation=[7]
FT conflict 44647..44649
FT /note="X02159"
FT /replace=""
FT /citation=[7]
FT conflict 44666..44668
FT /note="X02159"
FT /replace=""
FT /citation=[7]
FT conflict 44698
FT /note="Z15026 and X59352"
FT /replace="c"
FT /citation=[3]
FT /citation=[8]
FT conflict 44734^44735
FT /note="Z15026"
FT /replace="t"
FT /citation=[3]
FT conflict 44738
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT conflict 44913
FT /note="Z15026"
FT /replace="t"
FT /citation=[3]
FT conflict 45081
FT /note="Z15026"
FT /replace="g"
FT /citation=[3]
FT exon complement(45291..46331)
FT /number=4
FT /gene="LTA"
FT polyA_signal complement(45304..45309)
FT /gene="LTA"
FT conflict 45651
FT /note="X02911 and M55913"
FT /replace=""
FT /citation=[7]
FT /citation=[9]
FT conflict 45655^45656
FT /note="X02911 and M55913"
FT /replace="a"
FT /citation=[7]
FT /citation=[9]
FT conflict 46164
FT /replace="g"
FT /citation=[9]
FT exon complement(46579..46684)
FT /number=3
FT /gene="LTA"
FT conflict 46605
FT /note="X02911"
FT /replace="t"
FT /citation=[7]
FT exon complement(46771..46878)
FT /number=2
FT /gene="LTA"
FT CDS join(complement(46771..46869),complement(46579..46684),
FT complement(45919..46331))
FT /db_xref="PID:e1334708"
FT /gene="LTA"
FT /product="lymphotoxin-alpha"
FT /protein_id="CAA75071.1"
FT /translation="MTPPERLFLPRVCGTTLHLLLLGLLLVLLPGAQGLPGVGLTPSAA
FT QTARQHPKMHLAHSTLKPAAHLIGDPSKQNSLLWRANTDRAFLQDGFSLSNNSLLVPTS
FT GIYFVYSQVVFSGKAYSPKATSSPLYLAHEVQLFSSQYPFHVPLLSSQKMVYPGLQEPW
FT LHSMYHGAAFQLTQGDQLSTHTDGIPHLVLSPSTVFFGAFAL"
FT conflict 46833
FT /note="Z15026"
FT /replace="g"
FT /citation=[3]
FT repeat_region 46882..47024
FT /note="simple repeat (GA)n"
FT /rpt_family="microsatellite TNFc"
FT conflict 46960
FT /note="Z15026, X02911 and M55913"
FT /replace="c"
FT /citation=[3]
FT /citation=[7]
FT /citation=[9]
FT conflict 47076
FT /note="X02911"
FT /replace="c"
FT /citation=[7]
FT exon complement(47166..47327)
FT /number=1
FT /gene="LTA"
FT conflict 47170
FT /note="X59351"
FT /replace="g"
FT /citation=[8]
FT conflict 47248
FT /note="Z15026"
FT /replace="g"
FT /citation=[3]
FT conflict 47322
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT TATA_signal  complement(47351..47354)
FT /gene="LTA"
FT conflict 47621..47622
FT /note="Z15026"
FT /replace="ca"
FT /citation=[3]
FT conflict 47954
FT /note="X02911"
FT /replace="c"
FT /citation=[7]
FT conflict 48136^48137
FT /note="Z15026"
FT /replace="g"
FT /citation=[3]
FT conflict 48137^48138
FT /note="X02911"
FT /replace="g"
FT /citation=[7]
FT repeat_region 48253..48550
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 48534..48535
FT /note="Z15026"
FT /replace="aa"
FT /citation=[3]
FT repeat_region complement(48570..49016)
FT /note="L1ME2"
FT /rpt_family="LINE/L1"
FT conflict 48892
FT /note="Z15026"
FT /replace="t"
FT /citation=[3]
FT conflict 48986
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT conflict 48997
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT conflict 49025
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT conflict 49145
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT repeat_region 49386..49747
FT /note="MER39"
FT /rpt_family="Other/MER21_gro"
FT conflict 49783
FT /note="Z15026"
FT /replace="a"
FT /citation=[3]
FT repeat_region complement(49798..50101)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 49803..49804
FT /note="Z15026"
FT /replace=""
```

```
FT /citation=[3]
FT repeat_region 50304..50603
FT /note="AluY"
FT /rpt_family="SINE/Alu"
FT conflict 50593
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT conflict 50885
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT conflict 51165
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT repeat_region complement(51365..51903)
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT conflict 51437^51438
FT /note="Z15026"
FT /replace="g"
FT /citation=[3]
FT conflict 51491^51492
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT conflict 51498^51499
FT /note="Z15026"
FT /replace="a"
FT /citation=[3]
FT conflict 51512^51513
FT /note="Z15026"
FT /replace="t"
FT /citation=[3]
FT conflict 51516^51517
FT /note="Z15026"
FT /replace="g"
FT /citation=[3]
FT conflict 51519^51520
FT /note="Z15026"
FT /replace="ga"
FT /citation=[3]
FT conflict 51522
FT /note="Z15026"
FT /replace="t"
FT /citation=[3]
FT repeat_region complement(51909..51964)
FT /note="simple repeat (CA)n"
FT /rpt_family="microsatellite TNFa"
FT conflict 51927..51944
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT repeat_region 51970..52010
FT /rpt_family="microsatellite TNFb (GA)"
FT conflict 52005..52007
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT repeat_region complement(52036..52302)
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(52303..52598)
FT /note="AluSq"
FT /rpt_family="SINE/Alu"
FT conflict 52315
FT /note="Z15026"
FT /replace="a"
FT /citation=[3]
FT conflict 52492
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT repeat_region complement(52600..52640)
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(52653..52951)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 52952
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT repeat_region complement(52955..53251)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 53183
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT repeat_region complement(53255..53410)
FT /note="FRAM"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(53421..53713)
FT /note="AluSg"
FT /rpt_family="SINE/Alu"
FT conflict 53519
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT conflict 53559
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT conflict 53783
FT /note="Z15026"
FT /replace="a"
FT /citation=[3]
FT repeat_region complement(53806..54112)
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT conflict 54011
FT /note="Z15026"
FT /replace="t"
FT /citation=[3]
FT conflict 54459
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT conflict 54503..54505
FT /note="Z15026"
FT /replace="ccc"
FT /citation=[3]
FT conflict 54516^54517
FT /note="Z15026"
FT /replace="g"
FT /citation=[3]
FT conflict 54559
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT conflict 54562
FT /note="Z15026"
FT /replace="g"
FT /citation=[3]
FT conflict 54565
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT conflict 54575
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT conflict 54577
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT conflict 54582
FT /note="Z15026"
FT /replace="c"
FT /citation=[3]
FT repeat_region complement(54682..54988)
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT conflict 54836
FT /note="Z15026"
FT /replace="a"
FT /citation=[3]
FT conflict 54856
FT /note="Z15026"
FT /replace=""
FT /citation=[3]
FT repeat_region 55030..55332
FT /note="AluY"
FT /rpt_family="SINE/Alu"
FT mRNA 55463..56066
FT /note="gene fragment"
FT /gene="NB6"
FT repeat_region 55706..55992
FT /note="AluSq"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(56439..56731)
FT /note="AluSg"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(56734..57029)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region 57302..57481
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT conflict 57506
FT /note="AB000876"
FT /replace="a"
FT /citation=[14]
FT repeat_region 57568..57738
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT repeat_region 57785..57906
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT repeat_region complement(57907..58209)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
```

```
FT conflict 58054..58055
FT /note="AB000876"
FT /replace="cc"
FT /citation=[14]
FT conflict 59063
FT /note="AB000876"
FT /replace="g"
FT /citation=[14]
FT conflict 60286^60287
FT /note="AB000876"
FT /replace="c"
FT /citation=[14]
FT polyA_site complement(60780)
FT /gene="IKBL"
FT polyA_signal complement(60801..60805)
FT /gene="IKBL"
FT conflict 61431..61432
FT /replace="cg"
FT /citation=[10]
FT conflict 61646
FT /note="AB000876"
FT /replace=""
FT /citation=[14]
FT exon complement(61000..61589)
FT /number=4
FT /gene="IKBL"
FT exon complement(61762..61983)
FT /number=3
FT /gene="IKBL"
FT conflict 61940
FT /note="AB000876"
FT /replace="a"
FT /citation=[14]
FT repeat_region 62204..62504
FT /note="AluY"
FT /rpt_family="SINE/Alu"
FT conflict 62486
FT /note="AB000876"
FT /replace=""
FT /citation=[14]
FT conflict 62537
FT /note="AB000876"
FT /replace="a"
FT /citation=[14]
FT conflict 63375
FT /note="AB000876"
FT /replace=""
FT /citation=[14]
FT repeat_region 63653..63997
FT /note="AluSg1"
FT /rpt_family="SINE/Alu"
FT conflict 63701
FT /note="AB000876"
FT /replace=""
FT /citation=[14]
FT repeat_region 64000..64023
FT /note="simple repeat (GA)n"
FT conflict 64015
FT /note="Z15027"
FT /replace="g"
FT /citation=[3]
FT conflict 64028^64030
FT /note="Z15027"
FT /replace="g"
FT /citation=[3]
FT conflict 64049
FT /note="Z15027"
FT /replace="c"
FT /citation=[3]
FT conflict 64060
FT /note="Z15027"
FT /replace="a"
FT /citation=[3]
FT conflict 64061
FT /note="Z15027"
FT /replace="a"
FT /citation=[3]
FT repeat_region 64073..64375
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 64077^64078
FT /note="Z15027"
FT /replace="cg"
FT /citation=[3]
FT conflict 64079
FT /note="Z15027"
FT /replace="c"
FT /citation=[3]
FT conflict 64085^64086
FT /note="Z15027"
FT /replace="t"
FT /citation=[3]
FT conflict 64093
FT /note="Z15027"
FT /replace=""
FT /citation=[3]
FT conflict 64098
FT /note="Z15027"
FT /replace="t"
FT /citation=[3]
FT conflict 64138..64139
FT /note="Z15027"
FT /replace=""
FT /citation=[3]
FT conflict 64172
FT /note="Z15027"
FT /note="AB000876"
FT /replace="g"
FT /citation=[3]
FT /citation=[14]
FT conflict 64181
FT /note="Z15027"
FT /note="AB000876"
FT /replace="c"
FT /citation=[3]
FT /citation=[14]
FT conflict 64357^64358
FT /note="Z15027"
FT /replace="a"
FT /citation=[3]
FT conflict 65123^65124
FT /note="Z15027"
FT /note="AB000876"
FT /replace="g"
FT /citation=[3]
FT /citation=[14]
FT conflict 65129
FT /note="Z15027"
FT /replace="c"
FT /citation=[3]
FT repeat_region complement(65184..65486)
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT conflict 65184^65185
FT /note="Z15027"
FT /note="AB000876"
FT /replace="t"
FT /citation=[3]
FT /citation=[14]
FT conflict 65337
FT /note="Z15027"
FT /note="AB000876"
FT /replace="a"
FT /citation=[3]
FT /citation=[14]
FT conflict 65475
FT /note="Z15027"
FT /note="AB000876"
FT /replace="c"
FT /citation=[3]
FT /citation=[14]
FT conflict 65483
FT /note="Z15027"
FT /note="AB000876"
FT /replace="a"
FT /citation=[3]
FT /citation=[14]
FT repeat_region 65651..65948
FT /note="AluSg1"
FT /rpt_family="SINE/Alu"
FT conflict 66086
FT /note="Z15027"
FT /note="AB000876"
FT /replace="g"
FT /citation=[3]
FT /citation=[14]
FT repeat_region 66273..66571
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 66382
FT /note="Z15027"
FT /replace=""
FT /citation=[3]
FT conflict 66554
FT /note="AB000876"
FT /replace=""
FT /citation=[14]
FT repeat_region 66608..66918
FT /note="AluY"
FT /rpt_family="SINE/Alu"
FT conflict 66730
FT /note="Z15027"
FT /note="AB000876"
FT /replace="t"
FT /citation=[3]
FT /citation=[14]
FT conflict 66899
FT /note="Z15027"
FT /note="AB000876"
FT /replace=""
FT /citation=[3]
FT /citation=[14]
FT repeat_region 67025..67338
FT /note="AluY"
```

```
FT /rpt_family="SINE/Alu"
FT conflict 67038
FT /note="Z15027"
FT /replace=""
FT /citation=[3]
FT conflict 67172
FT /note="Z15027"
FT /replace=""
FT /citation=[3]
FT conflict 67321..67325
FT /note="Z15027"
FT /note="AB000876"
FT /replace=""
FT /citation=[3]
FT /citation=[14]
FT repeat_region 67383..67514
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT repeat_region 67515..67816
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 67613
FT /note="Z15027"
FT /note="AB000876"
FT /replace="g"
FT /citation=[3]
FT /citation=[14]
FT conflict 67735
FT /note="Z15027"
FT /note="AB000876"
FT /replace="t"
FT /citation=[3]
FT /citation=[14]
FT repeat_region 67836..68018
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 67993
FT /note="Z15027"
FT /note="AB000876"
FT /replace="t"
FT /citation=[3]
FT /citation=[14]
FT repeat_region 68180..68254
FT /note="MIR"
FT /rpt_family="SINE/MIR"
FT repeat_region 68290..68586
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region 68626..68924
FT /note="AluSp"
FT /rpt_family="SINE/Alu"
FT conflict 69034
FT /note="Z15027"
FT /note="AB000876"
FT /replace="t"
FT /citation=[3]
FT /citation=[14]
FT conflict 69219
FT /note="Z15027"
FT /note="AB000876"
FT /replace="c"
FT /citation=[3]
FT /citation=[14]
FT repeat_region complement(69395..69691)
FT /note="AluSg"
FT /rpt_family="SINE/Alu"
FT conflict 69485
FT /note="Z15027"
FT /note="AB000876"
FT /replace="g"
FT /citation=[3]
FT /citation=[14]
FT conflict 69692
FT /note="Z15027"
FT /replace="c"
FT /citation=[3]
FT conflict 69715^69716
FT /note="Z15027"
FT /replace="a"
FT /citation=[3]
FT repeat_region 70060..70362
FT /note="AluSq"
FT /rpt_family="SINE/Alu"
FT repeat_region complement(70380..70461)
FT /note="MIR"
FT /rpt_family="SINE/MIR"
FT repeat_region 70527..70742
FT /note="MIR"
FT /rpt_family="SINE/MIR"
FT repeat_region complement(70743..70876)
FT /note="MIR"
FT /rpt_family="SINE/MIR"
FT exon complement(71172..71448)
FT /number=2
FT /gene="IKBL"
FT exon complement(71860..71985)
FT /number=1
FT /gene="IKBL"
FT CDS join(complement(71860..71916),complement(71172..71448),
FT complement(61762..61983),complement(61000..61589))
FT /db_xref="PID:e1334709"
FT /gene="IKBL"
FT /product="I Kappa B-like protein"
FT /protein_id="CAA75072.1"
FT /translation="MSNPSPQVPEEEASTSVCRPKSSMASTSRRQRRERRFRRYLSAGR
FT LVRAQALLQRHPGLDVDAGQPPPLHRACARHDAPALCLLLRLGADPAHQDRHGDTALHA
FT AARQGPDAYTDFFLPLLSRCPSAMGIKNKDGETPGQILGWGPPWDSAEEEEEDDASKER
FT EWRQKLQGELEDEWQEVMGRFEGDASHETQEPESFSAWSDRLAREHAQKCQQQQREAEG
FT SCRPPRAEGSSQSWRQQEEEQRLFRERARAKEEELRESRARRAQEALGDREPKPTRAGP
FT REEHPRGAGRGSLWRFGDVPWPCPGGGDPEAMAAALVARGPPLEEQGALRRYLRVQQVR
FT WHPDRFLQRFRSQIETWELGRVMGAVTALSQALNRHAEALK"
FT conflict 71958
FT /note="Z15027"
FT /replace="c"
FT /citation=[3]
FT conflict 71961
FT /note="Z15027"
FT /replace="a"
FT /citation=[3]
FT conflict 72407
FT /note="Z15027"
FT /note="AB000876"
FT /replace=""
FT /citation=[3]
FT /citation=[14]
FT conflict 72824^72825
FT /note="Z15027"
FT /note="AB000876"
FT /replace="g"
FT /citation=[3]
FT /citation=[14]
FT conflict 72871
FT /note="Z15027"
FT /replace="g"
FT /citation=[3]
FT conflict 72901
FT /note="Z15027"
FT /replace="a"
FT /citation=[3]
FT exon 73034..73132
FT /number=1
FT /gene="ATP6G"
FT CDS join(73051..73132,73401..73501,74029..74202)
FT /db_xref="PID:e1334710"
FT /db_xref="SPTREMBL:O95670"
FT /gene="ATP6G"
FT /product="V-ATPase G-subunit like protein"
FT /protein_id="CAA75073.1"
FT /translation="MASQSQGIQQLLQAEKRAAEKVADARKRKARRLKQAKEEAQMEVE
FT QYRREREHEFQSKQQAAMGSQGNLSAEVEQATRRQVQGMQSSQQRNRERVLAQLLGMVC
FT DVRPQVHPNYRISA"
FT conflict 73145
FT /note="Z15027"
FT /replace=""
FT /citation=[3]
FT conflict 73161^73162
FT /note="Z15027"
FT /replace="t"
FT /citation=[3]
FT conflict 73190
FT /note="Z15027"
FT /replace="a"
FT /citation=[3]
FT conflict 73224..73227
FT /note="Z15027"
FT /note="AB000876"
FT /replace=""
FT /citation=[3]
FT /citation=[14]
FT exon 73401..73501
FT /number=2
FT /gene="ATP6G"
FT conflict 73523
FT /note="Z15027"
FT /note="AB000876"
FT /replace="a"
FT /citation=[3]
FT /citation=[14]
FT conflict 73828
FT /note="Z15027"
FT /note="AB000876"
FT /replace="t"
FT /citation=[3]
FT /citation=[14]
FT exon 74029..74372
FT /number=3
FT /gene="ATP6G"
FT conflict 74153
FT /note="Z15027"
FT /replace=""
FT /citation=[3]
FT repeat_region 74507..74685
FT /note="MIR"
FT /rpt_family="SINE/MIR"
```

```
FT conflict 74927
FT /note="Z15027"
FT /replace="g"
FT /citation=[3]
FT conflict 74929
FT /note="Z15027"
FT /replace="a"
FT /citation=[3]
FT conflict 74950^74951
FT /note="Z15027"
FT /replace="t"
FT /citation=[3]
FT repeat_region complement(75202..75502)
FT /note="AluSp"
FT /rpt_family="SINE/Alu"
FT conflict 75530
FT /note="AB000876"
FT /replace="t"
FT /citation=[14]
FT repeat_region complement(75746..76011)
FT /note="AluJo"
FT /rpt_family="SINE/Alu"
FT repeat_region 76016..76171
FT /note="LINE2"
FT /rpt_family="LINE/L2"
FT conflict 76529
FT /note="AB000876"
FT /replace="g"
FT /citation=[14]
FT repeat_region 76648..76924
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT misc_feature 77489..77768
FT /note="CpG island"
FT conflict 77955
FT /note="AB000876"
FT /replace="t"
FT /citation=[14]
FT conflict 78015
FT /note="AB000876"
FT /replace="t"
FT /citation=[14]
FT conflict 78032
FT /note="AB000876"
FT /replace="t"
FT /citation=[14]
FT conflict 78626
FT /note="AB000876"
FT /replace="a"
FT /citation=[14]
FT conflict 78925
FT /note="AB000876"
FT /replace=""
FT /citation=[14]
FT exon 78954..79288
FT /number=1
FT /gene="BAT1"
FT CDS join(79078..79288,80336..80463,80755..>80847)
FT /db_xref="PID:e1334711"
FT /gene="BAT1"
FT /protein_id="CAA75074.1"
FT /translation="MAENDVDNELLDYEDDEVETAAGGDGAEAPAKKDVKGSYVSIHSS
FT GFRDFLLKPELLRAIVDCGFEHPSEVQHECIPQAILGMDVLCQAKSGMGKTAVFVLATL
FT QQLEPVTGQVSVLVMCHTRELAFQISKEYERFSKYMPNVK"
FT repeat_region 79402..79606
FT /note="AluJ"
FT /rpt_family="SINE/Alu"
FT conflict 79504
FT /note="AB000876"
FT /replace="t"
FT /citation=[14]
FT conflict 79623..79624
FT /note="AB000876"
FT /replace=""
FT /citation=[14]
FT exon 80336..80463
FT /number=2
FT /gene="BAT1"
FT conflict 80533
FT /note="AB000876"
FT /replace="a"
FT /citation=[14]
FT conflict 80586
FT /note="AB000876"
FT /replace="t"
FT /citation=[14]
FT conflict 80643
FT /note="AB000876"
FT /replace="g"
FT /citation=[14]
FT conflict 80739
FT /note="AB000876"
FT /replace="a"
FT /citation=[14]
FT exon 80755..80847
FT /number=3
FT /gene="BAT1"
FT conflict 80763
FT /note="AB000876"
FT /replace="g"
FT /citation=[14]
FT repeat_region 81161..81453
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 81387
FT /note="AB000876"
FT /replace="a"
FT /citation=[14]
FT conflict 81500
FT /note="AB000876"
FT /replace="g"
FT /citation=[14]
FT repeat_region complement(81557..81799)
FT /note="AluSx"
FT /rpt_family="SINE/Alu"
FT conflict 81597
FT /note="AB000876"
FT /replace="t"
FT /citation=[14]
XX
SQ Sequence 81800 BP; 19886 A; 19955 C; 20623 G; 21336 T; 0
other;
Y14768  Length: 81800 September 24, 1999 21:14 Type: N Check:
6236

1    GATCCCGCGG GATCACGAGA TCAGGAGTTC AAGACCAGCC TGGCCAACAT
51   GATGAAACCC CATTTCTACT AAAAATACAA AAATTTGCCC TACGTGGTGG
101  CGGGCGCCTG TAATCCCAAC TACTCGGGAA GCTGAGGCAG AGAATTCCTT
151  GAACCAGGGA GGCGGAGGTT GCACTCCAGC CTGGGTGACA GAGCAAGACT
201  ACGTCTCAAA AAAAAAAAAA AAAAAAAAAA AAAAAGAAGG ATGGGCGCGG
251  TGGCTCACGC CTGTAATCCC AGCACTTTGG GAGGCTCAGG CGGGCGGATC
301  ACCTGAGGTC ACGAGTTCGA GACCAGCCTG ACCAACATGG AGAAACCCCG
351  TCTCTACTAA AAATACAATT ACCCAGGCGT GGTGGCGCAT GCCTATGATC
401  CCAGCTACTC GGGAGGCTGA GTCGGGAGAA TCGCTTGAAC CTGGGAGGCG
451  GAGGTTGTGG TGAGCCAAGA TCGTGCCATT GCACTCCAGC CTGGGCAACA
501  AGAGCGAAAC TCTATCTCAA AAAAAAAAAA AAAAGAAAGA AGTAGAGAAA
551  TAAAATATGA TTATGCCTGG GGTGATTGAG AAATGTTTCT AGAAAAACTC
601  TTATTAAGAG ATATGAGGCC GGGCGTGGTG GCTTATGCCT GTAGTCCTGG
651  CACTTTGGGA GGCTGAGGCG GGCAGATCAC AAGGTCAAGA AATGGAGACC
701  ATCCTGGCCA ACATGGCAAA ACCCCGTCTC TACTAAAAAT ACAAAAATTA
751  GCTGGACTGG TGGCGCGTGC CTGTAGTCTC AGCTACTTGG GAGGCTGAGG
801  CAGGAGAATC GCTTGAACCT GGGAGACGGA GGTTGCAGTG AGCTGAGATC
851  ACGCCACTGC ACTCCAGCCT AGCGACAGAG TGAGACTACG TCTCAAAAAA
901  AAAAAAAAAA GAAAAAGAAG AAAAAAAGGG CCGGGAGCGG GAGCAGTGGC
951  TCACGCCTGT AATCCCAGCA CTTTGAGAGG CCAAGGTGGG TGGGTCACCT
1001 GAGGTCAGGA GTTGGAGACC AGCCTGGCCA ACATGAGACT CCATCTCAGG
1051 AAAAAAAAAA AAAAATGATA TGAAGAATGA TCGGAGTAGA TATCTGTCAA
1101 AAATATATAA ATAAATAGAT ATGAAGAACG AGTTAGAACT TACCCAAGTT
1151 TCAGAAGAAG GTGTACATCT CAGGTGCAGA GAACAGATTA GTGGAGGCAT
1201 GAGTTTGGGG GCCTGTGCCT AGGCTGATAT GGCTAAATAA AGGAGGCGGC
1251 TTATTCATTA TGAAAAGTGA AGGCAGAAAA GTTTGCGGAA ATCAGATGAG
1301 AATTTTAGTC AGAGGCTCAG TAAGATTTAA GATGACTTCT GTTTTCAAAC
1351 TTCCAAAATG ATTCTCTTCT CCCTTCTCCC CGGCCGGCTA CCACACATAC
1401 CGTTTTGTTA CCCTTTGCAG CAAAATTCTT AAAAGGGATG TCCATAGTCA
1451 TTGTCTCCAA TGAGTTTCTC CTTCCAGTCT TTTTTTTTTT TTTTTTTTTT
1501 TTTTGAGACA GAGTCTCACT CTTGTCACCC AGGCTGGAGT GCCATGGCGT
1551 GATCTCGGCT CACCGCAACC TCCACCTCCC GGGTTCAAGC GATTCTCCTG
1601 CCTCAGTCTC CCGAGTAGCT GGGATTACAG GCACCTGCCA CCATGCCCAG
1651 CTAATTTTTT GTATTTTAGT AGAGATGGGG TTTCACCATG TTGGCCAGGC
1701 TGGTCTCCAA CTCCTGATTT CAGGTAATCT GCCCGCTTCG GCCTCCTGAA
1751 GTGCTGGGAT TACAGGCGTG AGCCACCACG CCCAGCTGTA AATTCTTTGC
1801 TTTATTTAAT AGCACCAGTT CTTTCATTCA TCTACCCATT CAAGTGTTTT
1851 TTGTTTGTTT GTATGTTCGG TTTTTTTTTT TTGAGACGGA GTTCGCTTTT
1901 TAGTTGCCCA GGCTGGAGTG CAATGATGCG ATCTCGGATC ACCACAACCT
1951 CCACCTCCTG GGTTCAAGAG ATTCTCCTGC CTCAGCTTCC CAAGTAGCTG
2001 GGATTACAGG CATGCGCCAC CACACCAGGC TAATTTTGTA TTTTTAGTAG
2051 AGACGGGGTT TCTTCATGTT CGTCAGGCTG GTCTCGAACT CCTGACCTCA
2101 GGTGATCTGC CCACCCCATC CTCCCAAAGT GCTGGGATTA CAGGCATAAG
2151 CCATCGGCCC CGGCTCATTC AAGTTTTAAA CATTATTGAG GGCTTATGAT
2201 ATGTCAGGCA CAATAGCATA GGTCACTTTA TCCCCATTTT TAGAAAAAAG
2251 CCTCACAGAG GCTGAGTGAT TTGTCAAAGT GACAAAGCTT ATAGTGGTGG
2301 AGAATGTTGG TTCATGGTAT TCCTTGCTCT TCATGAAATC TGTCATGGGC
2351 CTGAGTACCT TGTTTGGACA TGTGTTTTTG ACCTGCCTGC ATTCTTTGAG
2401 GTTAAGGCTG TCTTACACAG CTGGAATCCC CCTTGGTGCG TGAAACCATC
2451 TGCTGTAGTG GACACACTAT AGCGTTGTTG ATGGATTGTT CATATGGCCA
2501 AATCCTTGCC ACCTTTCCCT CAAGACCTGT CTCTCCTTCC CAAACCCAGG
2551 AAAAAAATCA CCCAAATAAA AACTAAGCTG ACTTGATCTG GAGGAGGGGG
2601 TAATTTATTT AATTTGTTGA CTCTGGCTTA CAATTTTGTC TTCTGTTTCC
2651 ATATCTGGGT CATTAGAAGC CCCTTCAATC CCATCATCCC TTTTCCCTTC
2701 AAATCAGGGC AACTCAGAGA TAGCTTTCTT GGCTGGGGGG CCTGTTGGCT
2751 TTTCCTTTTC TCTCGCTTTT TCCTCATACA TCAGGATCCT GGTTGAGATG
2801 GGAAGAGGCA CAGGGATCAG GGAACACTCC TCTCAAGGGA ATAAATCATA
2851 GGGTGCTAGG AAAGGCTAGA AGGGGATGGG GATGTGAGGG TGGAAGGGGG
2901 AAGAGGAAGG GCAGGGGTTG GAGTTTCTGG GGACCAGGAC TATGGATAGG
2951 GATGGGGTCG GGACCTTGGA AGGGTCCCTC TAAGGATGGA AGAAGGGATG
3001 ATGGGCAACG TGTAGAAATG TTGAGCCTGT GGACAAGGGT AGGGATGGGG
3051 GAGGAGAAAA CAGGTAAGTA CAGGGGAGGT TGGAAATTGA CACTCACATT
3101 TTTAGGATGG CAGATCTCTT GCCCAGCATC ATCCTGAGAA AGTCAGGGTA
3151 GCTGAACGTC TCCCCGGAGC CACTGGACAC CTCTCCAATT AATTTCTTTA
3201 GCTCTAGGTG AGTCTTGGGG ACTCCAAGTT TCTCCAGCAT TCGTTTCAGG
3251 GACATGATAT CTGGGGGCAG AGGATGGGGA TGGGGAGAAG GTGGGGAGAC
3301 CCTCTCTCTC CTCTCCCACT TCCGCATCCC CATTCTCCTT CCTCCCACTG
3351 TGGTAGGTGG GCCCCCTCCC AAATCATGAG GAATGGAGCA TGTAGGAGAG
3401 ACCTCTGTCT TCTTAGGCCT GCACACCACC CTGCCCCCGC AAATCACCCG
```

```
3451 TTTCTCACCA ATATCGCCAT TTCCATTAAG GTCAAACTCC ATGTATTTCT
3501 CTGGGTGAGG GGAGAGAGCA GAAGGGATAA GCGCTGACTG GAGGGTTTTC
3551 AGCAGGGGAC TGAGGAAGGG GTTGCCAACC AACTGCTAGG TGGAGGAGAC
3601 AACGTGGAAG GAGGAGGAGG GGCAGGCAGA GGAAGTGGGG TTGGGGAGTT
3651 GAGGGGCCCC GCTAGGGTGT AAGGAGCTAG AGGGGATGCT GAGAGAGGCA
3701 GGCCCGGGAA AGGGACCGCA GTGTGGCAGG AGGGCAATAA AGGCAGGTTC
3751 CCCTTTTGCT ACCACAAAAG TAAACTTTTT ATCTCTGCTG TTCCTGCCTT
3801 GTTCTTGTGG AAATCCTAAC CCTCCTACCC CCAGCCCTGT CTCCACCGCC
3851 TACAGTTTCC CCCTCACCTT TGAAGCCTTC CAGTTTGGAG GGCAGATCCT
3901 CATCACTGCT ATATTTGGGA TCGTCTAGGA ATTGCTATGG AGGGGAAAAT
3951 AAGAGCCAGG GTAGGCCCTC CCCATCCACG CCCCATCCCC CTGGCTCTGC
4001 CCCCTTAGTT CTTCCTTCTA CCTTGTTGAT CTCATCCAGC CTCTCTTCCT
4051 GCTGGGCCTT CAGCAGTCCG AAAGCTTTTC CTCCTGTGGG GTGAAAAAAA
4101 TTAATCAGCC CTCATCCCTC CCTTCCTCTC CTTACCCTGA CACCAGCCCG
4151 GGCTACCCTA GGCCAGCTGA CTGCCCTCCT AGCCCACAGC CCAGGTCTGT
4201 CCATCTCATT CCTCCGATCT AATGCCGCTG CCTATCCCAC CCTCCCTACC
4251 CTGTAAATCC CTGGTTTGGC TCATAGCTCA GCAGATGCTG GTAGAGGTGG
4301 GGAGACAGAC CAAGCTGGAG GCCTCTGTCT GCAGGCTCTC CTTCTCTCCC
4351 ACCAGTCTTC TCAGAAGCCT TCCTCCCAAA CTTCCCCTTT CCCCTCATTC
4401 CTTCCTCCTC TGCCTCCCAG ACCTCACTTC CCCATCAGGT TCTGCTGACA
4451 CTAGCACCCC AGGAAGGAGG GGGAGCCTGG TGTGAAATGC AGATTAGAAA
4501 ATAACATTCT CCCAAATTCC TCTTAAATGA GAGCCCAGTG TGCGCATGAA
4551 ACAAATATTT TGATAAGCCA CTGGAGGTGT GCAGGAAGGG CAGGAAGACA
4601 CCAGGGATAT TTATTCTCAG GAACTAGTAT TTTTGAACAG GTCCAAGGTT
4651 CAGGTAATGA TAAGACCTTG CATTTTGTCA GGTTTCATGA GTCTGGGCAG
4701 GCAGCTGTGG ATAGGAAGAC AAGGAAGGGC TGGCTGAATT TTTCTGCTCT
4751 GTAGATCGGA GAAGAATCTG AAGTCTGCTG GCAGCTCCAG GAGGGTGACA
4801 AGATGGGCAC AGGGTTTAGA AAGGGAGAGA AACCTCACTC CCAGCCTGGC
4851 AGCTCCAAGG AGGCCCTTAA AAATATGGAA ATGTCTAACA AAACATTTTA
4901 ATTTGTTTGC TTTAGAATGT GCAACGTAAC CCTCTCATCT TATCTCCTTT
4951 GAAATCTTTA GTAAAATCTA TAGAGATCAG GAATTTAAGA GTTTAAGTAG
5001 ATTCTTTGGA GGAACTGAAA GTGGTACTTA AAGAAACAGC GTTTGTGGAC
5051 AGAAGAGCCA GGGAGTGAGG CTGACATCAG GTTTGCACAC TGCACGTGCG
5101 TGCACACATG CAGGCATGTT TGTGTGATGG TGGTGCTGGT CACTGTGGGT
5151 TCAACAGAGT GGATGATAGC CTCTGACACA TGCCTCCCTA GGCATAACCT
5201 TTATCCCCTC TGTCACTCTG TTTGGCCAGT CGCTTCTTGG AGAGGTGAGG
5251 ACAGCACATT TCCCTCATTG AGATTTCCTG GTCCAAACTC CCTGCCCAGG
5301 GCTCCATCTA CCCTGCCCCG GCATTGATGA AGACACCTTT CCTTCTCCCT
5351 TCCCCCCAAC ATATTTCCTT TTCTGTAGCT GGGAGTTTTC ACTTCTCACT
5401 GCCCTCTCCC TGACCCAAGC CAGGTAGGGG AGATGCTGGA GTGGCCTCTG
5451 GCTGCTGCTG CTGCTTTTTT TTTTTTTTTT GAGTCGGAGT CTTGCTCTGT
5501 CGCCCAGGCG GGAGTGCAAT GGTGCAATCT CGGCTCACTG CAACCACTGC
5551 CTCCTGGGTT CAAGCAATTC TCCTACCTCA GCCTCCTGAG TAGCTGGGAC
5601 TACAGGCGCC CACCACCACG CCCAGCTAAT TTTTGTATTT TTAGTACAGA
5651 TGGGGTTTCG CCATGTTGGC CAGGATGGTC TCTATCTCTT GACCTTGTGA
5701 TCCACCCACC TCAGCCTCCC AAAGTGCTGG GATTACAGGC ATGAGCCACT
5751 GCACCTGGCT GGACTCTGGC TTCTTGACTT CCCTTTATTG CAGCTACATT
5801 CTCCTGCTGC AGAGAGAGCT GCAACCCTGT ACTCCCAATG TTCTCAGTGC
5851 CTGCTGGGAG GAACTCACTT CTTTATTCAT ACGTTCACTC AACAGATATT
5901 TACTGAATGC TTACTATATG CCAGACATGC CTCTAGGCCC AGAGAACCCA
5951 ACAAATGAAA ACTCTTGTCT TCATGGTTGA TCTAATTGGG GAAGGCAGAC
6001 AATCACAAAT AAATAAGTAT ATTAAAGGTG CTAGGTGGCC AGGCGCAGTG
6051 GCTCACGCCT GTAATCACAG CACTTTGGGA GGCTGGGGCA GGCGGATCAC
6101 AAGGTCAGGA GTTGGAGACC AGCCTGACCA ACATGGTGAA ACCCCATCTC
6151 TACTAAAAAT ACAAAAATTA GCAGGGCCTG GTGGTGCGTG CCTGTAATCC
6201 CAGCTGCTCA GGAGGCTGAC GCAGGCGAAT CGCTTAAACC CGGGAGGCAG
6251 AGGTTGCAGT GAGCTGAGAT CGCGCCATTG CACTCCAGCC TGGGCAACAG
6301 AGTGAGACTC CGTCTCAAAA AAAAAAAAAA AAAAAAAAAA AAGGCGCCAG
6351 GTGCCTATGG AGGAAAACAA AGTAGGGAAG AGGGGACATA GGAGTACTCT
6401 ACTCTGCCAA TGGTCCCAGT TTTTTGTTTT TTGTTTTTTG TTTTTTTTTT
6451 AAACGGAGTC TCACTCTGTC GCCCAGGCTA GAGTGCAATG GCTTGATCTC
6501 AGCTCACTGT AACCTCCGCC TCCTGGGTTC AAGCGATTCT CCTGCCTCAG
6551 CCTTCCGAGT AGCTGGGATT ACAGCCACCT GCCATCATAC CTAGCTAATT
6601 TTTGTATTTT TGTAGAGGCG GGGTTTCACC ATGTTGACCA GGCTGGTCTT
6651 GAACTCCTGA CCTTAGGTGA TCTACCCACC TCGGACTCCC AAAGTGCTAG
6701 GATTACAGGC GTGAGCCACT GCACCCAGCT TTTTTTGTTG AGACAGAATC
6751 TCGCTCTGTT GCCCAGGCTG GAGTGCAGTG GCACAATCTC AGCTCACTGC
6801 AACCTCTGCC TCCCGGGTTC AAGCGATTCT CCTGCCTCAG CCTAACGAAT
6851 AGCTGGGATT ACAGGCGATC ACCACTATAC CCGGCTAATT TTTGTATTTT
6901 TAGTAGAGAT GGGGTTTCAC CAAGTTGGCC AGGCTGGTCT TGAACTCCTA
6951 ACCTTGTGAT CCTCCCACCT TGGCCTCCCA AAGTGCTGGG ATTACAGGTG
7001 TGAGCCTGAC TTGTCCCCAG TTTTAAATAT GGTGGTCTGG CCAGACACGG
7051 TGGCTTACGC CTGTGGTCCT AGCCACTTGG GAGGCTGAAG TGAGAGGATC
7101 GCTTGAGCCC AGGAGATCAA GGCTGCAGTG AGCCATGACC GTGCCATTAC
7151 ACTCCAGTCT GGGCAACAGA GTGAGACCCT GTCTCAAAAA ATACAATAAA
7201 ATAATAAAAT AAATAGAGTG GTTAGAGAAG CGTTCTCTGA AAAAGTGATA
7251 TTTGAGCAAA GATTTCTTTT CTTTTTTTTG AGATGGAGTC TCACTCTGTC
7301 ACTCAGGCTA GAGTGCAGTG GTGCAATCTC CACTCACTGC AACCTCCACC
7351 TCCCGGGTTC AAGTGATTCT CCTGCCTCAG CCTCCAGAGT AGCTGGGATT
7401 ACAGGCACCT GCCACCACGC CCAGCTGGCT AATTTTTTGT ATTTTTAGTG
7451 AGACGGGGTT TTACTATGTT GGCCAGGCTG GTCTTGAACT CCTGATCTCA
7501 TGATTCGCCC ACCTCGCCCT CCCAAAGTGC TGGGATTACA GGTGTGAGCC
7551 ACCACGCCTG GCCGTTTGAT CAAAGATTTT AAGCAAGCAA AGGAACAAGC
7601 CTCTGTCTTC CTAGGCCTGC ACACTTCCCC AAATCACCTT TTTTCACTGA
7651 TATCTCTGTT TCCATTCTGG TCAAACTTTA TGTATTTCTT TGGGTAAAGA
7701 GAGAGAACAG ACAGGGTAAG CACTGGCTAG AGGGTCTTTG GCAGAGGATC
7751 GAGGAAGGCA TTGCCAACCA AGTGCTAGGT TGAGGAGAGG ATGTGGAAGG
7801 AGGAGATGGG GCAGGCAGAG TGGGCTTGGG GAGTTGAGGG ACCATGTAAG
7851 GTGGATATAC GGCTAGAAAG AAAAATACTT ATTTTTTTTT TTTTGAGACA
7901 GAGTCTCCAT CTGTCGCTAG ACTGGAGTGC AGTGGCGCGA TCTTGGCTCA
7951 CTTCAACCTC CACATCCTGG GTTCAAGCGA CTCTCCTGCC TCAGCCTCCC
8001 AAGTAGCTGG GACTACAGGT GCGCACCACC ATGTTCAACT AATTTTTGTA
8051 ATTTTAGTGG AGACGAGGTT TCACCATGTT GGCCAAGATG GTCTCGATCT
8101 CTTGACCTTG TGATCCACCC ACCTTGGCCT CCCAAAGTAC TGAGATTATA
8151 GATGTGAGCC ACTGCGCCTA GCCAGAAAAA TATTTATTGA TCCTAAAGTG
8201 GAGTCAGCGC CACTGTTGCT GCTAGTGTTG AATCCAGAAC CATAGCTGGA
8251 CATGGGATCA GAGGATGAAC AAGAGATGCT GACTGGGCCT AGAGATGGCA
8301 AGGAGGAGGA GGAAGAGGAA GAGGAGGAAT TAGTCAATTC CTAACAAAAG
8351 CGAGAGAGCA GTGCGTGCAG CTGAAGAAAG GAGTAAAGTC TCGGAGCAGC
8401 TGGAGCTCTG TGATGAGCAC GTGTCTTCCA GATCACAGAC AGGAGAGGAT
8451 TGCACAAAGC TCTTTGACTT CTCTAATGCA AGGGACCATT GCATGGCCCA
8501 CAAATCCTTT AACAGTTTGA AATGAATGTG CAGACTCCTT CACCCCAGCT
8551 TTCATCACCT GGGCATTTCT CTATGGTTTT AAATATAACA TTTGTTTCTA
8601 ATTTGTGTAA CTATAAGTCC ATGTGAACCT CATGGATTTT TGGCTTAGGC
8651 AAATAGCTTC TCTGTAATTG GCAGCAATTC CATCTTAATA AGGGAACTGT
8701 GATTTGCAAA ACGAAAAAAT ATACTGTTTT TTTCTTTTTC CTTTTTTTTA
8751 TTCAAGGGCT AGTGAATCTT GGGATCTCAG GGTTACATCG AGGTATATTT
8801 ATCATCATCA TTATAGTCCA GGACTAAAAG GAAGTTATTC CAGGCTCTTT
8851 GCAAAGAATC TCAGTTTTTT TTGTTTGTTT GTTTTATTTT GTTTTGAGAT
8901 GGAGTCTTGC TCTGTTGGCC AGGCTGGAGT GCAGTGGTGC AGTCTCTGCT
8951 CACTGCAACC TCCACCTCCC AGGTTCAAGT GATTCTCCTG CCTCAGCCTC
9001 CCGAGTAGCT GGGACTACAG GCACCCGCCA CCATGCCTGG CTAATTTTTT
9051 GTATTTTTAG TAGAGACAGG GTTTCACCAT GTTAGCCAGG ATGGTCTCAA
9101 TCTCCTGACC TCATGATCGG CCCGCCTCGG CCTCCCAAAG TGCTGGGATT
9151 ACAGGCGTGA GCCACTGCGC CTGGCCAAGA ACACCACTTC TTACACCAAT
9201 CATGTGGAGA CTTCTCCCAC ACCAATCAAT TCTTCAGTTC TTGGTAGATA
9251 TTAACTGGGA GTCCTGCAGT TTAATTCAGT TCTGACACTC TACCCATAAT
9301 TACCAGTTAT CTAATCCCCA AGTTGAAGGG CTCAGTTCTA CAAGGCTGCC
9351 CCCACTCCAG ATACCAATCA CAAACAAGTC CAGGCCTCCT AACTGTTTTT
9401 ATGACTCCCT CTTCTGGCTC AGTAATTCGC TGGAATAGCA CGCAGAACTC
9451 AGGTAAACAC TTTACTTATG TTAGCTAGTT TATTATGAAG GATAAAACTT
9501 AGGAACAGTC AGATGGAAGA GTTGCTCCGG ACAAAGTATC GGGGAAGGGG
9551 GATTTGGAGT TTTCTTGTCC TCTCTGGGCA CCTCCCTGCA CCTCCATGTG
9601 TTCATCAACC TGGCAGCTCT CCAACCCTCA TCTTTTTTGT TTTTTTTTAT
9651 GGAGGCTTCA TTATGCAGGC ATACTTGATT CAATAGCCAT TAGTTGGAGG
9701 TCAGAGGGTT GGAGCTAAAA GTTCTAACTC TCTAATCACA CAGTTGCCTA
9751 CCTTGGCAAC AAACCCCAAC CTTAGGGGCT TTCTAAAAGT CAACTTACTA
9801 ACATAAACTC AGTTGTGGTT GAAAAGAGCT TGTTATGAAT AACAAAAGCT
9851 CCCTTCACTT TATCATTTAG GAAATTACAA GGGCTTTAGG AGGTCTGGCC
9901 AGGAGCCAGG ATGGAGACCA AAACACACAC ACACACACAC ACACACACAC
9951 GCACGCACAC ACACATTAAT ATTTTGAGAC AGGGAGACAG GGTCTCGTTC
10001 TGTTATCCAG GCTGGAGTGC AGTGGCACAA ACATGGCTAC TGCAGCTTTG
10051 ACCTCCCAGG CTCAAGCGAT CCTCCCTTGT CAGCCTCCTG AGTAACTGGG
10101 ACTACAGGCA TGTGCCACCA CACCTGGCTA ATTTTTAAAA TTTTTTTTAG
10151 AGACAGGGTC TCCTTGTGTT GCCCAGGCTG GTCTTGACTG CCTGGGCTCA
10201 AGTAATCCTC CCACCTTGGC CTCCCAAAGT GCTGGGATTA CAGGCCTGAG
10251 CCACCGTGCC TGACCTTTTT TTTTTTTTTA TTTTTTTTTT TTTGAGACAG
10301 TTTCTTGTTC TGTCATCCAG GTTGGAGTGC AGTGGCACCA TCATAGCTCA
10351 CTGCAGCCTC AACTTCTTGG GCTCAAGCAA TCCTTCCATC TCAGCCTCCC
10401 AAGCAGCTGG GACCCGCACC ACCATGCCTG GCTCATTTTA AAATTTTTTG
10451 TAGAGATGAA GTCTCGCCAG GTTGCCCAGG CTGGTCTTGA ACTCCTGGGC
10501 TCAAGTGATC CTCTGGTCTC AGCCTTCTAG AGTGCTGGGA TTACATGTGT
10551 TAACCACCAC ACCCAACCCA TATATATACA TATTTATTTA TTTTTTCCCC
10601 GAGATGGAGT CTTGCTCTGT TGCCCAGGTT AGAGTGCAGT GGCGCGATCT
10651 CCGCTCACTG CAAGCTCCAC CTCCCAGGTT CACGCCATTC TCCTGCCTCA
10701 GCCTCCTGAG TAGCTGGGAC TACAGGCGCC TGCCACCACG CCCGGCTAAT
10751 TTTTTGTGTT TTTAGTAGAG ACGGGGTTTC ACCATGTTAG CCAGGATGGT
10801 CTCGATCTCC TGACCTTGTG ATCTGACCGC CTTGGCCTCC CAAAGTGTTG
10851 GGATTACAGG TATGAGCCAC CGCGCCTGGC CTTTTTTTTT GTGGGGGGAT
10901 GGAGTCTCGC TCTGTCGCCC AGGCTGGAGT GCAGTGGCGA TCTCGGCTCA
10951 CTGCAAGCTC CGCCTCCTGG TTTCACGCCA TTCTCCTGCC TCAGCCTCCC
11001 GAGTAGCTGG GACTACAGGT GCCTGCCCCC ACACCCGGCT AATTTTTTTT
11051 TGTATTTTTA GTAGAGACGA GGTTTCACCA TGTTAGCCAG GATGGTCTTG
11101 ATCTCCTGAC CTCGTGATCT GCCCCCCTCG GCCTCCCAAA GTGCTGGGAT
11151 TACAGGTGTG AGCCACCACA CCCGGGCCCA AAAATATATA TTTTTATTAT
11201 AACAATATCA CAATGGGAGA AGAACATTTC AGGTGGAGGG AACAGCAAGT
11251 TCAAAGGCCC TGATGTGAGA GCAGTGTGGA TATCTGGAGT GTTCATGGAG
11301 TAGCAGGGAG GCCTGTGCTG CTGAAGCCAG GTGAGGAGGA GGTCAGCTTG
11351 GGTAGGGGCC TGTAAAGACT CTGGCTTTTA TTCTGAGTGA GACTGCCTCC
11401 CAGACTGGTC TTTCCATTGA GAATGGCTTG ACAGACTACT GGGGTGAGAC
11451 AGTGCTTAGA TGCTCAGAAA AACTGCCTGG GATAGAGAGA ATGACAGCTG
11501 CTCTGACTGT GGGACAAAAA TGAAGACTAG GCCTGGCGCG GTGGCTCACT
11551 CCTGTAATCC CAGCACTTTG GGAGGCCGAG GCAGGCAGAT CACCTGAGGT
11601 CAGGAGTTTG AGATCAGCCC GACAAACATG GAGAAACCCC GTTTCTACTA
11651 AATGTACAAA ATTAGCCAGG CATGGTGGCG CATGCCTGTA ATCCCAGCTA
11701 CTTGGGTGGC TGAGGCAGGA GAATCGCTTG AACCTGGGAG GTGGAGATTG
11751 CAGTGAGCCG AGATCACGCC ATTGCACTCC AGCCTGGGCA ATAAGCGAAA
11801 CTCCATCCCC CCCACCAAAA AAAGTGAATA CTGAAGGGTA TATAGGAAGA
11851 CAAGCGGGTG GAGAAGGAAA CTCAGGGCCT TCTGAGAAGG AAACAGAGAT
11901 GAGGGGAAAA GAATAGACAT GCAAGAAGAT GCTGGGGCTC AGAAAGTGAT
11951 ACCCCAAAGA CTGGTGCTTT GAAATGCTGA GAGGCCTTAG AAGCTGCCTT
12001 AGAATTGTCC TACCACACTT TCTGGAATTT CCTTATCTGA CTGAAAAAAC
12051 TTCTTTGCAA AAGAAGTGCA ATTGTCTTAA GACCTCCTGC CTAGAGATCT
12101 TATCAAATGA CCAGATCAAC CACCAGAGAG AAGATATTAG GAGTCATCAT
12151 CATACCCAGA TAGACTTTTC TTCTCTTTCT TTTCTTCCTT TTGAGTCAGC
12201 GTCTTGTTCT GTTGCCCAGC CTGGAGTGCA GTGGCATGGT CATGGCTCAC
12251 TGCAGCCTCA ACCTCCAGGC TCAAGTGATC CTCCCACCCT AGCCTCCTGA
12301 GTAGCTGGGA CGACAGGCAT GTGCCACCAC GCCTTAGGTT TTTTTTTTTT
12351 TTTTTTTTGA GATGGAGTCT CTCTCTGTCA CCCAGGCTGG AGTGCAGTGA
12401 CATGATCTCA GCTCACTGCA ACCTCCTCCT CCTGGGTTCA AGCAATTCTC
12451 CTGCCTCAGC CTCCGGAGTA GCTGGGGTTA CAGGCGCCCG CCACCACGCC
12501 CAGCTAAGTT TTGTATTTTT TAGGAGAGAA GGAGTTTCAT CATGTTGGCT
12551 CTGGCTGATC TCGAACTCCT GACCTCAGGT GATCCACCTG CCTTGCCCTC
12601 CTAAAGTGCT GGGATTACAG ACATGAGCCA CTGCACCCAG CCATGGTGAC
12651 TAATTATCTT TTTTTTTTTT TTTTTGAGAC AGAGTCTCAC ACTGTCGCCC
12701 GGGCTGGTGT GCAGTAGCGC GATCTCGGCT CACTGCAACC TCCACCTTCC
12751 GGATTCAAGT GATACCCCTG CCTCAGCCTC CCAAGTAGCT AGGATTGCAG
12801 GCACCCGCCA CCACGTCCAG CAATTTTTTT TTTTTTTTGA GACGGAGTCT
12851 CGCTCTGTCG CCCAGGCTGG TGTGCAGTGA CAGGATCTCA GCTTACTGCA
12901 AGCTCCACTT CCTGGGTTCA TGCCATTATT TTTTTGTATT TTTAGTAGAG
12951 ATGGGGTTTC ACTATGTTGC CCAGGCTGGT CTCAAACTCT TGACCTCGTG
13001 ATCTGCCCAC CTTGGCCTCC CAAAGTGCTG GGATTACAGG CGTGAGCCAC
13051 AGTGCTGGTC TTTTTTTTTT TTTTTTTTAG ATAGAGTCTT GCTTTGTCAC
13101 CCAGGCTGGA GTGCAGTGGC ATGATCTTGG CTCACTGCAA CCTCTGCCTC
```

```
13151 ATGGGTTCAA GAGATTCTCC CGCCTCAGCC TCCTGAGTAG CTGGGACTAC
13201 AGATGCGCGC CACCAAGCCC AGCTAATTTT TTTGTATGTT TAGTAGAGAC
13251 AGGGTTTTGC CATGTTGGCC AGGATGGTCT GGATTTCTTG ACCTCGTGAT
13301 CCGCCCTCCT CAGCCTCCCA AAGTGTCAGG ATTACAGGCG TAAGTCACCG
13351 TGCCTGGCTT CTTTTTTTTT TTTTTTTTAA ATATAGAGAC AATGGATCAT
13401 AATATAGAGA CATTCTCATC CTAATGTTAG CACAACGGGG ATGCAGATGG
13451 TCAGGATCTG GAGACTTGCA GGAATAGAGC AGTGAGCCAA AGTGCAGCCT
13501 GGTTTATGCT TAATTATAAC TATATTTCTG CCCAGATCCC AGCATGCAAT
13551 TCTTGTCAAT CTTGGGACAG TTTCAATTTA CTCTATTGAA AGTTTTTATT
13601 AATATTTTGA TAACTGGAAT TTGATTTCGA TGCAATTTTT TTTTTTTGAG
13651 ACGGAGTCTT GCTCTGTCAC CCAGACTGGA GTGCAATGGC GTGATCTCGG
13701 CTCACTGCAA CCTCCGCCTC CCGGTTCCAG CAATTCTCCT GCCTCAGCCT
13751 CCCAGGTAGT TAAGATTACA GGCGCCGACC ACCATGCCTA GCTAATTTTT
13801 TGTTTTTAGT AGAGACAGGG TTTTGTCATG TTGGCCAGGC TGCTCTCGAA
13851 CTCCTGACCT CGGGTGATCC ACCTGCCTCG GCCTCCCAAA GTGTTGGGAT
13901 TACAGGCGTG AGCCACTGCG CCTGGCCGAT TTGTTTTTGT ATGTATATTT
13951 TTTAAATTTT ATGCTTTACT AGAATGCCAA AGGGATCTGT GGCACAAAAT
14001 AACTAAGAAC CACTGCTGTA ACCTCTCTCT CCCCTCACCC CTTTGAGCCT
14051 GGGAGTGGAA ATATCCCCTG CAGTTCTCTT ACATAACACT CCCACCTTTG
14101 TGTGTGCGTG TGTTTTGTTT GTTTTGTTTT GTTTTTGGGA TGGAGTCTCG
14151 CTCTGTCACC CAGGCTGGAG TGCAGTGAAG TGATGTCAGC TCAGTGCAAC
14201 CTCCACCTCC CAGGTTCAAG CGATTCTCCT GCCTCAGCCT CCCTAGTAGC
14251 TGGGACTACA GATGCATGCC ACCACGCCTG GCTAATTTTT TGTATTTTCA
14301 GTAGAGATGG GGCTTCACCG TGTTAGCCAG GATAGCCTTG ACCTCCTGAC
14351 CTCGTGATCC ACCCTCCTCA GCCTCCCAAA GTGCTGCGAT TACAGGCATG
14401 AGCCACCGCA CCCGTTGTCT TTAATAATTT TTTTGCCTGC TTAATATATA
14451 GTTTCTGAGA GATGTGAGTC TTCCTCTAAT ATTGTAGGTT ACTTTTCTGT
14501 CGCTGCTGCA CAAATTTTTT TTTTTTTTTG AGATGGAGTC TCACTCTGTC
14551 GCCCAGACTG GAGTGCAGTG GCGTGATTTT GGCCCACTGC AACCTCTGCC
14601 TCCTGGGTTC AAGCGATTCT TCTCTTGAGT AGCTGGGATT ACAGGTGCAT
14651 GCTACCACGC CTGGCTACTT TTTGTATTTT TAGTAGAGAC GGAGTTTCAC
14701 CATATTGGTC AGGCTGGTCT CGAACTCCTG ACCTCGTGAT GCACCCTCCT
14751 CGGCCTCCCA AAGTGTTGGG ATTACAAGCG GTAGCCACTG TGCCTGGCCT
14801 AAACTTTTTT TTTTAAGCAT CCTTGTGCAG AGACTCCAAG GGCAAACAAG
14851 AGTAGTGCCA ACCAGCCTGG CAGAGCTGAG GCAGGCCTGA ACTCTGGCCC
14901 CTAAAGGGAA GAGGTGACTA GTCACTAGCC TTCCTTCATT CTCTCACCCA
14951 GAGAGAAATG GGGAAGGAAG AGAAATGGAG TAAAAGAATA AGGTATGTAG
15001 ATGGGGAGAG ATGGGAAGGG AAAACATGGA GAATGATGGG GAGGCAGAAA
15051 GCGATGCGAA GAGAGATGGA GGAAGACAGA GAGAGAGAAA GACGGAGAGG
15101 CAGACCAGAC ACGGCAGCTC ACACCTGTAA TCCCAGCACT TTGGGAGGCT
15151 GAGGCAGGTG GATCACTTGA GGCCAGGAGT TCAAGACCAG CCTGGCCAAC
15201 ATGGCAAAAC CCCATCTCTA CTAAAAATAC AAAAATTAGC CAGGGGTGGT
15251 GGCATGTGCC TGTAATCCCA GCTACTCTGG AGGCTGGGGC AGGAGAATCA
15301 TTTGAACCTG GGAGGTGGAG GTTACAGTGA GCTGAGAGCG CAAAAAAAAA
15351 AAAAGAAAAA AAATAGAATG TCTCACGTGG AAGGTGCTCG CTGAATATTT
15401 GTTGAATGAG TGAATAAGCA GAAGTGTGTA TGCATGTGAC AGAGAGGCAG
15451 AGGGGTAGGG GTGTGTAGAG ATGTGGCGTT TGAGTGGACA CGGGGGAAGA
15501 AACAAGTAAT ATGAATAACA TGGTGAGACA GAAAGAGTTG TGGACAGAGC
15551 TGTGGGAAAT ATGAGAGATA AGGAGAGAGA TACTGAAAAG AGCGATTAAG
15601 AGAGATGAAG ATAGGGTGTC TGGCCCATGG AAGGCCCTCG TGAAGAGCAA
15651 TGCTGAATAG ATGAATGAAC CTCAATGCCC AGCAGTGGTG GGATGAAGGG
15701 GATGCTGTGC AGAAACCACA CTACCCATCA GAGAAGCAAC TCTGCTCGTT
15751 TCCCCTTGAG TTGATGAGGG ATTTAGCAAC CAGTGGAATG AAGAGCAAGG
15801 GAAAGACACC TGCGATCTTT CAGAAGCAAA ATGGACAGGG CCTGGAGATC
15851 TTTTGGATGG GGGAACTGAG GGAGAAGGCA GAGTAGGATG AATCCCAGGT
15901 GTTTGCCTTA AGTGACTGGG GCTTTGTGTT TCCATTCACT GGGACAGGGA
15951 TGACAGGAGG CTGACCTGGT CTGAGTGGAC ACTGATGACT TTGAAATGCC
16001 TGTGGAGTAT TCAGGGTAAG CCGCCTAGAG AGCAGTTAGA TAGTGTGGAC
16051 TTAGGTCTCA AGAAAGAGGC CTGAGGGGGA GATGTAGATT GGGGACTTGA
16101 CAGAGAATAA GCAGTAGTGA CAACTGTGGC TATGACAGGA TTGCCTAGGA
16151 AAGCCATGGT AGATGCGCAA AATTTATTGA ATGAAAGTGG GAAGCCAGGT
16201 GCATGCCTAT AATCCCAGCT ACGTGAAAGG CTGAGGTCGG AGGAACACCT
16251 AAAGCCCGGA GTCAGGACTT TGAGACTAGC CTGGGCAACA TAGTGAGATC
16301 TCATCTCAAA AATATATATA ATAAAAAATA TATAATAACA CTTTTTAAAA
16351 AGAAAGTGGG GCCAGGGGTG CTAGCTCATG CCTATAATCC CAGCACTTTG
16401 GGAGGCCGAG GCAGGCGGAT CACTTGAGGT CAGGAGTTCG AGACCAGCCT
16451 GACCGACATG GTGAAACCCC ATCTCTACCA AAAATACAAA ATTAGCTGGG
16501 TGTGGTGGCA CATGCCTGTA ATCCCAGCTA CTTGGGAGGC TAAGGCAGAA
16551 GAATCACTTG AACGCAGGAG GCGGAGGTTG CAGTGAGCCG AGATCGCACC
16601 ATTGCACTCC AGCCTGGGCA ACAAGAGCAA AATTCTGTCT CAAAAAAAAA
16651 AAAAAAAAAA AGTGGGAGAA ACAACCCAAA AATTTTCCAA CAGGACAGTG
16701 GTATAAAAGC ACACGTTTAG GGTACAGGAG TCTGGGGGAC TCAGGTTGGA
16751 TGGCTTTTAT GTCCTTGGGA AAGAGGAGGT GAGATTTCAA AGTGAAACAG
16801 CTTCCATGGT GACAGGATCT GAGATGTAGG TCATGGAGAG TCAAGGTGAT
16851 GATTGGAGTG GACAAGGTGA AATAACTGTG AAGTTGCACA CTTAGGGTAT
16901 TGAAAAGGCA TTTGCGAGAT GGGTACAAGA AGTTTCATAG CAGATTGAGC
16951 CATAATAAGA AAAACAAAAA ATAGAAACAA CCCAATTGTA CTTCATCAGG
17001 TGAATAGATC AACAAGCTGT GCTACATCAT ATAAAGGATT ACCACTCAGC
17051 AAAAAAAAAG GAATGAAACA CTGATACATA CAACAGCGTG GGCGAATCTC
17101 ACAGACATTA AGCAGAATGA AAGGAGCCAG ATACAAGAGT ATGTACTATA
17151 TGACTCCATT AATATGATGT TCAAGAACAG ACTAATCTAT AGAGGCCGAG
17201 CGCAGTGGCT CAGGCCTGTA ATCCCAACAC TTAGAGAGGT CAAGGCAGGA
17251 GGATCACTTG AGCCCAGGAG TTAGAATCCA GTCTGGGCAA CATAGAAGGA
17301 CCTCGTCTCT ACAAAAAAAA ATTTTTTTTT TTTTTTGAGA CAGAGTCTTG
17351 CTCTGTCACC CGGGCTGGAG TGCAATGACG CGATCTCGGC TCACTGCAAC
17401 CTCTGCCTCC CAGGTTCAAG CAATTCTCCT GCCTCAGCCT CCCGAGTAGC
17451 TGGGACTATA GGCATGTGCC ACCACGCCCG GCTGATTTTT TTTTTTTTTT
17501 TTGTATTTTT AGTAGAGACG GGGTTTCACT GTGTTAGCCA GCATGGTCTT
17551 GAGCTCCTGA CCTTGTGATC CCCCTGCCTC GGCCTCCCAA AGTGCTAGGA
17601 TTATAGGCCT AAACCACTGC ACCCAGCCTA CAAAAAAACT GTTAAAAAAT
17651 TAGCTGGGTG CGGTGGCATA TGCCTGTAGT CGTAGCTACT CAGGAGACTG
17701 AGGTGGGAGG ATGGCTTGAG CCCAGGAGGT CAAGGCAGCA GTGAACTATA
17751 ATTGCACTAC TGCATTCCAA CCTGGGTAAC AGAGCAAGAC TCTGTCCCAA
17801 GAAAAAAGTC TACAGAGGTT AAAAATGGTA AAATGGTTGC ATCTAGGAGG
17851 GGAAAGGGAT TTACCAAAAA GGGGCATGGA GGAATTTTCC AGGATAATGA
17901 GAATTATCTA TATCTCTATT ACAGTGTGGT TGCATGGGTG TACATATTGG
17951 CCAAAGGTCA TCAAACTGTA CATTAAGATC TGTGCATTTT ACTGTATGAA

18001 ATTATATATC AATGTTTAAA AATGTCATTT ATGGCTGGGC ACGGTGGCTC
18051 ACGCCTGTAA TCCCAGCACT TTGGGAGGCC GAGGCAGGTG GATCACGAGG
18101 TCAGGAGATT GAGACCATCC TGGCTAACAT GGTGAAACCC TGTCTCTACT
18151 AAAAAATACA AAAAAATTAG CCAGGTATGG TGGCTGGTGC CTGTAGTCCC
18201 AGCTACTAGG GAGGCTGAGG CAGGAGAATG GCGTGAACCC AGGAGGTGGA
18251 GCTTGCAGGG AGCCGAGATC ACGCCACTGC ACTCCAGCCT GGGTGACAGA
18301 GCAAGACTCT CTCTCAAAAA AAAAAAAAAA AAAAGTCATT CATATGGACA
18351 TTAAGTCCAT TAAGGAGTTG GCAGAGGTTG AAATAGAGAG GAAGTGGGCT
18401 GGGCATGGTG GCTCACGCCT ATAATCCCAG CACTTTGGGA GGCCGAGGTG
18451 GGCGGATCAC AAGGTCTGGA GTTCGAGACC AGCCTGGCTA ACATGGTGAA
18501 ACCTTTTCTC TACTAAAAAT ACAAAAATTA GCCTGGCATG GTGGCAGGTG
18551 CCTGTAATCC CAGCTACTCT GGAGGCTGGG GCAGGAGAAT CGCTTGAACC
18601 TGGGAGGCAG AGGTTGCGGT GAACCAAGAT CACGCCATTG CACTCCAGCA
18651 ACAACAGCAA AACTCTGTCT CAAAAAAAAA AAAAAAGAAA AAAGAAAAAA
18701 GAAAGAAAAA ATAAATAGAG AAGAGGCTGC AGAGTCAGGG ATGGGCTCTT
18751 GATGAATGTG AGGGCGAGTG GACGACACCA AGGAGAGATT GAAGGTCGTG
18801 AGGTCAGATG CTATGAGCCT CAGAGGAGAA GCTGTTCTCA GAAGCTGAAA
18851 GTATCATCAC CTGGGAGCCA TTCATTAGTT GAAAAATATT GATAAGACTT
18901 GGCCAGAGAA GTCACCCAAT CTCACCCCTG CCAGCACGTA GAGGAACCCT
18951 GGTGGGGGTG GGGGAAAGAT GAGATCACCT GTGCTGGATT GTGGACTTTT
19001 GTGTTTTCAC CTGTTTTCAT GGTCTGAGTT GGGAGTTTAT TGGCCTCAAC
19051 TACATCCAGC CAGCCCACAG CTTGTTGTTA AAGAAGTCAG ACCAGAGACT
19101 GGGTGCGGCG GCTCACGCCT GTAATCCTAG CACTTTGGGA GGCTGAAGCA
19151 GGTGGATCAC TTGAGGTCAG GAGTTCAAAA CCAGCCTGGC CAACATGGTG
19201 AAACCCTGTC TCTAGTAAAA ATACAAAAAA GTTAGCCAGG CATGGTGACA
19251 GGCGCTGTAG TCCCAGCTAC TTGGGAGGCT GAGGCAGGAG AATCGCTTGA
19301 ACCCGGGAGG CAGAGGTTGC AGTGAGCGGA AATCACACCA CTGCACTCCA
19351 GCCTGGGCGA CAGAGCGAGA CTCCCTCTCA AAAAAAAAAA AACAAAAAAA
19401 ACACACACAA AAACAGAGAC CAGTGCACAT CCACTTCAAC ACTGTTTTAT
19451 TTGTTGCCTT TCCTCTTGAG TGCCTGGGAA TAATTTCTTT AGAAAAGTTC
19501 CATCCTTGGC CAGGATGGAT GCAGTGGCTC ACGCCTGTAA TCCCAGCACT
19551 TTGGGAGGCC GACGTGGGTG GATCACGAGG TCAGGAGTTC AAGACCAGCC
19601 TGGCCAAGAT AGTGAAACCC CGTCTCTACT AAAAATATGA AAATTAGCTG
19651 GGCATGGTAG TGGGTGCCTG TAATCCTGGC TACTCAGGAG GCTGAGGCAG
19701 GAGAATCTCT TGAACCCGGG AGGCGGAGGT TGCAGTGAGC CGAGATCGCA
19751 CCACTGCACT CTAGCCTGGG TAACAGAACA AGACTCTGTC TCAAGAAAAA
19801 AAAAAGAAAA AAAAAGTTCC ATTCTTGTGC CCACTGTGCT GATTAGCGCA
19851 GCAGTTCCTG TCTTGCAGGT TAGCCTTCCC TAGAGGTTCT CAGAGTTATT
19901 AAAAGGGGTC CCTTGAATGT ACACACATAT CAGGATTTAT TCTAGATGGA
19951 CTGAAAAAAA AACCCCACAT ACTGTCTTTG GAATCTGGAA CATTTTTAGC
20001 TATGACAAAT GTTTAGAAGT CTTAGATATA GAAAAGAGCT GGGAGAGCAG
20051 GATGCCCTTG AGTCCAGGTG TAGCAATGGC ACTGCCACCC TTGCACAGGA
20101 CGGGGGAAGC ACTGGAGAAG AACCACACAT TGAACACCTA GGGCAATGGG
20151 GTCTGGAAGT TTACAGGTGG TACCCCCAGT TTACAGTAGT GGCACCAGGA
20201 CTCCAACCTA GGCTGATTTG TCTTCTAAGC CACATTGCCT TGGGACGTGG
20251 AGTAGAACCA TGTGATTGAG TGTACACTGT GGAAGAGTGG AGTTTGTTTT
20301 TTTTTTTTTT TGAGACGGAG TTTTGCTCGT TGCCCAGGCT GGAGTGCAGT
20351 GGCACAATGT TGGCTCACTG CAACCTCCGC CTCCTGGATT CAAGTGATTC
20401 TCGTGCCTCA GCCACCCAAG TAGCTGGGAT TACAGGCACG CACCACCACG
20451 TCTGGCTAAG TTTTGTATTT TTAGTAGAGA TGGGGATTCA CCATGTTGGC
20501 CAGACTGGTC TCGAACTCCC GACCTCAGGT GATCTGCCCA CCTCGGCCAC
20551 CCAAAGTGCT GGGATTACAG GCGTGGGCCA CTGCTCCCGG CCGGAAGAGT
20601 GGAGTTTTAT CCCCAGCCAG TTACTTATTC CACGTCCCCA CTGGTGGTGA
20651 CCTCTTCTCC CTCCCCAAAC CGTGGGCTTC TTACAGCTGA GATAGTAGAG
20701 ACAATTTTTC CTATATTGAT CCTTTGGGGT GTGAATGGGG TGTTGATAGA
20751 AGGAGGCAAG CAAGGTAGGG GATCGGGCAG AAGCTGGAAT CAAGGCCCAG
20801 TTGAGAAAAG CCACCATTGG TAGCTCGGGC CTCCCATGCT CTCTGACCAC
20851 CAGGGGGCAG GCCTTGCCTT GCCATGACGC CTCCATCCCA CATCCTCTAG
20901 TTAGATATCC CTGGGGTGGC AGCCACTTCG CTGCTGGAGA ATGGCTGAGT
20951 TGGTCCTTCC CAGGTTGTGT GGCATGTGTG GTTGTTCCGG TAGCTTAGGG
21001 AGGTAAAATA ATTTGTTCAA GATTAGACTG GGCACAGTGG CTCACACCTG
21051 TAATCCCAGC ACTTTGGGAG GCCGAAGTGG GAGGATTGCT TTAGGCCAGG
21101 AGTTTTAGAT CAGCCTGGGC AACACGGTGA GACTCTGTCT GTACAAAAAA
21151 TAAAAATAAG TGAGCCAAAC GTGATGATGT GTGTCTGTAG TCCCAAATAC
21201 TCTGGAGGCT GACGTGGGAG GGTCACTTGA GCTCAGGAGT TCAAAGTTCA
21251 AGGCCACGGG GAGCTATGAT CATATGATCA CACCGCTGCA CTCCAGAATG
21301 GTTGACAGAG TGAGACCCTG TCTCAAAAAA AAAAAGGTTG TATAACATAA
21351 TTGTGGGGAT TGCGGGGGAG CTGACATTTG AACCTAGACT CATTCCAAAC
21401 CAGCATTCTT AACCATGAGA TGACTGCTTC TCCAAAAGAA AGCTCTGAAT
21451 TACTTCAGTG ATGGGTGGGA ACTGGACAGT CTACAAAAGT ACTGTGACAC
21501 AGAGCTAGAA ACACAGATAC AAAGGAAAAA ACATATATTA GGTCAGGTGT
21551 GGTGGCTCAC ACCTGTAATC CCAGCACTTT GGGAGGCCAA GGCGGGTGGA
21601 TCACTTAAGG CCAGGTGTTT AAGACCAGCC TGGCCAACAT GGCAAAACTC
21651 CGTCAGTACT AAAAACACAA AAATTAGCCG AGTGTGGTGG CACACGCCTG
21701 TGAACCCAGA TACTCGGGAG GCTGAGGCAC GTGAATCGCT TGAGCCTGGG
21751 AAGCAGAGGT TGCAATGAGC CAAGATTGTG CCACTGCACT CCAGCCTGGG
21801 CGACTGAACG AGACTGTCTC ACAACAACAA CAGCAACAAC AACAACATGT
21851 ATTCATATAA AATGAGCCAG GCTGGGCGCG GTGGCTCACG CCTGTAATCC
21901 CAGCACTTTG GGAGGCCAAG GCGGGCGGAT CACAAGGTCA GGAGTTTGAG
21951 ACCAGCCTGG CCAACATGGT GAAACTCAGT CTCTACTAAA AATACAAAAA
22001 TTAGCTGGGT GTGGTGGCGC ACGTCTGTAC TCCCAGCTAC TCGGGAGGCG
22051 GAGGCAGGAG AATCGCTTGA ACCCGGGAGG CAGAGGTTGC AGTGAGCTGA
22101 GATTGAGCCA TTGCACTCCA GCCTGGGCAA CAAGAACAAG ACTCCATCTC
22151 AAAATAAATA AATAAATAAA ATAAATAAAA GGAGCCAGAC CAGGAGTTTG
22201 AGACCTGGAC AACATAGCAA GATGCTGTCT TTCAAAAATA AATAAAAGTA
22251 AAAAGTCAGG TGTGATGGCA CATACCAGTA TTCCAAGCTA CTTGGGAGGC
22301 TGAGGCAGGA GGATTGCTTG AGTCCAGGAG ATCGAGGCTG CAGTGAGCAA
22351 TGATCACACC ACTGTATTCC TGCTTGAGTG ATAGAGTGAG ACCCTTGTCT
22401 CTAAAAAATG TAATAGTAAA AGGAGCCAGA AGGGGCATG CGGTAAGGGG
22451 TGGACACATG GCATATTACA TGTCATCCTT CACTTCTATG TGCTGTATTA
22501 TCTCCTGTAG CTGTGTGATA CACATACAAG AGTGTCTGGC ATACTCTTGT
22551 ACCAATCCCT GTAGAGGTAT TACCTTGTTT TATTGTGTTC CACAGATGTG
22601 CATTTTTTTA TATATTGAAG TTTGTGGCAA CCCTGTGTTG AGCAAGTCTG
22651 TAGGTGCCAC TTTTCCAATA GCCTGTGCTC ACTTTGTGTC TCAGTGGCAC
22701 ATTTTGGTAA TTCTTGCAAT ATTTCAAACT TTTTCTTGAT TATTTTATCT
22751 GTTATGATCT GTCAGTAGTG ATTTTTTTTT TTTGAGACAG GGTCTCACCC
22801 TGTCATTCAG GCTGGAGTGC AGTGGCAAGA TCATAGCTCA CTGCAGCCTC
```

```
22851 GAAATCCTGG GCTCAAGTGA TTCTCCTGCT TCAGCCTCCC AAGTAGCTAA
22901 GACTACTGGA ACACACCAGC ACACCTGGCT AATTTTATCT ATTTATTTAT
22951 TTAGGCAGAG TTTTGTTCTT GTTGCCCAAG CTGGAGTACA ATGGCACGAT
23001 CCCAGTTCAC CACTACCTCC GTCTCCCAGT GATTCTCCTG CCTCCGCCTC
23051 CCGAGTAGCT GGGATTACCG GCATGCGCCA CCACGCCCGA CTAATTTTGT
23101 ATTTTTAGTA GAGACGGGGT TTCTCCATGT TGGTCAGGTT GGTCTTGAAC
23151 TAACGACCTC AAGTGATCTG CCCTCCTTGG CCTCCCAAAG TGCTGGGATG
23201 ACAGGAGTCA GCCACTGTGC CTGGACTTTT TTTTTCTTTT TTGTAGAGCT
23251 GAGGTTGGCC AAGCGCAGTG GCTCATGCTT GTAATCCCAG TACTTTGAGA
23301 GGTCGAGTGG GCAAATCACT TGAGCCCAGG AGTTTGAGAC CAGCCTGGAC
23351 AACATGGTAA GACCCTGTCT CTACAAAAAA TACAAAAATT ATCTAGGCAT
23401 GGTGGTGTGT GGCAGCTACT CAGGAGGCTG AGGTGGGAAG ATGGTGTGAG
23451 CCCATGAGGT TGAAGCTGCA GTGAGCCGTG ATTGTGCCAC TGCACTCTAG
23501 CCTAGGCAAC AGAGTGAGAC CTTGTCTCAA AAAGAGAGAG AGAGAGAGAA
23551 GGTCTTGCTC TGTTGCCCAG GCTGGAGTGC CATGGAGCCA TGGCTTTCAC
23601 AAGCAAAATC GTAGCATACT GCAGTCTTGA ACTTCTGTCC TCAAGTGATC
23651 CTTCCACCTC AGCCTTCTGA GTAGCAGAGA CATAGGCACA TACCACTGCA
23701 CTTGGCAAGA ATTAGCTAAA TTTACTCTGC CACGCTGTGT AAATGGAACA
23751 ATAAAGCCTA GATGGCTGGA TGTGGTGGCT CATGCCTATA ATCCCAGCAC
23801 TTTGGGAAGC TGAGGTGGGC GGATCACCTG AGGTCAGGAG TTCAAGACCA
23851 GCCTGGCCAA GATGGTCAAA CCCTGTCTCT ACTAAAAATA CAAAAATTAG
23901 CCGGGTGTGG TGGCAGTCAC CTGTAATCCC AGATACTAGG GAGGCTGAGG
23951 CAGGAAAATC ACTTGAACCC AGAGGGCGGA GGTTGCAGTG AGCTAAAATT
24001 GCACCACTGC ACTCCAGCCT GGGTGACAGA GCAAGACTCC ATCTCAAAAA
24051 AATATATAAA TAAATAAATA AATAAATAAA AGCCTAGATG ATAGTACATC
24101 TGTTTACAAC GTGGTTTACT GATTTTAAAA TTTTTCTTTT TTTTGTTTGT
24151 TTGTTTTTTT TTGAGACACG GTCTCAATAT GTTGCCCAGG CTGTAGTACA
24201 GCAGCATGAT CTCAACTCAT TGAAACCTCT GCCTTCCAGA GTCAAATGAT
24251 CCTCCTACCT CAGCCTCCTG AGTAGTTGGG ACCACAGACA TGAGCCACTA
24301 TGCTCAGCTA ATTTTTGTAT TTTTTGTAGA GATAGGGTTT CGTCAGGTTG
24351 TCCAGGCTGA TCTCCAACTC TTTTTTTTTT TTTTTTTTTT TTGAGACAAA
24401 GTCTTGCTCT GTCACCCAGG CTGGAGTGCA GTGGCGCAAT CTCGGCTAAC
24451 TGCAACCTCT GCCTCCAAGG TTCAAGTGAT TCTCCTGCCT CAGCCTCCTG
24501 AGTAGCTGGG ACTACAGGCA GATGCCACCA CACCCAGCTA ATTTTTTGTA
24551 TTTTTAGTAG AGATGGGGTT TCACCGTGTT AGCCAGGATG GTCTCGATCT
24601 CCTGAACTCG TGATCCGCCC GCCTCAGCCT CCCAAAGTGC TGGGATTACA
24651 GGCATGAGCC ACCATGCCTG GCCTGGCCTC CAACTCTTTA ACTCAAGTGA
24701 TCTGCCCTCC TGGGCTTCCC AAAGTACTGG GATTACAGGT GTGACCCACC
24751 ATACCTAGCC AAAATTATTT TTTTTTAAAC TCCATTTACT ACTATTCCCA
24801 ATTTAAATTG TATTTTATTT AGAGACAAGG TCTCCCTCCA TTGCTCAGGC
24851 TGAAATGCAG TGGCAAAATC ATAGCTCACT GCATCCTTGA AACTCCTGGG
24901 CTCAAGGAAT CCTCCTTCCT CAACTTCCTG AGTAGCTGGT ACTATAGGTG
24951 TGTGCCACTA TGTCCAGCTA AGTTTAAATT TTTTTGTAGA GACAGGGTCT
25001 TGCCATGTTG CCCAGGCTGG TCTCAAACCC CTGGCCTCAA GCCGTCCTCT
25051 CACCTTGGCC TCCCAAAGTG CTGGGATTAC AAGTGTAAGC CACTATGCCC
25101 CGCTGGTTTA CTGAATATTT TAAGCCTACT GTTGAGGTCT ACTGCTCCGA
25151 AAAACAAGAT TTCTTTAAAA TGATTACTGC TCATTGACAA TGTACATAGT
25201 TATCCAAGAG CTCTGATGGA GATGTCCAAA GAGACGAATG TTGTTTTCTT
25251 TTTCTTTTTT TTTTTTTTTG AGACGGAGTT TCGCTCTTGT TGCCCAGGCT
25301 GGAGTGCAAT GGTGCGATAT CAGCTCACTG CAACTTCCGC CTCCTGGGTT
25351 CAAGTGATTC TCCTGCCTCA GCCTCCCGAG TAGCTGGGAT TACAGGTGCG
25401 CCCCACCACA CCCGGCTAAT TTTGTATTTT TAGTAGAGAC GGGGTTTCTC
25451 TATGTTGATC AGACTGGTCT CGGACTCCTG ACCTCAGGTG ATCCTCCCAC
25501 CTCAGCCTCC CAAAGTGCTG GGATTACAGG CATGAGCCAC CACGCCCAGC
25551 TGAATGTTGT TTTCCTGCCT GCTAACATAA CATCCATCCT GTAGTCCATG
25601 GTTTAAGGAG TAATTTAGAC TTTCAAGTCT CATTATTTAA GAAATACATT
25651 TCGCCAGGCA TGGTGACTCA TGCACTTTGG GAGGCCTAAG TGGATGGATT
25701 GTTTGAGCCC AGGAGTTGGA TACCGGCCTG GGCAACACAG GGAGACCCCA
25751 TCTCTACAAA AAATACAAAA ATTAGCCTGG TGTGGTGGTG TGCGCCTGCA
25801 GTCCCAGCTA CTCAGGAGAC TGAAGTGGGA GGATCAACTG AGCCCTGGAG
25851 GTTGAGGCTG CAGTGAGCCG TGATGGCGCC ACTGAACTCC AGCCTGGGTG
25901 AGACAGAGCC AGACCCTGTC TCAAAAAACC AAAACAGGAA TCCCAGCACT
25951 TTGGGAGGGC GAGGCTGGTG GATCACTTGA GATCAGGAGT TGGAGACCAG
26001 CCTGGCCAAC ATGGTAAAAC CCTGTCTCTA CAAAAAATAC AAAATTTAGC
26051 TGGGCTTGGT GATGTGCGCC TGTAATCCCA GCTACTTGGG AGGCTGAGGC
26101 AAGAGAATTG CTTGAACCCA GGAGGCAGAG GTTGCAGTGA GCCGAGATTG
26151 TGCCACTGCA TTCCAGCCTG GGCAACAGAG TGAGACTCCA GCTCAAAAAA
26201 AAAACAAAAA ACAAAAAAAC CCAACCAAAC ACACACACAC ACACACACAC
26251 ACACACACAC ACACACACAC ACACACACAC ACACAAGGTG ATTGCTATGC
26301 AAATTAAATA TAGCTTCACC AATCAGCTTG CTTCTCCTAA TCCCTCATAG
26351 AAATCTTGGA GTAGCCCTGG GGTGCAGTGA GGCCAGATTC CAGGAAGGGA
26401 ATGGGTGGGA TGGGTCTGGG TACTGGTGAG CAGCAGGGAG GGTAAGTAGG
26451 TTAGCTTCTG GGCCAAGGAG GGCTCCTGGA GGCTTGTTCT GGGTAAAAGC
26501 TGGTCTCCTT GAACCTCGGA AGGTGGGGGA ACCCACATCC TGGTCTCACC
26551 CGCCCAACCG CCCACTCACC CTTGGGTTAG CAGCCATCTA AGCCCCACCC
26601 TCCCCCTCAC ACGCTTAGCT AGCCTGCCAC AAGCTGGCCC CTTGGCCTCC
26651 TAGAGACCCT GACATCTCCT CCAGCAGCAT CTGTCCTCTC TCCTCAGGGA
26701 GGCAAGCATT TGATGCTCGA GGTCCCTGGC AGTTGTGGTC CTTGGCAAGT
26751 GATGTGTGAG TCCCGTGTGT CATAGGAAGC TCCCCATCCC CATCTGGTGA
26801 CCAAAGGCCT GGCTACAAGT AGTGAGTCCT TCCTCCTCCA CCCAGACCTC
26851 ACTGCTCAGA TCCCCTTCGC CAACTGGGAC ATCTTCCGAC ATGGCCTGGA
26901 TGCTGTTGCT CATCTTGATC ATGGTCCATC CAGGTGACAG GGCGTGCGCT
26951 CAGGACCCCA AGGAGTGTGG GTGGGAGGAG GGAGATCCAG GAGGCTGGAC
27001 TAGATGCTAT AGGGAACGGG CTTGGTGGGG GCTGAAACAC ACGGCTCTGA
27051 GGGAGGAGTG GGACTGCTCC AAAGGTGACA CTCAGTGGAC TCCCCCAATT
27101 CACAGTCTCT CTGTGTCACC CACTGTGAGG CTTTTTAGTT TGAGAGATCC
27151 AAGTAAATCT ATGATAAATT TCTTGGTGAA ACTTTTTTTT TTTTTTTTTT
27201 GAGACTGAGT CTCGCTCTGT CGCCTGGCTG GAGTGCCGTG GTGCGATTTT
27251 GGCTCACTGC AACCTCTGCC TCCCATGTTC AGGTGATTCT CCTGCCTCAG
27301 CCTCCTGAGT AGCTGGGACT ACACGCACGC ACGACCACAC CTGGCTAATT
27351 TTTTTTTTTT TTGAGACGGA GTCTTGCTCT GTCACCTAGG CTGGAGTGCA
27401 GTGGCATGAT CTTGGCTCAC TGCAACCTCT GCCTCCTAGA TTCAAGCAAT
27451 TCTCCTGCCT CAGCCTCCTA GGTAGCTGGG ATTACAAGCG CGCACCACCA
27501 CACCCAACTA ATTTTTGTAT TTTTAGTAGA GACAGGGTTT TACCATGTTG
27551 GCCGGGCTGG TCTTGAACTC CTGACCTCAG GTGATCTGCC TGCCTCGGCC
27601 TCCTAAAGTG CGGGGATTAC AGGCCTGAGC CACTGCGCCC AGCCATTTTT
27651 GTATTTTTAG TAGAGATGGG GTTTTGCTAT GTTGGCCAGG TTGGTCTCAA

27701 ACTCCTGATC TCAAGTGATC TGCCTGCCTC GGTCTCCCAA ATTACTGGAA
27751 TTATAGGCAT GAGCCACTGC GCCCGCCTGG TGGTGAAACT TTTTTTTTGA
27801 GACAGTTTCA TTCTGTTGTC CAGTCTAGAG TACAGTGGCG GTATCTCAGC
27851 TCACTGCAGC CTCCACCTCC TGGGTAAAAA TGATTCTCCT GTCTCAGCCT
27901 CCCAAGTAGC TAGGATTACA GGTGCATGCC ATTACTGCTG GCTAACTTGT
27951 GTATTTTTAG TAGAGACGAG GTTTCACCAT GTTGGCCAGG CTGGTCTCAA
28001 ACTCCTGACC TCAGGTGATC CACTCGCCTT GGCCTTCCAA AGTGTTGGGA
28051 TTATAGGCGT GAGCCACTGC ACCCGGCCGA AACTGTTTTT AATGAACTGA
28101 GAGACCTTAA CTATCAGGCA TATTATTAAC TAAATGCAGG AGTTCTCAAA
28151 ATGTGGATTT CTGGGAACCT AGAACAAATT CTTAGGCCCC ACTCCAGACC
28201 TACTGAGTCA AAAACTCGGG GGGTAGGCCC AGGAATCTGT TCTAGCAGAT
28251 GCTCCAGGTA ATTTCCATAC ACACTCAAGT TTGAGAACCA CTGAAATAGA
28301 ATGACTTGTA AATTTCCCTA AGTAGATAAT TTAATCAGGG ATTTTAATAT
28351 TTGGTTTAAT TCATCAACAT GGACGGGTTA AATAGGAATC TCAACCAATT
28401 AAGGCCACGT CTTCACCTGG AGATTTAATT AGTTACTTTC TTTAACGAAA
28451 ACCAAGAGTG GCTGGGTGCA CTTTGGGAGG CCGAGGCAGG TGGATCACTT
28501 GAGATCAGGA GTTTGAGACC AGCCTGGTCC GAGGTGGGTG GATCACTTGA
28551 GGTCAGGAGT TTGAGACCAG ATGGTGAAAC CCTGTCTCGA CTGAAAATAC
28601 AAAAATTAGC CAGTCGTGGT GGTGGGTGCC TGTAGTCCCA GCTACTTGGG
28651 AGGCTGAGGC ACCAGGATCG CTTGAACCCA GGGGACGGAG TTTGCAGTGA
28701 GCCAAGATTG CACCACTGCA TTCCAGCCTG GGCAACAGAG CAAGACCCCA
28751 TCTCAAAAAA AGAAAAAAGA AAAAGAGGAA GGATGGCTTA CTGTACAATG
28801 CCATTTGTAC TAAAATAATA CCTGGATAAT ATAATGAGTG ATATTAGTTA
28851 ACTAGGCAGC TGCATTCATT AATGAGCTTA ATTTCACCAT GATGGTTTAC
28901 ATTTCAGCTA GACAAGTTAC TACTGAACTG GCTGGAGAAT GATGGCAGAG
28951 GGTGAGAGTG AGAGTTGGTA TAGGAAGAAT TTGAAAAATG ATTTTTATTT
29001 TTTATTTTTT GAGATGGAGT CTTGCTTTGT CGCCCAGGCT GGAGTGCAAT
29051 GGCGCAATCT TGGCTCACTG CAACCTCTGC CTCCCGGGTT CAAGCGATTC
29101 TCCTGCCTCA GCCTCCCGAG TAGCTGGGAT TACAGGTGCA CACCACCATG
29151 CCCGGCTAAT TTTTGTATTT TTTTAGTAGA GACGGGGTTT CACCATGTTG
29201 GCCAGGATGG TCTCGATCTC CTGACCTGGT GATCCGCCCA CCTCGGCCTC
29251 CCAAAGTGCT GGGATTACAG GCGTGAGCCA CTGCACCCAG CCGAAAAATG
29301 ATTTAATAAG CAATGTTAAG AAATCAGATT GGTTAAGAGG GAAGGGTTTA
29351 ATGAGGCACC CAAAGTATCT ATTCCCATGC ACCCTATGCC TGGAGAAAGC
29401 TGGGATGTTC TACTCCAAGC TCTGTTGTCT TTTCTCTTTG GAATAACTGG
29451 GGAGGTGTTT CTCTGGGTCT TCCTTCTGCC CCCAGGATCC TGTGCTCTCT
29501 GGGTGTCCCA GCCCCCTGAG ATTCGTACCC TGGAAGGATC CTCTGCCTTC
29551 CTGCCCTGCT CCTTCAATGC CAGCCAAGGG AGACTGGCCA TTGGCTCCGT
29601 CACGTGGTTC CGAGATGAGG TGGTTCCAGG GAAGGAGGTG AGGAATGGAA
29651 CCCCAGAGTT CAGGGGCCGC CTGGCCCCAC TTGCTTCTTC CCGTTTCCTC
29701 CATGACCACC AGGCTGAGCT GCACATCCGG GACGTGCGAG GCCATGACGC
29751 CAGCATCTAC GTGTGCAGAG TGGAGGTGCT GGGCCTTGGT GTCGGGACAG
29801 GGAATGGGAC TCGGCTGGTG GTGGAGAAAG GTGAGATGCT GGGAGGTGGT
29851 GTCTCCTCCT GGCTGGAGGC CCCAAGAGGC AATGTCCTTG GGAGGCAGGG
29901 ATGCTCCTCT GAGGCCCCTT CCCTCCCTGA GCCTGTGTGC ACTTCTTCCC
29951 CAACCCCCGT CTCCATTGCC CCATGCAGAA CATCCTCAGC TAGGGGCTGG
30001 TACAGTCCTC CTCCTTCGGG CTGGATTCTA TGCTGTCAGC TTTCTCTCTG
30051 TGGCCGTGGG CAGCACCGTC TATTACCAGG GCAAATGTGA GTAATGGAGC
30101 CAGGGGCAAT AGTGGACGGG ATGGGAGGGG CAGTAAGAGA GTGGGAGGAG
30151 GGAGGACAGA GACCAGGAAG AGGAGAGCCT CGGGACTGCA ACACTGAGCA
30201 GCTCCTGTCC TCTCTCTGAC CAGGCCACTG TCACATGGGA ACACACTGCC
30251 ACTCCTCAGA TGGGCCCCGA GGAGTGATTC CAGAGCCCAG ATGTCCCTAG
30301 TCCTCTTCAA AAGACCCCAA TAAATCTGCC CCACCACTAA CTCCTCATGA
30351 GTCTCAAGTG TTTTCTTCTC CATTCTCCAG ATGCCAAATC TACTCTCTCC
30401 GGATTCCCCC AACTCTGAAC TTTCCCTTCC ACCAGGTCTG ACCTGGAAAG
30451 GTCCAAGAAG GCAGCTGCCG GCTGTGGTCC CAGCGCCCCT CCCACCACCA
30501 TGTGGGAGCT CAGCACATCT GCTTCCCCCA GTCCCAGGAG GCTGAGCCTG
30551 ATTGTCCTGA GAAATGGGAA GGATCAGATA TGACTCCTCC TTGGCAACTG
30601 CCCTTTCCTG CCAGGCCCAC ACATACCCTC TTCTGGCTGT TAGGGGAGCT
30651 TGGGTCCCTG AACACTGTCA TTCACCCAAT AAATTACTAT TTGACCCCAG
30701 AGTGGGTGGA AGGGTGAGCC ATGTGTTTTT TTTATTTTAA TTTTTAAAAA
30751 ATTTAAAAAA TTCCCTATTC AAAGGTCAAA AAGCCACATA AGTTTTGATG
30801 ATGATCAATT TGAACGGAGG CTCGAGATGG ACTGAGAGGA CTGAGACACA
30851 GAAGTGGGGG GACCATGGTT TTTACTGGCT GGACCACAGG GGGACCCTGT
30901 CCACCCGCCT GGGTTGAGGA AGGTGTCTGG GGTGCTCAGG TGGGTTTGTT
30951 CTCAGCAATG CAGGCATAGT CAGCTCTTGG ATCCTCCTTG GTGCCTCTCT
31001 TGTCTCTGCC CCTGAGGTCA GGTCCCTCAC TGCTGGGCAC TGGCAGCCTC
31051 TGCAGAGATG CATAGTGGAG TTCCTGCTCT GAGGAGCCCT GGGCCTGGGA
31101 CCAGGACAGA AGGTGCTGAT GGGAGGCGAT GCCGTCAGAT CCTTCCCTGT
31151 GAGTTCTGCT CCCACCTCCA GCCTTTCTTA CTTCTCTCCC TCTCTCTCTC
31201 TCTCTCTCCC TCTCTCTCTC TCTCTCTCTC TCTCTCTTTT TTCTTTGGAG
31251 ACAGAGTCCC ACTATGTTGT CCAGGATTGT CTTCAACCCC TGAGCTCAAG
31301 CTATCTTCCT GCCTCAGCCT TGCGATAGCT GGAATGACAG ACGTGAGCCA
31351 CTGTGCCTGG TTCTGGAGCC TCTCTCTCTC TCTCTCTTTC ATTGCTCTTC
31401 TCTTTGTGTC TCTCTCACTC TCATTTTCTC CCTGTCTCTC CTATCCTCTG
31451 TCTCACTTTT TCTCTTGGTT TCTGTCTCAT TTTCTCTTTC TCTTTTGCCT
31501 CGATTTTCTC TGCCTCTCTC ATGCTCCTAC TTTCTCTCTC CTTGTCTCCC
31551 GTCCCCAACC CTCCTCTCAG CGCTCAGCCA TGCTTCTCCC CACTCACCCA
31601 CTCAGGATCT CTCTTGCCCT CCCCCTTCCC TGTCCCCAGA CTCACCCAGC
31651 TCCTCTCCAG CCTCTTTACT GGAAGAAAAG AAGAAGCTCA ACACAGCCCA
31701 CCCTTTGTGC TTCTCCCGGG CCCTCCCGGG CTCCCCCCAC CAGCAGGCGT
31751 GGACTCCCCT GTTGGCTTCC CAGTGGCTCC AGGGCCAGGC AGTGTTCTGG
31801 GAAAGCAGTG GGAAAGCGTG TGGGGGTGGG GGCACAGGGG GCACTGCTGC
31851 AGGGGGAGGG AGGGAGTGCA GCGCTCACCT CTTCGATGCA GCCAACACAG
31901 GCAGGCGGAC AGAAGGACCA CTGCCAGAAG CAGGAGCCCG CCCAGCCCCA
31951 GGCCCCCGTA GATACATATA TCTTCAGGGA AGAGGGCTCA AGGTTAGGAA
32001 GCCCATTCCT TCTCCAGCGT ACCCCAGCCT CCTGGTTGGT TGCAGCTTTC
32051 TCAGATTCCC TCTCCAACAG TTTTAGAGGC AGAAAAATAT ACCCTCAGAG
32101 CTTCTCCATC CCAGACCTTA ATACCTCACC TCTTACCAGG TTTCTGGGCC
32151 TCCCGTGAGG TCCCTTTCCC TCCTGTACCA GCTGTCCCCA GAGGCCTGTC
32201 CACCTAGTCA TGAGCTGCAT ACATCACTGT TCCCCATCAC TTCCTAAGCT
32251 CCCAGGACTC TCTCTATGCA GATGCAAGAG ACACTTTACT TACCATCATT
32301 CCGCGATAAC ATTAGGTCAG GGATCAGGGA CTGGCCTGGA GGTCAGGAAC
32351 TCTAGTCCTT GCTCTTTTAG GCGAAATGAT CAGGGGCTGG TGACTTGCCT
32401 CAAGTTCCTC ATCTGTGAAG TGAGGGGCCC TCTGTTATAA TCGTTCCCAG
32451 GCTGGGGAGC CTCCCTTATG TGCCAACCCT GAGCTGGGCA CTTTCCATTC
32501 CTCACTGCTA ATCCCCAGAA CATAGGGTAT CATGGTGCCC GTTGCCCAAG
```

```
32551 TAAAGACCCG AGATTCAAGC CTGGACTTTA CTAGGTCACC CATCCCAGAA
32601 AGGTGGAGCT GGGGTTTAAG GCCAGCTCTG CCCAACTCCA GAGCCCAGCC
32651 CTTTCCTCTG CCCTGGGCTG ACCACGTGGT TTGGAGGGGA CTTTTCAGCC
32701 CTGGATGGTT CTAGGTGCTG GTAAGGGGAT GATGGAGGGG AAGGAGCCTG
32751 GGCCTGGGTG GGTGTGGGCA CTGGGGGGAA GAAGGGAGGG TGATATCAGC
32801 ACACCCAGCA GGTGGGCTGC TCCCTGAGCC GCAGAGCAGC GCGGGAGTGT
32851 GGGGGCCCCC TTGGCTGGTG TGGAGAGCTG CTTCCCACAG GCAGATGCTG
32901 CTAGGGCTGA AGTGGGGCAT GGAGGAGTAT CTGGGGCCCC ATAACTTCCC
32951 CTCAGGCACT TCCTCCCCTC CAACCACTGG TTCCTGTTTG AGGGTGAAGA
33001 GGGGGCCGTT CTCTTCACCC CAGAGCCAGA TATACTGACT AGGGTCTGGA
33051 AACTGGGACC CTTCTGGGTT TAAGAGGAAT TCTGGGGGGT GGGGAGCAGA
33101 AGTGCAGGTG GAGGCCATAA GGGCCGTGGG CACAGAAATG AATTGTCTTT
33151 AATTTCTTTG GGGAGCAGAG ACTCAGAGGA TTCCTCGACG GCCCAGGGAA
33201 ACTCAAACCC ATACTCTCCC TCCCCTCATC TTAGCTTCAC CCCACTCTGG
33251 GGTGTGACCA TCCTTCCACC AAGGTCCCTG CCCATTCCCA GCTTACCCAG
33301 AGCTTGTGGC CTGCAGGATG GACAGACTCC AAACTGGCCA GTGCTGTCTG
33351 GCTGAAGAGA AGTGTTGCCA CCTCAGACAT TCCTGCCCCT CCTCTGGCTT
33401 TAACTTCTCC CCCAGCCTGG GTTCCTCCCC AGCATTGTTA GGAGAGGAAG
33451 TTCGGCTCCA GGGTTAGGGT TACAACATCT TTCTTTTCAA ACTTCTGGGC
33501 CTGTAGCTAG GGCCATAATC TGCCTCAGCT CCAGTCATTT CACTGATCCT
33551 GCCACTAGTC AGCAAACACC CACCACCTTC AGGTGTCTTC CTTGATTAAC
33601 TCTTCTCTAT TACCTCTTAC ACTGCATCTG TTGAGACCTC TTGATCTTGG
33651 CATATGTCTG TGGACACATG TACCCTGCTC CAGTTCAGAC TGGGAGCTCT
33701 GGCATATTTG GTGTTCTGGT GACCCACTTG GTGTTCTGCA GCTCTATGCA
33751 GCAGGCTGCT TGGCTGTCTG ATTAACAAAC ATGAGAGTTA AAGGCAGCCA
33801 TACTAGATCT ACATGAGCCC TGTGGATAGA TCAAGGACAG AAAGTTGGTG
33851 ACAAGTTGGT GACAGGAAAG GAGGTGTGGG CAATACAGGT GGACTTCCTG
33901 TAGGAGGCAG CATTCTGGCA TTGAACACAG AACTAAGTAA AGGCAGCGGC
33951 CTGAGGCTCT CTGGAGAGAC CTGGGTGAAG GCTTCTTAGT GGCACTGTGA
34001 TGGAGAAGGA GATGGGTGCT GGAGGATTGC ACACCCACTC CTTGAGGAGG
34051 GTGAAGACTG GGGAGCGCCT GTAAGGCAAG GGGTGAGGAG GAGATGTGGT
34101 GCCTGGGTGG GTCCATATTA ATTTGCGTTT CCCCTTCCAA ATTCAAGAAC
34151 CTTCTACTTC CTCCCCCACC CTTCTCCCCA TCCTCCATGT TCCCCATGCT
34201 GAATAATATT CAGGGTTTTT TGTTTGTTTG TTTTGTTTTT TTTACATTTT
34251 ATTATTAAAA AGTGCAAACA TAGGGTGAAA TGAGAATAAT TGTACAGTGA
34301 ACATATTCTT ACCACCTAGA TGCTACTATT AACATTTTTT GTTTTGTTTT
34351 GTTTTTGAGA TGGGGTCTCA CTCTGTCACT CAGCCTGAAG TGCAGTGGTG
34401 AAATCATAGC TCACTGCAGC CTTGAATTCC TGGGCTCAGA GGTCCTCCCA
34451 CCTTAGCCTT CTGAGTAGCT AGGACTACAG ACACCAGCTA CCACATGAGG
34501 CTTTGTAGAA ATGGGGTCTT ACTATGTTGC CCAGGCTGAT TTTGAACTCC
34551 TGGTCTCAAG CAATCTTTCC ACCTTAGCCT TCCAAAGTGC TGGAATTACA
34601 GGAGTGGGCC ACTGCACCTG GCTCTATTAA CATTTTTTAT TTGCTTTATC
34651 ACATATTTAT CAATCCATCT CACTTTTAAA TATCTTTTAA AATTACAAAT
34701 ATCAGTACAT TTTACATCTA AACCCTTCAG AAGCTTAACA TTGACTGGAG
34751 TTCAGTATTT ATTTCCCCAT TTCTTTTCTG GCCTGAGGAA GGCAAATTTT
34801 ACATACAAAT CTCAAGTCAG TACTCTTTTT TTTTTTGAG ACGGAGTCTT
34851 GCTCTGTTGC CCAGGCTGGA GTCCAGTGGT GTGATCTTGG CTCACTGCAA
34901 CCTCTGCCTT CTGGGTACAA GCGATTCTCC TGTCTCAGCC TCCCAAGTAG
34951 CTGGGACTAC AGGTTTGTGC CACCATATCC AGCTAATTTT TGTATTTTTA
35001 ATGGAGAAGG GGTTTCACCA TGTTGGCCAG GCTGGTCTCA AACTCTTGAC
35051 CTCAAGTGAT CCACCTGCCT TGGTCTCCTA AAGTGCTGGG ATTATAGGTG
35101 TGAGCCATCT CGCCTGGCCT AATACTGTTT TGTTTGTTTG TTTTTGTTTT
35151 TAAGACAGAG TCTTGTTCTT GTCACCCAGG CTGGAGTGCA ATGGCATGAT
35201 TTCGGCTCAC TGCAACTTCC GCCTCCTGGG TTCAAGTGAT TCTCCTGCCT
35251 CAGCCTCCCA AGTAGCTGGA ATTAAAGGTG CCTACCACCA CGCCCCGCTA
35301 ATTTTTATAT TTTTAGTAGA GATGGGGTTT CACCATGTTG ATCAGGCTGC
35351 TCTCGAGCTC CTTACCTCAG ATGATCCACC TTCCTTGGCC TCCCAAAGTG
35401 CTGGTATTAT AGGCAAGAGC CACTGCGCCC AGCCCCAGTA TTCAGTTTTT
35451 AAACTGTCTT GTTATCAAGG CTCTGGAGCC AGATGCCTGG GTTCAAATTC
35501 TGGTTCTGCC ACTGACTCTG TGAGCTCCAT AAGTTTCTTA ACCTCTCTGT
35551 ACCTCAGTTT CCTCTTAGGG TTTTTGTCAG GATTATAATT ATTGGCTGGG
35601 CATGATGGCT CATGCTTGTA ATCCCAGCAC TTTAGGAGGC CAACACGGGC
35651 AGATCACGTG AGTCCAGGAG TTTGAGCCCA GCCTGGGCAA TGTGGCAAAA
35701 ATCCATCTCT ACAAAAAATG CAAAAATTAG CTGGGCATGG TGGCATGTGC
35751 CTATAGTCCC AGCTATTCAG GAGGCTGAGG TAGGTGAATC CATAGATCCT
35801 GGGAGGTCAA GGCTGCAGTG AGCCATGATC CTGCCATTGC ATTCCAGTCT
35851 GGGTGACATA GCGAGACCCT GTCTCAAAAA AAAAAATTAT TAAAGTGTGT
35901 AAATCAGTGG CATAAACATG TTAAGTGCAT TTTGTGGGTC AGCTATATTA
35951 TTATTAGTAT TACGGAAACA CATAGAGATG TTACCAAGAA GGGGAGATGA
36001 TTGGAGCCAC TTCCAGCTTC CTTGGACCTG GTCTTTCTTC CCTTGACTCT
36051 TTTTTTTTTT TTTTTTTTTT TTTTGAGAGA GAGAGTCTCA GCCTGTTGCC
36101 CAGGCTGGAG TGCAATGGTG CAATCTTGGC TCACTGCAAC CTCTGCCTCC
36151 CAGGTTTAAG TGATTCTCCT GCCTCAGCCT CCTGTGTAGC TGGAATTACA
36201 GGCGCGTGCC ACCACGCCCG GCTAACTTTT TGTATCTTTA GTAGAGACAG
36251 GGTTTCACCA TGTTGGCCAG GCTGGTCTCG AACTCCTGAC CTCAAGTGAT
36301 CCACCTGCCT CAGCCTCCCA AAGTGTTGGG ATTACAGGTG TAAGCCACTA
36351 CCCCGGCTAC TCCCTTGACT CTTAACCACT CATGCTGCCT ACATCTACCA
36401 TTCATGTGGT CCTTGCTGCT TTGTTTTGGT TATTCCTGCA TTTATTTGTC
36451 CTTTTATTCA TTTATGTATA AACATTTAGT AAGCACCTAC TAATGGATAG
36501 GGCTCATTGT AGACTTGGAA GCTCTCTGAG GGTGGGAGTA TGCCTCGTCC
36551 ATCTGTCTTT ACTTTTTGTA GCAAGGGAGG TAAAGCTCCA TTTCCATCCC
36601 TCCTTAGTGA GTCAGTAGTC AGTGGTGAGG CTAAGGCTTA CCTCTCCCTT
36651 TCTCACTCAG CACAGGGGGC TGGAGATGAG CAAGGGAACG GGAGGAGGTC
36701 AGCCCAGTAT GGGAATCAGT TCTTCTCAGG GAACCCAGAC ATCCATCCCT
36751 CAAGATTCCA GTCCTTGTCC TAGTCCGGCC CTTGACCTCA GAGACGGGAT
36801 CAGCTCTTCC TCCAGCACCT ACCTTGAGGG TATAGAAGAA TGCAAACCAC
36851 ATTGGAAACC TGGAGATCTG TGTTCTCATT TCAGCTCTGC TGACTGGCTT
36901 CCTGCAAGCT ACCTTCCCTC CCTGGGCCTC AGTTTCTCTC TCTGCTGAGC
36951 CAGAAGATGT CTAAAGACCC CTTTGGTTCC ACCCTGAGAG CCTGTCTCCC
37001 TAACCTCAAC TTCTTCCCCA GTTCAGAGAA CCCAGGCATC CAGCTGCCCC
37051 ACCCCAGCTC TGGGTAAACA GGAAGCTGGG TGAGGGGAGC AGGGGTGTGC
37101 GGAAAGTCCC AGCCAGGTGT GCAGGTCTAC AGGGAGGGGG TGGGCCCGTC
37151 CCTGAGGTAT GAAAGCCCCC TGCTCTGGCT CTGGTTCAGT CTCAATGGGG
37201 GCACTGGGGC TGGAGGGCAG GGGTGGGAGG CTCCAGGGGA GGGGTTCCCT
37251 CCTGCTAGCT GTGGCAGGAG CCACTTCTCT GGTGACCTTG TTGCTGGCGG
37301 TGCCTATCAC TGTCCTGGCT GTGCTGGCCT TAGTGCCCCA GGATCAGGGA
37351 GGACTGGTGA GTGGCTGCAA CAGGCCCTGG TGGAGAGTTG TATCTTGCGG

37401 ATGCTTGGCT CCCTCTGGTT GTGCCTGTGG TCTTTTGCCC CCTCTGGCTC
37451 AGCTGGCTCG GCTGTCCCTG GTGGGGATGT CTTGTCTCTT TGCTGACTCT
37501 CTTTCCATGT TCCTGTGATG TTGTGCTTGT GTCCCGACAT AAGCCCCTTG
37551 TGTCTCCTCT CCTCTTCCCG AGGTACATCT GTTTCTCCGC CCAAGTACCT
37601 ATGCCTTGCT TGTTCTCCCT TCTAAGGAGG TGTGTGTTGG GGATGGTGCT
37651 GGTAGGAGAA ACCCCAGGCC TGCAGCTTGG GTCCACTTTC AGAGGGGTAG
37701 GGGTGACATG AGCTGAATCT GAACTCTGGG CACTGTGACC CCACCCAACC
37751 AGGTAACGGA GACGGCCGAC CCCGGGGCAC AGGCCCAGCA AGGACTGGGT
37801 AAGAGCAGAC TGTCTCTCCT TCCCCGCTTC AGACCCTCAG GGGCTCCCAG
37851 CTCCCTGCTG CGTCCCCAGA TACCTCTTCC TCTAGGAATC CAGGCTCCCC
37901 ATCCCTGCGC CCTGTTCTCT CAAGGGTAGC CTGCATGGGT GGCTGCCCTG
37951 CCCCCAATCG TGGACTCTTT GCCCCTTCCA GGGTTTCAGA AGCTGCCAGA
38001 GGAGGAGCCA GAAACAGATC TCAGCCCCGG GCTCCCAGCT GCCCACCTCA
38051 TAGGTAAGGA CCTCCAAGAC CTGAATAAGA GTGTAAATAA TCCGAAGGTT
38101 CCAGTTCTGC TCGCCCAGAG TCCTTCGGCT CCATGATTCC AGTGCTCGGT
38151 TTCCCACCCG CTTCACGACC TTTTGTCGCT CGTGCCCACT CTTACGCTCG
38201 TCCCCGCAGT GTAGTTTCTT CTTCCCTCCG GTGCAAGCAA AAGCCGGCCT
38251 GGAGGTCCCC ACTACAGCGT TCTGCACCCC ACATCCGTGT TCCCTCGGCC
38301 CCCAACTCGC ACTCATCCCA GAAACAGCAC CATCCCTCCT CCCCCGGCCC
38351 GGCTCGGCTC CCGCAGGGGC TAAAAGCCGC CACTTCCCCA GAAGTCCCAA
38401 GCCTTTAGGA TCGCATTCCC AAGAGCGCGT CGGCCCGTGT CTCCGCAGGC
38451 GCTCCGCTGA AGGGGCAGGG GCTAGGCTGG GAGACGACGA AGGAACAGGC
38501 GTTTCTGACG AGCGGGACGC AGTTCTCGGA CGCCGAGGGG CTGGCGCTCC
38551 CGCAGGACGG CCTCTATTAC CTCTACTGTC TCGTCGGCTA CCGGGGCCGG
38601 GCGCCCCCTG GCGGCGGGGA CCCCCAGGGC CGCTCGGTCA CGCTGCGCAG
38651 CTCTCTGTAC CGGGCGGGGG GCGCCTACGG GCCGGGCACT CCCGAGCTGC
38701 TGCTCGAGGG CGCCGAGACG GTGACTCCAG TGCTGGACCC GGCCAGGAGA
38751 CAAGGGTACG GGCCTCTCTG GTACACGAGC GTGGGGTTCG GCGGCCTGGT
38801 GCAGCTCCGG AGGGGCGAGA GGGTGTACGT CAACATCAGT CACCCCGATA
38851 TGGTGGACTT CGCGAGAGGG AAGACCTTCT TTGGGGCCGT GATGGTGGGG
38901 TGAGGGAATA TGAGTGCGTG GTGCGAGTGC GTGAATATTG GGGGCCCGGA
38951 CGCCCAGGAC CCCATGGCAG TGGGAAAAAT GTAGGAGACT GTTTGGAAAT
39001 TGATTTTGAA CCTGATGAAA ATAAAGAATG GAAAGCTTCA GTGCTGCCGA
39051 TAAAGATGCT GAGTTGCGAC ACACGTCTTA ATTCAGGGTG GGTGCACGGG
39101 TGCGGGTTAA ATATTCTCAG TACTCTTCTG GTTGCTTGAA ACAATTCATC
39151 ACAACACAGT GTATGGCCTT TGCTCCTAGG GATGATGGTC TGCCTGTCCC
39201 ACCCCCTCCC TGCCTCTGAA TGGCCAGGCC CCACCATTAG CCCAGTTGGA
39251 GGGTGGGAGG AAGGGGGACT TCTCAAACTC CGAAGCTTCT CTAGGCATCC
39301 TGATTTTCAG GGCCACATGG TCCCAACCAG ACTCTGCACC ATACTCTTTT
39351 CTCTTGGGTA CCCCCCAACA GTGAGAGGGG TCATTACAGA GCCCAGCAAG
39401 CACCACTCAG AAAGGCCCAG CAGCAGAGTA AGCCCCTATC ATGACAGAGG
39451 AATGAAGCCT GGAGGGGCCC CGCACTTCTC CCCCTAGAGC TGCCTGAAGG
39501 CCTCTCTGTC TCCTACCCGA CAGTCAACTC TTCTCCTCCA AGGAGCTTAA
39551 TTCAAGGCTC ATGGGGTCTG AAGGGAGGAG GCTGAAGGAG AAAGAAGGGG
39601 AGAATATTAG AGAGAGATGG GGATGGCAGG AAGGAGCCTG TGGTGCCTGA
39651 AAACACCAGG AAGTTCTGGG GAGGAGGAAA AACCGATGCC CCACTTAGGG
39701 TGTCCCATTT AGGGTGAGAC GGAAAATCCT CACCTTTTTT TCACACTTTA
39751 GGTCCCCCTT CCCAAAAGTG AGTAAGTGTG GGTGCTTCTG GGATGAGTAA
39801 CAGTGTCCCC CATTACTTCA TGGCTGACTT TCAGCCACAG GCTGGAGGAG
39851 GCAGAGGGTG ACCCAAGGCC CTATCTAGGT CACCCCAATG GGTCACCCTA
39901 CCCCCTCAGC CTACCACATG GTTTTCTCCT GCCTGGCACC CCAGGGCTGG
39951 AGGTAAAGCC TAATTTCCGA ACTCAGTGGG GGCTCCCAGT CTAGGGGGGC
40001 TCAATTTCCG TCTCCATATT TGTTTTTGGA ATTATTATTT TTTTGAGACA
40051 GGGTCTCGTT CTGTCACCCA GACGGGGGTA CAGTGGCATG ATCATAGCTT
40101 ACTGTAACCT CAAACTCCTG GGCTTGAGTG ATCCTCCTGC CTCAGCCTCC
40151 TGAGGAGCTA GGATTACAGG CATGCACCAC TACACCTGAC TAATCTTTAA
40201 TTTTTTTTCT AGAAACAAGG TCTTGCTATG TTGCACAGGC TGGTCTTGAA
40251 CTAGTGGGCT CAAGTGGTCC TCCCACCTCA GCCTCCCAAA GTGTTGGGAT
40301 AACAGGCATG AGCCACTGCG CCCCACCCTT ATTTGTCTTT GACTCTCTCC
40351 AGAAGAGCCT TCATCCAGGG AGGGGGTGCT TTTCTCTTTC CGGATTACCC
40401 ACCTCTCACC TCTCCCCTCC TTCACCACAA AGACCAGTGG GACCAAGCCG
40451 GCATGTGAGT CCTTCACCCA CATCTTATTC CTATGTTTCA TTCTTTTTTA
40501 AAAAATAGAG ACAGGATCTC ACTATGTTGC CCAGGTTGCT CTGGAACTCC
40551 TGGGTTCAAG CGATCCTCTC ACCTTGGCCT TGCAAAGTGG TAGGATTACA
40601 GGTGCATGCC ACCACGTCCG GCAGTTCGGT TCCTTGTTCT TTATTGTCCT
40651 CAGTCTCTTC GATTTCACCC ACTGAGAGAA TGGAAGGGGA TAGAACAGCT
40701 GGAAACTGGT TGAAGGAAGC CAGAATTCAC TAAGTGCCCA CTGTGCCAAG
40751 GGCTGAGTGA GGTCCTCTGA TGGAGGTCAG GCCTTCTCTC ACATGCCCTA
40801 TGTGTGGTGG ACATTCCTAT CCCCATTGGA TAGATAGGTT AAGTGGCTGG
40851 TTCAGGTTGC AGAGTTAGGA CAGGGTGATT TGAAGCCTAG ACACCCGAAT
40901 CTCTGGAAGT CCCTTGGCTG TGTGATTCAG GTACCTGAGA ATGCGGCTCC
40951 TCTCCAGCTC TCTCCGGACT GCTGGCCAGC TGCAACAGCC GGAAATCTCA
41001 CCTGAGCTGC AGGATTTTCC CAGCAAGGAT TGGAATTCCC AGAGTTGGAA
41051 ATTCCCATGC CCTGAGGGAG AGGTAATTAG GTTCAGGCTC TTGTTTCCTG
41101 GGGGATGGGG AATATTCTGT TGGGCTTTGT TTATGTAGGG TCTCCAGGGC
41151 CCTAGGAGTC TAAGGATGGG ACTGGGTCCG AGGGATCTTA AAGCCTGTGG
41201 AGAGAGGACT TAGGGAGCTT CTTCCCACCC ACAAGAAGAG GCAGATGCAG
41251 AATTAATTCC AAGAAGGAGA CCATGTTTCT TTTCTAAGCA AACTTTATTT
41301 CTCGCCACTG AATAGTAGGG CGATTACAGA CACAACTCCC CTGGGGAGCA
41351 GAGGCTCAGC AATGAGTGAC AGTTGGTCAC CAAATCAGCA TTGTTTAGAC
41401 AACTTAATCA GATAAATATT TTAAAAAACA TAATCAAAAG AAGGCACAGA
41451 GGCCAGGGGG CTACATGGGA ACAGCCTATT GTTCAGCTCC GTTTTCACGG
41501 AAAACATGTC TGAGCCAAGG CAGCTCCTAC ATTGGGTCCC CCAGGATACC
41551 CCGGTCTCCC AAATAAATAC ATTCATCTGT AAATAAATAA ATAATAAATA
41601 AATAATAAAT AATCACAAGT GCAAACATAA ATAGAGGGAG CTGGCTCCAT
41651 GGGGAGGGCT GGGCTCCGTG TCTCAAGGAA GTCTGGAAAC ATCTGGAGAG
41701 AGGAAGGCCT AAGGTCCACT TGTGTCAATT TCTAGGTGAG GTCTTCTCAA
41751 GTCCTGCAGC ATTCTGGCCA GAACCAAAGG CTCCCTGGTC TCCAGATTCC
41801 AGATGTCAGG GATCAAAGCT GTAGGCCCCA GTGAGTTCTG GAGGCCCCAG
41851 TTTGAATTCT TAGTGGTTGC CAGCACTTCA CTGTGCAGGC CACACATTCC
41901 TGAATCCCAG GTTTCGAAGT GGTGGTCTTG TTGCTTAAAG TTCTAAGCTT
41951 GGGTTCCGAC CCTAAGCCCC CAATTCTCTT TTTGAGCCAG AAGAGGTTGA
42001 GGGTGTCTGA AGGAGGGGGT AATAAAGGGA TTGGGGCAGG GGAGGCGTTT
42051 GGGAAGGTTG GATGTTCGTC CTCCTCACAG GGCAATGATC CCAAAGTAGA
42101 CCTGCCCAGA CTCGGCAAAG TCGAGATAGT CGGGCCGATT GATCTCAGCG
42151 CTGAGTCGGT CACCCTTCTC CAGCCTGGAG ACCCCTCCCA GATAGATGGG
42201 CTCATACCAG GGCTTGGCCT CAGCCCCCTC TGGGGTCTCC CTCTGGCAGG
```

```
42251 GGCTCTTGAT GGCAGAGAGG AGGTTGACCT TGGTCTGGTA GGAGACGGCG
42301 ATGCGGCTGA TGGTGTGGGT GAGGAGCACA TGGGTGGAGG GGCAGCCTTG
42351 GCCCTTGAAG AGGACCTGGG AGTAGATGAG GTACAGGCCC TCTGATGGCA
42401 CCACCAGCTG GTTATCTCTC AGCTCCACGC CATTGGCCAG GAGGGCATTG
42451 GCCCGGCGGT TCAGCCACTG GAGCTGCCCC TCAGCTTGAG GGTTTGCTGG
42501 AGGGAGGGAG AGAGGGAGAG GAGAGTCAGT GTGGCCATGT CGGTTCACTC
42551 TCCACATCCT GGCCCTCGAG CTCTGCCCAC CCCACATCCG GTTCCTGTCC
42601 TCTCTGTCTG TCATCCCACA TCCCACCTGG CCATGACGTT CTGAGTATCC
42651 CACTAAGGCC TGTGCTGTTC CTCCACCCTT CCCTTGAGCT CAGCGAGTCC
42701 TTCTCACATT GTCTCCAAGT TCTGCCTACC ATCAGCCGGG CTTCAATCCC
42751 CAAATCCTAG CCCTCCAAGT TCCAAGACAC ATCCTCAGAG CTCTTACCTA
42801 CAACATGGGC TACAGGCTTG TCACTCGGGG TTCGAGAAGA TGATCCTGAA
42851 GAGGAGAGAG AAAAGAAAAA GCTGAGACCC TTAAACTTCC TAGAAAATAC
42901 CCCCCTACTT TCACCTCCAT CCATCCTCCC CCAAGACCAA AACTTTAAAT
42951 TTCCCCCACT GCTTCCATAC CGGTACTAAC CCTACCCCCA AACCCAAACC
43001 CAGAATTAGG AAAGAGGTTT GGAGACACTT ACTGACTGCC TGGGCCAGAG
43051 GGCTGATTAG AGAGAGGTCC CTGGGGAACT GTTGGGGAGA AGGAGAATGG
43101 TTAACATCGA GGGAGTCACC CTTAAAGGAG GAACAGCTGG CTGCCTGTCT
43151 GGCCTGCGCT CTTAGCCCTG AGGTGTCTGG TTTTCTCTCT CCATTCATCT
43201 GTGTATTCAC CTTCCAGGCA TTCAACAGCT CTTTCCCTGA GTGTCTTCTG
43251 TGTGCCAGAC ACCCTATCTT CTTCTCTCCT TATCTCCCCC ATCTCTCTCC
43301 TTAGCTGTCA TATTTCCCGC TCTTTCTGTC TCACCATCTT TATTCATATC
43351 ACTTGTTTCT TCCCCCATCT CTCTTCTCAC ACCCCACATC TGTCTCCATA
43401 TCTTATTTAT ATATCTGCTT GTTCATTCAT TCATTCATTC ATTCACTCCA
43451 TACACACTTA GTGAGCACCT TCCATGTGCC AGACATCCTG TCTCTCCATC
43501 TTTCTCTCTC TCTCTTCCCC ATCTCTTGCC ACATCTCTTT CTGCATCCCC
43551 GTCTTTCTCC ACGTTTTTTT CTCTCCATCC CTCCCTATCA GCGCACATCT
43601 TTCACCCATC CCATCTCTCT CCCTCTCTTG CGTCTCCATT TCCCCTTGGG
43651 TGGGAGAGTG GATGAAGGCT GGCCAGGCAC TCACCTCTTC CCTCTGGGGG
43701 CCGATCACTC CAAAGTGCAG CAGGCAGAAG AGCGTGGTGG CGCCTGCCAC
43751 GATCAGGAAG GAGAAGAGGC TGAGGAACAA GCACCGCCTG GAGCCCTGGG
43801 GCCCCCCTGT CTTCTTGGGG AGCGCCTCCT CGGCCAGCTC CACGTCCCGG
43851 ATCATGCTTT CAGTGCTCAT GGTGTCCTTT CCAGGGGAGA GAGGGTGGAG
43901 CCGTGGGTCA GTATGTGAGA GGAAGAGAAC CTGCCTGGCA GCTTGTCAGG
43951 GGATGTGGCG TCTGAGGGTT GTTTTCAGGG GGGGTCTGTA GTTGCTTCTC
44001 TCCCTCTTAG CTGGTCCTCT GCTGTCCTTG CTGAGGGAGC GTCTGCTGGC
44051 TGGGTGTGCC AACAACTGCC TTTATATGTC CCTGGGGCGA GAGGAGGGCG
44101 GGGAAAGAAT CATTCAACCA GCGGAAAACT TCCTTGGTGG AGAAACCCAT
44151 GAGCTCATCT GGAGGAAGCG GTAGTGGGCC CTGCACCTTC TGTCTCGGTT
44201 TCTTCTCCAT CGCGGGGGCG GGGATTTGGA AAGTTGGGGA CACACAAGCA
44251 TCAAGGATAC CCCTCACACT CCCCATCCTC CCTGCTCCGA TTCCGAGGGG
44301 GGTCTTCTGG GCCACTGACT GATTTGTGTG TAGGACCCTG GAGGCTGAAC
44351 CCCGTCCCCA TGCCCCTCAA AACCTATTGC CTCCATTTCT TTTGGGGACC
44401 AGGTCTGTGG TCTGTTTCCT TCTAACTTCC AGACAGGATG CAGGAAAAAG
44451 ATAGAACTAG AACTGGGAGG GGCTTCAGAA AGCTGAGTCC TTGAGGGAGA
44501 GAAAACGGGG TTGGAGGGAA AAGCTGTGTT GAGTCCTGAG GCCTGTGTTT
44551 GGGTCCCTGC GGGGAGAAGG AGCTGGGGGC TTGGTGGCAG GCTTGAGGCC
44601 TCAGGAAAGG CTGGGTGGGG GTAGCAGGGA CAAGCCTGGG ACAGCCCCGG
44651 GGAGTGAAAT CACCCCCGGG AATTCACAGA CCCCACTGGG GCAGGCCTTC
44701 TTCTTTCATT CTGACCCGGA GACTCATAAT GCTGGTTTCA GTCTTGGCTT
44751 CCAAGGAACT CTGGGGTCCC TGATTTTTTT CATGAAGCTC TCACTTCTCA
44801 GGGCCCCAGT GTGTGGCCAT ATCTTCTTAA ACGTCCCCTG TATTCCATAC
44851 CTGGAGGTCC TGGAGGCTCT TTCACTCCCT GGGGCCCTCT ACATGGCCCT
44901 GTCTTCGTTA AGGGGGGGTC CCCATACTCG ACTTCCATAG CCCTGGACAT
44951 TCTCCTACCC ATTGCTGTGG TCACATCTCC CCAGAGGTCT CCTGTAACCC
45001 ATTCCTCAGA GCCGCTACAT GTGGCCATAT CTCCCAGGAG CTCCCTGACC
45051 CCCGCCCCTC CAGACCCTGA CTTTTCCTTC ATCTTCTCAG CTTCTCCTTT
45101 GCTTCCCCTG CAGCAGTCTG GCGGCCTCAC CTGGTGAGTC CATCACATAT
45151 CCCTGAAGCT CTCTGAGCCC TTATCCTTTT GTTCTCCCCA CAGCTCTTGC
45201 TCCCTTTGAG CCCTCTGTCC CTCCCTCCAT TCCTTAGATA GACTGGGAAG
45251 TTCTCACTCC CAGACACACA CACACAAGCA GACAGCATTT CAGAGAAAAG
45301 AGGTTTATTG GGCTTCATCG AGGGTGCAGA TGCCTCCGTG TGGGGCTCTG
45351 GTCGGCAGCT GGCTTTCAGA GCCTTTCCCT GCCTTCTGGG GCCCTGTGAT
45401 CCCTCATGCC TACTTCTTTC TCTCTTGGTC AGCCTTGTGC GCATGCCCTC
45451 TCACTCTTCA TCTCTTGGGC CTGTCTCTGT TTCTCCTTGG ATGTTCTTCT
45501 ATTATTCCCC TCTCTCCATC CTCCATAAAT AAATAATTTA ATTTTTTTGC
45551 CTTCATAAAT AGTCCCCTCC CTGCCTCTAG TCATCCCCCA AGCTCCTCCA
45601 TGTGCCTGCT CTTCCTCTGT GTGTGGATCT AGGCCCCACC TAGCTGGTGG
45651 GACAGACCAA CAGCTTTGGG CTGGGAATTC CTAGGCAGGC TTGAAATCCT
45701 CAGCCAGACA GACATCAGGG ATGGTTCAGG GAGGTGTGGT CCCCTGGGAT
45751 GCCTAGAATT CCTTCTTTGA AAGCTCCGGT GACTTGATCA GGGAAGACTT
45801 GAGCTGTTGG AATGGCCAAA GGAGAGGTGG TGACGACCCC TGAAATGGTC
45851 AGAATGGAGG CAGAATGGGG AGAAGGTCTT GAAATCAATT ATTTTTTCTT
45901 TCTGGATTTT TCCAAGTTCT ACAGAGCGAA GGCTCCAAAG AAGACAGTAC
45951 TAGGGCTGAG GACTAGGTGG GGGATGCCAT CTGTGTGGGT GGATAGCTGG
46001 TCTCCCTGGG TGAGCTGGAA CGCAGCCCCG TGGTACATCG AGTGCAGCCA
46051 GGGTTCCTGC AGCCCTGGAT ACACCATCTT CTGGGAGCTG AGGAGAGGCA
46101 CATGGAAGGG GTACTGGGAG GAGAAGAGCT GGACCTCATG GGCCAGGTAG
46151 AGTGGGGAGG AGGTGGCCTT GGGAGAGTAG GCTTTCCCAG AGAAGACCAC
46201 CTGGGAGTAG ACGAAGTAGA TGCCACTGGT GGGGACCAGG AGAGAATTGT
46251 TGCTCAAGGA GAAACCATCC TGGAGGAAGG CACGGTCCGT GTTTGCTCTC
46301 CAGAGCAGTG AGTTCTGCTT GCTGGGGTCT CCTAGGAAGA GCCATAGGGG
46351 ATGGGGGTGG GAGATCAGGG GTCTGGATCA GAGGTCTCAA TCCCTGAGGA
46401 AGTGGGCACT GAACAACTGA GTTCCTGGGG GATGGCAGGG GGAGGCATAG
46451 GAGTGGGCTC CCTCTGTTTT TTTTAGCGTG GGGGAAGTTG GGGGAGAGGG
46501 GTGGATGCTT GGGTTCCTGA GGCAGGGGTA GGAGGAGAGC TGGTGGGGAC
46551 ATGTCTGGGA GGTCAGGTGG ATGTTTACCA ATGAGGTGAG CAGCAGGTTT
46601 GAGGGTGCTG TGGGCAAGAT GCATCTTGGG GTGCTGACGG GCAGTCTGGG
46651 CAGCTGAAGG TGTGAGGCCA ACACCAGGGA GCCCCTAGGG GAGAACAGAG
46701 TTGAGGGGGG CTCTAGGGCT CAAGGTTTGG CTGAGCCACC CCAGCAGCCC
46751 CCATTCTCCT GCTGCCTCAC CTGGGCCCCA GGCAGCAGAA CCAGCAGCAG
46801 CCCCAGAAGG AGGAGGTGTA GGGTGGTGCC ACACACCCTT GGGAGGAAGA
46851 GACGTTCAGG TGGTGTCATG GGGAGAACCT GCAGAGAAAG AGAGAGAGAG
46901 AGAGAGACAG TGAGCGGGGC GGGGCACGCG GCGGAAGACA GACCTCCCGC
46951 CCTGGGAGAG AGCACCCCCC GACCCCCGAG AGAGAGATCG ACAGAGAAGG
47001 GGACAAGATG CAGTCAGAGA AACCCCAAGG TGAGCAGAGG GAGACAGAGA
47051 GAGACAGGAA GGGAACAGAG AGGAATCATG GCAGAAACAG AGAATGTGTG
47101 ACAGAGACAA TGAGACTGAC AGATGGAGAG TCAGAGACAG AGAAGGAAAC
47151 CAAAACCAAA CCCACCAAGG CCCAGGCCCA GGCAGGCCGG GGATCCAGGC
47201 AGCAGGTGCA GGAGGGACCG AGGCCCAGGC AGAGGGCAGG ACACTGCTGG
47251 GCGGTAGTCC AAAGCACGAA GCACGGGCAG CCCAAGGAGA TGGGGCAGGA
47301 GAGCCTCACC TGCTGTGCGG AGCCCCTGGG CCCGGACGCT CAGGTCCCTT
47351 TATAGAGGAA GCGGCAGTGG CAGCGTGGCA GGCAGCGGGC GGGTTCTAGG
47401 TCGGGGCTGG GGCCCGGGGA AGCCCCCAGG GCTTAGAAGA TACTGCTGTT
47451 TCAGTCAAAG GCAGGAAAGG CTGAGGCCTA GGAGAGAACC ACAGGCTGGG
47501 GGTTCAGGCG ACTGAGTTCT GGGAAAGGGA GTCGGGTCAG GGGAATCGTG
47551 GGCTGGGAGG GCCAGGGAGT GGGGTCAGGC CTAGAGTTCC AAAGAAGGGA
47601 CAGTCAATTC AGAGAGGAGG TGGTTGAGCA GCTGGGGTGT GAGCTGGAGG
47651 CCCGGTTCCC TGAAGAGCAA TCATATATAA CATCTCTGCA CCCTTGGCTG
47701 AGTACAGGCT TCTCTCTTTG CCCATTTCCT TCTCTTGTAC CCCTGTCCTT
47751 GTCCCAAACA ACTCAAATCA TACTTGTCCC AGTATACGGA CTTTCCAGCC
47801 CATCTGGCAG GTTTCACATC AAGAAGGTCC ATTATATATC CCCTTCATCG
47851 GGGACATTCT GGTGTTTGCC TCTTGTCCAG GTGAAAATAT TAATGGATCC
47901 CCCTTCCACT CTTGAAAGTG TCCTAGTTTG GAAGATAAAT TTTTTGGTCC
47951 CCTTACTCAG AGATGATGCT GGACAGTCAA GTATGATGAG GTCTCATCTC
48001 ACTCCTGAAG GATGCCCTCC ATCTCTTCCT GACTTCAGGT GGCTTCCACA
48051 GAACAGATTT ATATAACCCC AAATAAACAC ACATTCCAGG ATCAATGTGG
48101 CAGACACCTT CCAAAGTTTT CTACAAAGGA AGTTTCAGAA TTCCACATGG
48151 GTGAGGCTTA AGGGTGGTGA CTTAGGGTGG GGTGGGGACA GCTAAGGGAC
48201 TTGTTCTGAA GCTGCATTTG CAGAGCCAAT ACATTATTTT AAAATTGTTC
48251 CAGGCCTGGT GCAGTGGCTC ACACCTGTAA TCCTGGCACT TTGGGAGTCT
48301 GAGGTGGGCG GATTACTTGA GGTCAGGAGT TTGAGACCAG CTGGCCAACA
48351 TGGTGAAACC CCATCTCTAT AATAAATACA AAAATTAGCC AGGCGTGCTG
48401 GTGCGCATCT GCAGTCCCAG CTACTCGGGA GGCTGAGGCT GGAGAATTAT
48451 TTGAACCTGG GAGGCGGAGG CTGCAGTGAG TCAAGATCGC ACCACTGTAT
48501 TCCAGCCTGG GGGACAGAGC AAGACTCTGT CTCAAAAAAA AAAAAAAAAA
48551 GTTTCCAAAA TTTTTTTCCT TTTTATTGAT ATGTAACTTA GGTATGAGGT
48601 GGACACCTCT TAAGTGTACA GCTGATGAAT TTTTCCATCT GTATGTAGCC
48651 ACCACCCAGC TCCACGTATT TTCAGGTCCC CAGCAGGTTC CCTCATGCCC
48701 CCTCCCTGCT GATACCCTCC AAAGATAACC AACCACTCTC TCACTTTTAT
48751 CACCATAGAT TATTTCCTCC TGGTTTGGGG CATCATATAA ATGAAATCAC
48801 ACAGAATGTA CTCATTTCTG TTTGACTTCT TTCAGTCAGC ATTATGTTTG
48851 TGAGATTCAT ATACGTGGTT GTATGTATCA GTAATGTTTT TAAAAAAAAT
48901 TATGATGTAA TATTCTATTG TATCAATATA TTCTAATATA TTCTGTTGAT
48951 GGGGATTTGG TTTGTTTCTA GTTTTTGTCT AACACAAAAA ATGTCTAACA
49001 CGAACATTCT CATACATGAG CCCATTTTCA CTTAGCTTGC AAGTTACAAG
49051 GTTTAAAAAA AGCAGTTTCT AAAGATGACT TTATTCACTT TTCCTAAGTT
49101 TGAGATAATA CCAACTTGTC ACCTCAAATA TTATTACTGC TACTGATGTG
49151 ATTTTACTTG GAGAGTGTTA GAGGGGTTGG AGCTGGGGCT GGTGGTGACC
49201 CGGGGTAAAG CCCACTGTTG CATGGGGCAG ACCACTTCTC TCCCCAGGCA
49251 CAAGGTCCCT GAGGGGTCTT GGTGTAACAT GGAGGGACAT GTAAGTAACA
49301 TTCTGGGTGT GTATGAGCTA TTTCTCCTGT TCTTCTCTAC TTGAGGACCT
49351 GCCCCTTTGC CTTTTATGCT TTACTAGTCT TAGCTATTAT TTCTGTAGGG
49401 GCAGAAAGGG GTGATCCCTT CCTGACCCAT CATAAGGGTT ATGGCCAATA
49451 CTCCTATAAT AAAAGACAGA TTAACAAGAG AAAAGCATAA TACATTTATT
49501 TAATCAAAGT TTTAGGTGAC ATGGGAGCCT TCAGAAATGA AGACCCAAGG
49551 ACCCAGGGGA AAACTATTTT TATGCTTAGA TTTGATGAAG AATGAACAGC
49601 TGTGCAGAAA TGTAATTGAA CAAAAGGAGT ATCATCTAAT GGTCACAGAC
49651 TGGGACTGGG GGGATCCCAG CAAGGCCTGG CCATATTCTT CTTGGTCTCT
49701 CTGTACAGCA TTCCTTCCTC CCAGGTATAG GGCAGAACCT CTTCTGGAAT
49751 GAGGGTCTTA TAACCTACTA TCAGATGAGA TAGGTCAGAA AATTTCTTTT
49801 CCTTTTTTTT TTTTTTTGAG ACAGTTTCTC ACTGTCGCAC AGGCTGGAGT
49851 GCAGTGGCAC GATCTTGGCT CACTGCAACC TCTGCCTCCC AGGTTCAAGC
49901 TATTCTCCTG CCTCAGCCTC CCGAGTAGCT GGGATTACAG GCACACGCCA
49951 CCAAGCCCGC CAAATTTTTT TTTGTATTTT TAGTAGAGAC GGGGTCTCAC
50001 CATGTTGGCC AGGTTGGTCT TGAATTCCTG ACCACAGGTG ACCCACCAGC
50051 CTTGGCCTCC CAAAGTGCTG GGATTATGGG CGTGAGCCAC TGCGCCCAAC
50101 CTTCTTCTCA TTCTTTTAAC CTTATTATCT CTTGTTGTCA TGTGGGTTTC
50151 CCTTTTAGCC CCTGCTCCTT TCTTTTTCTC TGTGTTGCCC TTTCTCTCAG
50201 GGTCCTTTTT GCTTCCTGTT GTCTCTTTCT GCTTCTCTAA TGGTATGAGC
50251 TGATGGACTG GGACCCCAGC TGAGCTATAT TAAAATATAA AATGTTATTA
50301 CAAGGCCAGG AGCAGTGGCA CATGCCTGTC ATCCCAGCAC TTTGGGAGGC
50351 TGAGGCGAGC AGATCACAAG GTCAGGAGAT AGAGACAATC CTGGCTAATA
50401 CGGTGAAACT CCATCACTGC TAAAAATACA AAAAATTAGC CGAGCATGGT
50451 GGCACGCGCC TGTAATCCTA GCTACTAGGG AAGCTGAGGC AGGAGAACTG
50501 CTTGAACCCA GGAGGCGGAG GTTGCAGTGA GCCGAGATCG TGCCACTGCC
50551 CTCCAGCCTG GGCAACAAAG TGAGACTCCA TTTCAAACAA ACAAACCAAA
50601 AAACAAAACA AAACAAAACA AGCAAACAAA AAAAGGTATT GCAATTAACA
50651 GTGAGACACA GAGAGAAATT TAAATTAAAG AGGAAGAATG GGACATTGAA
50701 AGACAAAAAA GGGAAGGCAA GAAGGGTGAT GGGGAGACAT GAGAGACACA
50751 GAGGAAGGAA GGGTAAGACT GGGCTGAGGC TCAGTGTCAC GTGCATGTGA
50801 GATATGCGAA GGATGCTCCT TGAGATGGGC CAATCTTGGT TTCAATCTCA
50851 GTTTCGGAGG TTGTATGAAT TTGGTTTCTT TCTTGGGCAG GCCAGCAGTT
50901 GGTTTGGGAC TTTCCCTGGG TGGGAGAGCT GATGACTGGA GTCTTGTGCC
50951 CCAGACTCAG GGAAATACAG TCTTTATAGT GGTCTTTGTG GAGAAACTAG
51001 TGAAATCTCT GAAGCCTCCA AATGAGACTG AAATGACATT AGCTTCAAAC
51051 TTGAACTTAG CCTCAAAACC TGAATTGGGA TTTAATACCA ACATCAACCC
51101 TAACCCAAAT TTAACCTCAA CCCAAATCAC AACTCAAACT CAACCCCAAC
51151 TGTAACCCTA ACCTTAAATC TAAACACATC CCAATTAATA ACCCCCTAAA
51201 TAAAACTTCT CCTCTACCCC AACCCAACCC TGTTTCTAGG GCTAATCTTG
51251 AAACCAGTTT ACCACCACTC CTAACACTAA ACTTAAATCT GACTCTAAAT
51301 GTAAGTCCAA TCTGAGCCAC AAGCCTAAAG TTGAACTTTA TCCTGCTTTA
51351 TGAATTATTC ATCCATTCCT CCATTTAGTG AGTATCTGCG TGCCTAACAC
51401 ATGCTGGGCA TTGTCCTAAG GCAGGAGGGA CATGGAGGCA AAGGGATCAG
51451 AGAAGGTACC AGCACCTGTG GAGCTTGTAT TCCAGTGAGG CCAGACGGAA
51501 AAGAAAGAAA CTGAAGAAGA AATTGGTACT ATGAAAAAAT AAGACAGGCT
51551 GATGTTGTAA GAGTGGCAGG GAGCTACTTT TAAATACAGT AGTCAGCAAA
51601 ATCCTCTTTG AGTGTTTGGG TGGCACTGGA GCTGAGACCC AAATGACAAA
51651 AAATAGTGAC CAGGTAAAAG TTTGGGAGCA AAGCATTTCA GGTAAAGGGA
51701 GCAGCTACTG CAAAGGCTGG AAGGCGGAAC CAAGCTGGGG GTGTTGACGA
51751 CAAACAGAAG GCCAGTGTGG CTGGAGCAGA GAGAGAGACT GGGAGGCGGG
51801 TGGGAGATGA GGTCAGAGAG GAGGGCAGGG GCCAGGTCAT GCAGGGCCAT
51851 GCAAGAAGGG TAAAGCCTCT AGATTTCATC CAGCCACAGG AAGCCTTTAA
51901 AGGTCGTCAG AGTGTGTGGT GCGTGCATGC GTGCGTGTGT GTGTGTGTGT
```

```
51951 GTGTGTGTGT GTGTGTTGCA GGGGAGAGAG GGGGAGGGAG AGAGAGAGAG
52001 AGAGAGAAAG AGGGAGGTGA GCAGAGGTGA TTGGATTTTT TTTTCTTTTG
52051 ACATGGTGTC TTGCTCTGTG GCCTAGGCTG GAGTGCAGTG GCACCATCAT
52101 AGCCCACTGC AACCTCAAAA CCATGGGCTC AAGTCATCCT TCCACCTCAG
52151 CTTCCCAAGT ATCTAGGACT ACAGGTGTGT GCCACTGTGC CTGGCTAATT
52201 TTAAAAAATA TTTTAAAATT TTTGTTGAGA CAGGGTCTAT GCTGCTCAGG
52251 CTGGTCTCGA ACTCCTGGTT TCAAGTGATC TGCCCATCTT GGCCTCCCAA
52301 AGTTTTTTTT TGTTTGTTTG AGAGGCGGTT TCGCTCGTTG CCCAGGCTGG
52351 AGTGCAATGA CTGATCTCAT CTCACTGCAA CCTCTGCCTC CTGGGTTCAA
52401 GCGATTCTCC TGCTTCAGCC TCCCAAGTAG CTGGGATTAC AGGTGCATGC
52451 CACCATTCCC GGCTAATTTT TTGTATTTAG TAGAGATGGG GTTTCACCAT
52501 GTTAGTCAGG CTGATCTCAA ACTCCTGACC TCAGGTGATC CGCCTGCCTC
52551 AGCCTCCCAA AGTTTTGGGA TTACAGGTGT GAGCCACCAT GCTGGGCCAG
52601 CCTCCCAAAG TTTTGGGATT ACAGGCATGA GTCACCACAC TGGCCCTGGA
52651 TTTTTTTTCT TTCTTTTTTT TGGAGACGGA GTCTCACTCT GTTGCCCAGG
52701 CTGGAGTGCA ATGGCGTAAT CTCAGCTCAC TGCAACCTCT GCTGCCCGGG
52751 TTCAAACGAT TCTCCTGTCT TAGCCTCCTG AGTAGCTGGG ATTATAGGTG
52801 CATGCCACCA TGCCTGGCTA ATTTTTGTAC TTTTAGTAGA GAAAGTACAC
52851 CATCTTGGCC AGGCTGGTCT CGAACTCCTG ACCTCAGGTG ATCCACTTGC
52901 GTCGGCCTCC CAAAGTGCTG GGATTACAGG CGTGAGACAC CGCACCCAGC
52951 CTTTTTTTTT TTTTTTTTTC TTTTAAGACA GAATCGCTCT GTCACCCAGG
53001 CTGGAGTGCA GTGGCACAAT CTCGGCTCAC TGCAACCTCT GCCTCCCAGG
53051 TTTAAGCAAT CCACCTATGT CAGTCTCCCA AGTAGCTGGG ATTATAGGTG
53101 CATGTCACCA TGCCTGGCTA ATTTTTGTAC TTTTAGTATA GAAAGTACAC
53151 CATGTTGGCC AGGCTGGTCT TGAACTCCTG ACTTCAAGTG ATCCGCCTGC
53201 CTCAGCCTCC CGAAGTGCTG GAATTACAGA CATGTGCCAC TGCACCCGGC
53251 CTGGTTTTTT TTTTCTAAGA GATGGAGTCT CACTTTTCTG CCCAGGTTGG
53301 AGTGCAATGG CACCATCATA GCTCACTGCA GCCTTCAACT CTTGGCCTCA
53351 GGCAATCCTT GCACCTTAGC CTCGCAAAGT GTTGGGATTA CAGGCATGAG
53401 CCACTGAGCC TTGCCTGGAC TTTTTTTTTT TTTTGAGATG GCGTCTCGCT
53451 CTGTTGCCCA GGTTGGAGTG CTACGGCATG ATCTTGGCTC ACTGCAACTT
53501 CCACCTCCCA GGTTCAAGTG ATTCTCTTGC CTCGGCCCCC CGAGTAGCTG
53551 GGATTACAGG CATGCGCCAC CGTGCCTGGC TAATTTTGGT ATTTTTAGTA
53601 GAGATAGGGT TTCATCATGT TGGGCAGGCT GGTCTTGAAC TCCTGACCTC
53651 GTGATCCACC CACCTCGGCC TCCCAAAGTG CTGGGATTAT AGGCATAGCC
53701 AACGCGCCCA GCCTGGACTT GTTTTTAAAA GATCACTGTG GCTCCTGTGT
53751 TTAGGCTGGC TGGTAGGAGA CAGGTGGCAG TGGCATTGAT GGTGAAGAGA
53801 AAATAGTGGC AGCCATGGAG ATGGAGAGAA GTAGACAAGT TTGGGATATA
53851 TTATACATTC CAGGGGTAGA AACAACAGGA CTAGATGATG GATTGATGGG
53901 TGGGAGATGT AGATACTGGG AGAGAAGCAG GATTCTGATG GATGGAAAAA
53951 CTAAAAAATT CTATTTTGGG TGTGGTAAGT CTAAGTCTAT TAGACATGCA
54001 AGTAGAGATG CCACTGGGCA GATACACATC TGGATTTCAG GGGCAAGGTC
54051 CAAGCTAGAG AAAGAAACCT GGGCATGGTC AGCATGAGGA TGGTGTTTAA
54101 AGCCATGGAA CTTATCTTGT GCATCCCTAT AAGACCCCTT TGAGGCACTT
54151 GTTTCCCCTC ACAATGGATG CAGTGCATCT TCCATTCTGA ATTCCAGAGG
54201 CAACAACCTC CTGCTCCTAG AAGCTAAACT CTCCAGACTT AGTCTTCTGA
54251 ATTCCCACTG GGATTTAACC TCCCTGGATT CAATTCCCTA CCCCACAAGG
54301 ACCCTTCTAC CAATCCATTT CACAATATTT GGTGGAACTC TTCACATTTT
54351 CAAATCCTGT CTTCTATGTT TGAAAGCATT TTCCTCTGGG CATTTTGAAA
54401 CCTGGCTTTC CCCTAATGAT GCTACTCCTC TCTGTGGTGA AGGTTGCTCT
54451 TTCTTCGACC TCCATCACTA GGGGAATGAA GGCAGAAGAG AGGGGTCGTT
54501 ATTTTCTCTC CCCCAGACCA CTGATAAACT CCTCCTCATC TCCCAAAGCA
54551 CGTGCCTTTT GCATTTGCCA TGCTTTTCCT GTCTTCACAG CACAATCTTA
54601 TCCATCAAAA TATTTATTAT GCTCTTAACA ACTTGTATAA CTAGCTCAGT
54651 CTTTCTCCAC ATCTTATCCC TACCATCATC CTTGTTTGTT TGTTTTTTGA
54701 GACAAGGTAT TGCTCTGTCA CCCAGGCTGG AGTGCAGTGG CGTGATCTTG
54751 GCTCACTGCA GCCTCGACCT CCTGGGCTCA AGTGATCCTC CCACGTCAGC
54801 CTCCTGAGTA GCTAGGACTA CAGGCGCATA CCACCGTGCC CGGCTAAGTT
54851 TTGTATTTTT TTTTTTTTTC CAGAGATAGG GTTTTGCTAT GTTGCCTAGG
54901 CTGGTCTTGA ACTCCTGAGC TCAAGCAATC CACCCGCCTC AGCCTCCCAA
54951 AGTGCTGGGA TTACAGGCAT GACCCACCGC ACCCGGCCTC CTACCACCAT
55001 CTTTAGTGAC TTTAAAAGCC ACAAAGATCG GCTGGGCGCA GTGGCTCACA
55051 CTTGTAATCC CAGCACTTTG GGAGGCTGAG GCGGGTGGAT CACGAGGTCA
55101 GGAGATCGAG ACCATCCTGG CTAACAAGGT GAAACCCCGT CTCTACTAAA
55151 AATACAAAAA CAAAATTAGC TGGGCGTGGT GGTGGGTGCC TGTAGCCCCA
55201 GCTACTCGGA AGGCTGAGGC GAGAGAATGG CATTAACCCG GGAGGCGGAT
55251 CTTGCAGTGA GCCCAGATCA CACCACTGCA CTCCAGCCTG GGTGACAGAG
55301 CGAGACTTCG TCTCAGAGGA AAAAAATAAA AATAAAAAAA GAAGCCACAC
55351 AGATCAGATG AGTCTTACAG TCTGGCCTCT CAATTCTCCA AGTTTCTCAA
55401 GTCTACTGGT TTTACAGCCT CTTCTTCCCT CCTCCATTTT AGCCACCCAG
55451 GAGTAAGGTT GTACCTCAAA CCCCAGCATC ACTTGGAAGA ATTATGCCTT
55501 CAAGATCTTG ATCTCTGAAT TACCCTTCTC ATCACAACCA CTCACTTTTC
55551 ACATCTCCCA CTCTCTGCCT CCCCTTGAAT CAGTTCTTTA TTCTTACCAA
55601 GACTTCCGGC TGCTTCTGCT TTCCTGGTTT TGCTTGTTGA TTTAACATGT
55651 ACTGCATGGT CAACAATTTC AACTCCACTT TGTCTGCATT TTAGAATCTC
55701 TTACCGCCGG GCACCGTAGC TCACACCTGT AATCCCAGCA CTTTGGGAGG
55751 CCAAGGTGGG CGGATCACCT GAGGTCAGGA GTTCGAGACC AGCCCGGCCA
55801 ATGTGGTGAA GCCCCATCTC TACTAAAAAT ACAAAAATTA GCTGGGCGTG
55851 ATGGTGGGCA ACTGTAATCC CAGCTACTCT GGAGGCTGAG GGAGGAGAAT
55901 CGCTTGAACC CGGGAGGCAG AGGTTGCAGG GAACCAAGAT TGAGCCATTG
55951 CACTCCAGCC TGGGCAACAA GAGTGAAACT CTGTCTCAAA AATTGACTTG
56001 GGTAAACTAA GTGGTCTCAT TCTTGGGGCT TTGTCTTAGG CTGCCAAGCG
56051 CTGCAGTAGA CAGCTGAGTC ACCATAAGAG GCCTATAACT AATCCTGGCC
56101 AGCTAATCTC AGAGAGTCAT CAGTGTTCTT GGACACAACT TTTACTCATT
56151 CTTTGTTGAC TCCCCATTAT ACTACTGTGC GTGATGTTCT AAATCTCCGT
56201 GCCTCTGATG TTCTTCACTC TCATCTATTA ATATTTCATT GAGCTTAAGG
56251 CCATCAAGTG TGGATCTCAC TTCCCTCTTT TTCCTTAATT TTGTTATATT
56301 TAATCTTTCC TTACATACTG ACTCACAGTA CTTTTTCTTG TTTCCAAGAC
56351 TGATCCTTTT ATCTGTACTC TATCCCACTT CCTCTATCTT TACTATATTC
56401 TCTCCCCACT TCTTGCTTAT GTTTTCACTA TCTCTCCCTT TCTTGTTTGT
56451 TTTTGGGATA GGCTTGCTCT GTTGCCCAGG CTGGAGTGCA GTGGCGCGAT
56501 CTCGGCTCAC TGCAACCTCT ACCTCTCGGG TTCAAGTGAT TCTCCTGCCT
56551 CAGCCTCCTG AGTAGCTGGG GTTACAGGCG TGTGCCACCA CGCCCAGCTA
56601 ATTTTTGTAT TTTTAGTAAA GACAGCATTT CACTATGTTG GCTAGGTTGG
56651 TTTTGAACTC CTGATATCTA TAGTCTGCCC ACCTCAGCCT CCCAAAGCGC
56701 TGGGATTACA AGGGTGAGCC ACCGCACCTG GGCTTTTTTT TTTTTTTTTT
56751 TTGAGACTGA GTCTCGCTCT GTCACCCAGG CTGGAGTGCA GTGGTGTGAT

56801 CTCGGCTCAC TGCACCCTCT GCCTCCTGAG TTCAAGCGAT TCTCCCACCT
56851 CAGTGAATAG CTGGGACCAC AGGCATGTGC CACCACGCCT GGCTAATTTT
56901 TGTATCTTTC GTAGAGATGC GGTTTTGCCA TGTTGGCCAT ACTAGTCTTG
56951 AACCCCTGGC CTCAAGTGAT CTGCCCACCT TGGCCTCCCA CAGTGCTGGG
57001 ATTACAGGTG TGAGCCACTG CACCTGGCCT ATCTTTGTCT CTTTAAGCAT
57051 CTTCATTTTA CCTTCAATTT TATTTCTTGA AAATAAACAA CAAAAAAGAT
57101 CCTTCCTTGG AATATTTCTC CCTCAAATGT CCAGTCTATC TCACCTTCCG
57151 TTTTCGTGTG ATGTCTCAAA AGACTAGCCC CCTTAACGAA CCACTTTTTC
57201 AGGAACTGGT AAAATTTCCC TCAACTGTAA AATCTGTGGG TAGAACTAAG
57251 AGATGATTCT CAGTTATTCT GTGTTCACCA TTCTGACCTC ATGCCCACCA
57301 CCTCCTTTTC CTCACCTCCC ACTCATTCCT CAACCTATGT AAAGTGATTT
57351 CTGCCCCTCT ATTAAATGGA AACGGTTTTC TCAAAATTTG TAATCGTCTA
57401 GTTCCTGAAC CCAACAATCT TTTCAACCCT CATTTACTAG ACCTCTTTGA
57451 AGCCTCAAAC ATGGCTGCCC TGTCTCTGCT TGGCATCCCC CGCCCCCCGT
57501 GCCCTGCTTC TCTGACTGCT CCCCAACCTC TCCTTTATTG GTTCTCTTCC
57551 TTTGATCATG CTCAGGGCTC TTCTGGTCAC TCCGCCCCCT CCCTTGAAGA
57601 TCCTGTCCAC TCAAAGGCTC AACCTCTACT GGGAAGGGAA TGATGCCCAA
57651 ATCTGTTCCT CCAGACCCAA CTTCTCTTGA GCTCCAGACC TGCTTTCTGG
57701 AAATCTCCAC CTGAGAGTCA TGCTGGTACC TCAAACCTTC ATTGAGCACT
57751 TTATTACCCT AAGCAGGCAT ACCTATACAT TATAACTATA ATGTATAGTT
57801 ATTATTTGTA TACTGTCTCA TTAGACTGCA AGTTCTATGA AGGCAGTGAC
57851 TTTATTATTT ATACCTATTG TCAAGCAGTG CCCTGCTCCT AATAGGTGCT
57901 TAATAATTTT TTTTTTTTTT GGATACAGAG TCTTGCTCTG TCACCCAGAC
57951 TGGAGAGCAG AGGCATGATC TCAGCTCACT GCAACCTCTG CCTCCCGGGT
58001 TCAAGTGATT CTTGTGCCTC AGACACCTGA GTAGCTGGGA TTACAGGTGC
58051 CCAAGACCAC ACCCAGCTAG TTTTTATATT TTTAGTAGAG ACGGGGTTTC
58101 ACCATGTTGG CCAGGCTGGT CTCGAACTCC TGGCCTCAAG TGATCCTCCC
58151 AAAGTGCTCC CACCTTGGCC TCCCAAAGTG CTGTGATTAC AGGAGTGAGG
58201 TGCCTGGCCA CCACCACCCA CCTTTTGGAC AGACGTTGAG AGGATTTGTG
58251 GTGTTAGGAA AAGGGCTGGT TAACACACAG TAGGTATTAA TTCACAAGAT
58301 GATGGCCGTT TTCTTACTCC TTGTCTGTAC TGCAGCACTC ACCTGTTCTC
58351 CCTCTTTTTC ATCGTCACCT GGACCACTGC CCACAGGATG CAGTCCAATC
58401 TCTGACCTTG ACTTTCTGTG GTCCTTTCTT CATCACATGT TCCCTCACAC
58451 CTGTGCTTCT TCTCCCTCAG GACGCTCCTT TCCCAGAACA TACTGCATTT
58501 TCTGACTTGT CAGGCTCGCT CAAGACGTGT TATTCGCCTA TGGAACCTTT
58551 CCATCCACCT CCACCTAGTG AACCCATTCA TTAGAGCAAG TTCAAATGCT
58601 GTCTCCTCAG AGGCTCCCTG GATGCTTTCA AATGGACTTA ATCTCCCCTC
58651 TTTCTGTAAT CATCTTAAAG TTTACTCATC ATGCCTTGAA TCAGTTAACT
58701 AGGTATGGGT CTGCCTCCCC ACTAGATTGT AAGCTTGAAG CAGGGTTTAA
58751 ATTTATCTTT GGGCTACATA ATGCTCAGTT GACCTACACT GAACAGAGTC
58801 CCTCCCAGAT CCTTTTCCTA ACTTAATCCA AACCTGCCCC AACCTGAGTC
58851 CTCAAACTGA CCCTCTTCAA TCTCATTATT CTTTTCCAGG CCTGTTCCTC
58901 CAATTGAATC CCAATGCATC CTCCACTGAG ACTCAACTCC TGGTGCCAGC
58951 TCTAGAGAGC AGAGATACAC TTCCACCTAC CCACTGTTTC ATGGCAGGCA
59001 CTGAGTCACC CCTTGGGAAT CCCTGGCCCC TGTCCCTGCC CCCACCATGC
59051 TGGACAGGAT GCCTGCTGCC TGCCCCAAGG GGTCAGAGGA AGCCAGGGCC
59101 TCTAGGGCTC AGTATTTTTA TCAGCCTAGG AAACAAGAGA CACAGAAGCG
59151 GAAACCATGT GGTCAGAGAA AGTGAACAGC CCTTATCTCA GGGGAACCCC
59201 CACCAGGGCT GGGGGCCAGA CACCAGGAAG GGACGTTGAA AAGACTGTGG
59251 TCCGGTGGGC CCAGGTTTCA GGGCTTATTT TGCTTTGGTT TTATGTAACT
59301 CTTGCAGAAG GGGAAATGAT GTGCTTAAAA ACCAGTGAGA AATCTCCTGG
59351 GTGGAGGAGC AAATTTGGTC ACCAGAGTTT TCTGTAATTT ACTTCTCAAC
59401 CTCCAGCTTA GGGATCCTAA CCCCAACTCC CCCCAAGGCC AAACCTGGAC
59451 TCTTTGCATG CCACTTCCGA AAAGCTCCAG CCTGTGAGAG GAACCGGACG
59501 CCAGGGTCAC ACACAGCTTG AAGTCTCTTC TCTGAATCAC AGCAGCTTCC
59551 TCTCACAGGA TCTTTCTCAC CAGCCCTTCC CCCTCCTGCA CCCTGAACAT
59601 CTGCCTGGTC CTTCAGAGGC CTTGGTCTCT TCCGAACTTC AGGCTGTCTT
59651 ATTCAAGGAT CTCATGCCCA CCCATGTTCC CGTGTTCCCA GAGGCGTGCC
59701 CTCCCTGTCT TCTACTGCCT TCCACTTCCA CTCCTGGGAT TCTCCTGTAG
59751 AGAACTGGGC TCAAAACCTG AAAGATAAGT TTCTGATTGA TAAGATGCTC
59801 TGGTGTAACT ATGCAGACAG TGCACTCACC AGCACATGTC TGTACTCTCC
59851 TTGCAGCCCT GTCCTTTCTT TTTGCCAAGT TCCCAGTTGT GTCCTGGGCG
59901 ATGGTAGTGG GGTCTACCTA TCTTAGTTTC CAGGGAGCCC CCCTCTCTCC
59951 TCCGTGGGAA TCCCAGACTC TCCCTCCACC TCTCAGGACT TCAACTACTG
60001 GTTCTTGGAA AAGGAAATGT TTATAGGGAA GGCCATGCCT GGGCTGCCCA
60051 GAGATGAACT CAGGTTTAGG ATGGGGCAGC AGCTGAATGG CGGAGATGAG
60101 GATTGGGGCC CTGGAGGAAG AGAGATACAA CTGACAAGCT AGAGGGCAGA
60151 GATGGATGGA GCCAGCAGCC TGGCTGGCTT TGGGTGTGGA GCTCCAGTTG
60201 CCCAAGAACA GACAACCCAT GTTACATAAT TTTGCTCTTT CCACCTCCTT
60251 CCCCTGCTCC TAATTTTAGC AACTGAAATT CCACTCCCCA CAGGGCTCTT
60301 CCTCCAGCCA TTCTGCTTGT ATTCCCCATC CTCCCCCAGC CTCCTCATGG
60351 CCCCCCAGCG CCGGCTACCC CCAACTCGCT CATTCAGTGA CTCAGCAAAG
60401 GGAAAGCCCT GTGTTGGGGG AAGGGGGTGT TAAGTGGGGA GGAGGATGTG
60451 GTTCAGCTTG GGGAGCCAGG GAGGGGAGGG AAACTGGTTT TGTTCCCCTT
60501 TTTGCCTCTG TTCGCATCTA TTTCTCCAGC CACAGATTGC AGGGTGGACA
60551 AGACCCCAGT GTAGGGAGTG GAGTGAGACT GGATGGCAAA CTGGGCCTTC
60601 TTAGGGCTGA ACTGAGGGGG CATGCAGAGG CTTCTTGCCT CCAGCTCAGA
60651 GCCCGTGTTA GCGTGTGCGT GTGTGTTGCA GGTGTGTGGG AGCTGTGGGT
60701 TTCAGCAGGC GGGGCGGGAA AGGAGATGTC ATGGGTGATG AAAACCACAG
60751 CAATGGAAAA AGACAGAGGA GACACTGGGA GTGAGATGTG GAACCAGGTC
60801 TTATTAGGAA ACGGCAGAGT CAAGGAGCAG TGGTGGCGGC CCGCACCCCC
60851 CACCCCCACT TGTTACAACT TTCGCTCCCA CCCATATCCT CTGCTGAGCA
60901 GCACCTGGCT CTTCCTTGTG GTCCCCATCC TTTCCCTTAC CCTTCCTTCT
60951 GCCTCATCCT GAGGCTGCAG CCCCGAAGTT TCTTGCTTCT TCCCTAGGGT
61001 CACTTGAGGG CCTCTGCATG GCGATTCAGG GCCTGAGAAA GGGCTGTCAC
61051 TGCTCCCATC ACACGGCCCA GCTCCCAGGT CTCAATCTGG CTTCGGAATC
61101 GCTGCAGGAA GCGGTCAGGG TGCCAGCGGA CCTGCTGGAC CCTCAAGTAC
61151 CTCCTCAGAG CCCCCTGTTC CTCCAAAGGG GGGCCCCTGG CCACCAGGGC
61201 TGCAGCCATG GCCTCTGGGT CCCCTCCCCC AGGGCAGGGC CAGGGCACAT
61251 CACCAAATCG CCAGAGGCTG CCCCTCCCCG CTCCTCTGGG GTGCTCTTCC
61301 CTGGGCCCGG CCCTGGTTGG CTTGGGTTCT CGGTCCCCTA GAGCCTCCTG
61351 CGCCCTCCTG GCTCGGCTCT CACGCAGCTC TTCCTCCTTG GCCCGGGCTC
61401 GCTCCCTGAA GAGCCGCTGC TCCTCCTCCT GCTGTCGCCA GCTCTGGCTG
61451 GAGCCCTCAG CACGTGGGGG TCGACAGGAT CCCTCTGCTT CTCGCTGCTG
61501 CTGCTGGCAC TTCTGGGCAT GTTCCCGGGC CAGGCGATCT GACCAGGCTG
61551 AGAAGGACTC AGGTTCCTGG GTTTCATGGG AGGCATCACC TGTTTGGGAG
61601 GGTGGGATGG GGGTAGTTGG AGAGAGGCTG TGAGAAGGTG ATGGGCCCCA
```

61651 AGAAGCAGGG AAGCACAGAG GAGGGGGCAG AGGAGCCTAG TGTGACGCAT
61701 TCATGCAGAT GGAAAGCAGC CAGTGGGGAA GGGCAGCTGT GGATAGCAGT
61751 GGACTTCTCA CCTTCAAACC TCCCCATGAC TTCCTGCCAC TCGTCCTCCA
61801 GCTCACCCTG GAGCTTCTGT CTCCATTCCC GCTCCTTGGA GGCATCATCT
61851 TCTTCCTCCT CTTCAGCAGA ATCCCAGGGG GGTCCCCAGC CCAAAATTTG
61901 GCCAGGGGTC TCCCCATCCT TATTCTTTAT TCCCATGGCG GAGGGACAGC
61951 GGCTTAGCAG CGGGAGGAAG AAATCGGTGT AGGCTGTTGG TGGGAGAGCA
62001 GGGAGCAGAA GTTAGCTGGG ACTCCCACTC CTCGGCAGAG CTGCCATCTT
62051 GAATCCTGCT GGTCACTCGC TCCCTTGGTC TTAAAGCAAT GGAAGGGTGG
62101 CTCACATGGA TGCCTCCTGG GTATTAGGAA ATAACCATCA CTGATAAATA
62151 CAGACTGATA GTTTACCACA TGCATGGTAA GTGTGGTAGT TTAAAATATT
62201 CTCGGCCAGG CGCGGTGGCT CACGCCTGTA ATCCCAGCAC TTTGGGAGTC
62251 CGAGGCGGGC GGATCACGAG GTCAGGAGAT CAAGACCATC CTGGCTAACA
62301 TGGTGAAATC CCATCTCTAC TAAAAAATAC AAAAAATTAG CCAGGCCTGG
62351 TGGCGGGCGC CTGCAGTCCC AGCTACTCGG CAGGCTGAGG CAGGAGAATG
62401 ATGTGAACCC GGGAGGTGGA GCTTGTAGTA AGCCGAGATC GTGCCACTGC
62451 ACTCCAGCCT AGGTGACAGA GCAAGACTCC GTCTCAAAAA AAAAAAAAAA
62501 AAAATTCTTA CTTGTGTCTC CACAAACTAT TTCAGCGTCC TAGAAGTTCT
62551 ACCATGCCAA CTGACATTTT CTGACCTGCC TCCAGCCACC AAGGCAGTGC
62601 CCAAATTCTT TGACTATGTA TTCCACTTTT CACTTAAAAA TATTGTCACT
62651 ATATAAATCC CTTATTTTCT CACCTGAAAC CATATATCAG ACTGGGAAAA
62701 TATCTGAATG CTCTACTCTG GGCCATTAAT AAATACACAA ATATGTAATA
62751 CTTATCTTTT ATTATCACTT ATCAATAAAC TGAGTAAAAT AAATGTTTTC
62801 AGGGGAATTT CTCTCAGCCA GCCTTACCAG GGGATGATGG GAGAGGGGTG
62851 GGGAGGTGAA CCGGCAACAA CTATGGCCGG CGGCAGAGCA AGCTCTTTCC
62901 AAATGACTGC TGACCTAGGG CAGGGGAAAG GGAGTGGAGT GTGACAGAGG
62951 GTCTCACCCA TGGGCTGAGA GAAAACAGGA GAGGAACCGA CGTTCCTGAA
63001 CTCCCCTTTT CTTCAGTCCC AACCTTGCTG CATCTGGCCC AAGGTTAGCT
63051 GAGTGCCATG CTACTTCCTT CACTGCCAAC CCAGGCATCC TGGCCAGGCC
63101 CACCTGCTGT GGCCACCAAC CACCTCTTTC ACTTGGGGGA TAGAAGAAGG
63151 GGAGGGAGGC AGCCTTCCTT CCTGTGGACC TACTTTCTTT CCCCGGGGTA
63201 AGAGGAAATG GGCTAGCAGT CCTTAAATCT TTATTTGGTA GTGCTGGAAA
63251 GTACTGTTTA CCTGGCAGAA AGCTGGAATA GGGGAAGGCA AGGCCAGGAA
63301 GGCAAGAAGA CAGAATGGCC CAGGTGTGGC TGGCGAAGGC CCACCATCCC
63351 TACCCAAATC ACATCAGGGT TGGTGGGGGG GGGCACTTCT CCCTAGTGCT
63401 GCTGTGACCT GTCACAGACC CTCTCAACTT GTCCCACCCA GAAAGTACCT
63451 GGTCCTGTCT CTCATTCGCT TGTTCCCCAC CTGAGCTCAG GTGGTGAGCA
63501 TGGTGAGTGC TCAGGCTTGC ATGGGAGGTT TACATTCATA GGTTTTAAGG
63551 AGTAGGGCCT CCAACTATAA AAACATAATA TTAAACAGCC ACTACAACTG
63601 AGGCATGTGT TTGAAAAAAG CTGGCTACAA AACTGTAGAG AGGATCAGAT
63651 GTGGCCAGGC ACGGTGGCTC ACGCCTGTAA TCCCAGCACT TTGGGAGGCT
63701 GGGGGGGGGG GACAATGGAT CACAAGGTCA GGAGTTCAAT ACCAGCCTGG
63751 CCAAGATGGT GAAAACCCGT CTTTACTAAA ACAAACAAAC AAAATATATA
63801 TATAATTATA TTTTATATTA TATATAAAAT TAGCCAGGCG TGGTGGCTGA
63851 CGCCTGTAAT CCCAGCTACT TGGGAGGTTG AGGCAGAGAA CTGCTTGAAC
63901 CTGGGAGGCG GAGGTTGCAG TGAGCCGAGA TCGCGCCTCT GCATTCCAGC
63951 CTGGGTGACA GGGTGAGACT CTATCTCAAA AAAAAAAAAA AAAAAAAAAA
64001 AGAGAGAGAG AGAGAGGATC AGATAATAAC TGTCTAAAAC AAGAGACCAA
64051 ATCCTATGGT TGGAAAAAAA GAGGCCGGGC GTGGTGGCTC ATGGCCTGTA
64101 ATCCCAGCAC TTTGGGAGGC TGAGGCGGGT GGATCACCTG AGGTCAGGAG
64151 TTCGAGACCA GCCTGGCCAA CATGGTGAAA TCCTGTCTCT ACTAAAAATA
64201 CAAAAATTAG CCAGGCGTGG TAGCATGTGC CTGTAATCCC AGCTACTTGG
64251 GAGGCTGAGT CAGGAGAATC ACTTGAACCT GGAAGGCAGA GGTTGCAGTG
64301 AGCTGAGATC ATGCCAGTGC ACTCCAGCCT GGGCAACAGA GGGAGACTCC
64351 CATCTCAAAA AAAAAAAAAA AAAAAGAAAA AAAATAGACT GGCAAAAAAT
64401 ATGTAAAAAT TCTAGCCTAT GGGTGGTGGG GCTGTAGGTG GCTTTTCCCC
64451 CAATTTTATT TGACATTTTA ATGTGAAAGG GATATAATTA AGTTGAAATT
64501 TTCTTTTTAA AGAAGGACGA ACAAATTTGA TAACTATTAA CAGTGTACTG
64551 TGATCCCTTC CAGTCCCACC TTGGTGTACC TAAAAGCATA CATATGACAC
64601 ATTTATTTGG AGATATGTAA CCATACTACA AATATTGCTT GGCAACTTGC
64651 TTCTTTCATT TAACATCTCA TTTATATTTT TCCACAGGAA TATATACAGT
64701 CTACCTCATA CTTTTTGACG ACTTTATAAT TAGTGTTCCA CTGTGTGTAC
64751 AAAACAAATC TCCTTACCCC AATATCGATA GACTTTGTTT CAAAAAATGT
64801 TCAGCCTCAT GAGTATCTTA CAGTCTTTCT GTGGGGGTAG ATTTTCATAA
64851 GTGGAATTTC TGAGTCAAAG GACATGCGCA ATTTGGTCAA ACTGCTCTCA
64901 ATAAGTTTGT GCCAATATAC AAGTGTGCCT ATGTCTTCCT TTACCAAAAT
64951 TGAATATCAT CAACCCTTTT AATTTTGCCA GTTGGATAGA TTAAAAAATT
65001 ATTTTATTAA CAATGTTTTT TTCTTTTTCA TGTCTTTTGG CTATTTGTAT
65051 TTTTTGTGAA TGAATTGCTC ATATTCTTTA TCCACTTTTT CTTTGGAAAT
65101 TATATATATT GATATGAGGG TTCTATATAC TGTATGTGTT ATACATATTG
65151 CAAACATTTT CTTCTATCTA TCTTTAAAAA AGCTTTTTTT TTCTTTTTGA
65201 GACAAGGTCT CACCCTGTTG CCCAGGCTGC AGTGCGGTGG CATGATCTTG
65251 CTCACTGCAA ACTCTGCCCC ACCTGGGCTC AAGCGATCCT CCCACCTCAG
65301 CCTCCTGAGT AACTGGGACT ACAGGACGAC AGCCACGTAC CACCACGCCT
65351 GGCTAATTTT GTATTTTTTG TACAGACCAG GTCTCACTGT GTTGTTCAGG
65401 CTGGTCTTGA ACTCCTGGCT CAAGTGATCT GCCTACCTTG GCATCCCAAA
65451 GTGTGGGATT ACAGGTGTGA GCTATCGCAC CCGGCCTAAA ATTTTCATGT
65501 AGTTAAATCT ATATCAATAT ATTATTTTCT GACTTTTTAT TTCACTTCAC
65551 TCTTAGGAAG GCTCTATTTT AAGATTATAA AACATTCTCT GCATTTTCTT
65601 CTACTACTTT AGATGTATAC TCATTTTTAT TTTTAGAAAT TAAAAAACTC
65651 GGCCAGGCGC GGTGGCTCAT GCCTGTAATC CCAGCACTCT GGGAGCCCAA
65701 GGTGGGTGGA TCATGAGGTC AGGAGTTCAA GACCAGTCTG GCCAAGATGG
65751 TGAAACCCCG TCTCCACTAA AAATACAAAA ATTAGCCTGA CGTGGTGGCG
65801 GATGCCTGTA ATCCCAGCTA CTTGGGAGGC TGAGGCAGAG AACTGCTTGA
65851 ACCCGGGAGG CAGAGGTTGC AGTGAGCCAA GATTGCGCGA CTGGACTCCA
65901 GCCTGGGTGA CAGAGTGAGA CTCTGTCTCA ACAACAACAA AAAAGAAATT
65951 AAAAAACTCT CAAAAATACA AAGTATAATA CAATACACAA CCCCATTCCT
66001 ACCACAGGGT TTGTTAATGT ACTGTCATGT TTGTAGTATA ATTTATTTTA
66051 AGGAAATAAA ATCATCACAG ATAAAGGTAG TTTACCGTGA TACCGCCTTC
66101 CAAGTTTCAT TCCTCCCCAG GCCATCTGTC CCATGAATTT GGTGTGTACC
66151 TTCCTGTCAT ATTCAGGTAT TTAATCTGCT GGAATTTACT TTTTGATTAG
66201 GTGTGAGGTG GGACTCTTGT TTTTCCCTAG ATGAGCCAAA TGTTCCAATA
66251 TTATTTATAA AATAGTTCAC CTGGTCAGGC ACAGTGGCTC ATGCCTGTAA
66301 TCCCAGCACT TTGGGAAGCT GAGGTGGGTG GATCACCTCA GGTCAGGACA
66351 GGAGTTTAAG ACCAGCCTGG CCAACACGGT GAAACCCCCA TCTCTACAAA
66401 AATACAAAAA TTAGCTGGGC ATGATGGCAG GTGCCCAGCT ACTCAGGAGG
66451 CTGAGGCAGG AGAATTGCTT GAACCCGGGA GGCGGAGGTT GTAGCAAGCT

66501 GAGATTGCGC CATTGCACTC CAGCCTGGGT GACACAGCGA GACTCTGTCT
66551 CGGAAAAAAA AAAAAAAAAA AGTTCACCTT TCTCCAGTAT TAGAAATGCC
66601 CCCTTTAGGC CTGGCGCGGT GGCTCACGCC TCTAATCCCA GCATTTTGGG
66651 AGGCCGAGGC GGGCGGATCA TGAGGTCAGG AGTTCGAGAC CAGCCTGGCC
66701 AACACAGTGA AACCCTGTCT CTACTAAAAA TACAAAAAAT TAGCTGGGCG
66751 TGGTGGCGGG CGCCTGTAAT CCCAGCTACT TGGGAGGCTG AGGCAGGAGA
66801 ATGGCTTGAA CCTGGGAAGC GGAGCTTGCA GTGAGCCGAG ATCACACTAC
66851 TGCACTCCAG CTTGGGCGAC AGAATGAGAC TCCATCCCCC ACTTCGCCAA
66901 AAAAAAAAAA AAAAAAAAAA AAAAAGAAAT GCCGCCTTTA TAGCGATTTA
66951 CCAGATCAAC CGTTCTCAAT GCTCTTTAAT ACGCTGGAGT TTCATACTAA
67001 GAAAAATAAA CATAAAAACA TTTTGGCCAG GCGCTGTGGC TCACGCCTGT
67051 AATCCCAACA CTTTGGGAGG CTGAGGTGAG CAGATCACAA GGTCAAGAGA
67101 TCGAGACCAT CCTGGCTAAC ATGGTGAAAC CCCGTGTCTA CTAAAAATAC
67151 AAAAAATTAG CCAGGCGTGG TGGCATACGC CTACAGTCCC TACGCCTATA
67201 GTCCCAGCTA CTTGGGAGGC TGAGGCAGGA GAATCTCTTG AACCTGGGAG
67251 GTGGAGGTTG CAGTGAGCCG AGATTGTGCC ACTGCACTCC AGCCTGGGCG
67301 ACAGAATGAG ACTCCTTTTC AAACAAAACA AAACAAAACA AAACAGAAAA
67351 CAAAAACAAA ACCAAAAGAC ATTCTGTGGG ATGGGCACGG TGGCTCATGC
67401 CTATAATCCC AACATTTTGG GAGGCTGAGG TGGGTGGATC ACTTGAGGTC
67451 AGGAGTTTGA GACCAGCCTG GCCAACATGT TGAAACCCCA TCTCTACTAA
67501 AAACACAAAA ATTAGGTCGG GCATGGTGGC TCATGCCTGT AATCCCAGCA
67551 CTTTGGGAGG CCGAGGCAGG TGGATCATCT AAGGTCAGGA GTTCGAGAGC
67601 AGTCTGGCCA ACATGGTGAA ACCCCATCTC TATTAAAAAT ACAAAAGTTA
67651 GTCGGGCATG GTGGCAGGCT CCTGTAGTCC CGGCTACTCA GGAGGCTGAG
67701 GAAGGAGAAT CACTTGAACC CAGGAGGCGG AGGTGGCAGT GAGTCAAGAT
67751 ACTGCCACTG CACTCCAGCC TGGGGAACAG AGGGAGACTC CGTCTCAAAA
67801 ATAAATAAAC AAATAAAAAT TTAAAAATTA ATAAATAAAA ATAAAAAAAT
67851 TAGCTGGGCA TGGTGGTGTG TGCCTGTAAT CTCAGCTACT TGGGAGGCTG
67901 AAGCAGGAGA ATCGCTTGAA CCCAGGAAGC AGAAGTTATA GTGAGCCGAG
67951 ATCATGCCAC TGCACTCCAG CCTGGGCGAC AAAGCAAGAC TCCGTCTCAA
68001 AAGGAAAAGA AAAAGGAAGC TGGAAGCTGA ATGAGATGGG CCTTTCAACC
68051 AAGGAGTTAG AAGGCCATCT GGTGGCAGGG CTGGCAGAGG ACCAGGAGTA
68101 AATAAGGCCA GAGAGGACAC CAGGGTCTGG GAGTGAAGGC ACTGAGCTTG
68151 GGTCCCCCTT TGGAAGACAA TGACCTGAGA GCTGTGAGAT TTCAGACAAG
68201 TTCCCGAACC TTTTGGGCCC TGCTTTCCTC ATCTGTAAAT GGGATAATAT
68251 CAGTCTCACC AGCTTCTTAA AATTCAATAC AATGGAGTTG GGTGTGGTGG
68301 CTCACGCCTC TAGTCCCGGC ACTTTGGGAA GCCGAGGTGG GCAGACTGTT
68351 TGAACTCAGG AGATGCAGAA CAGCCTGGAT AACATAGCAA AACAGTCTCT
68401 ACCAAAAATA CAAACAATTA GCTGGGCATG GTGGTGTGTG CTTGTAGTCC
68451 CAGCTACTAG GGAGGCTGAG GTGGGAGGAC TGCTTGAGCC CACGAGGTAG
68501 AGGCTGCAGT GAGCCATGAT TGCACCACTG CACTCCAGGC TGGGAGACAG
68551 AATGAGACCC TGTCTCAAAA CAAACAAGCA AACAAACAAT AAAGGAAATC
68601 CCTACCACAC TATCAGGGGC ATTTTGGCTG AGCGCGGTGG CTCACGCCTG
68651 TAATCCCAGC ACTTTGGGAG GCTAGGCTGG CAGGTCACCT GAGGTCGGGA
68701 GTTTGAGACC AGCCTGACCA ACACGGAGAA ACCGTCTCTA CCAAAAATAC
68751 AAAATTAGCC GGGCGTGATG GTGCATGCCT GTAATCCCAG CTACTTGGGA
68801 GGCTGAGGCA GGAGAATCTC TTGAACCCAG GAGGCAGAGG TTGAGGTGGG
68851 CTGAAATCGC GCCATTGCAC TCTAGCCTGG GCAACAACAG GGAAACTCCA
68901 TCTCAAAAAA ACAAAACAAA ACAAAAAACA AAACTCCCAT TTTTGCGAGG
68951 CAAATTGGGC TCACAGAGGT AAGCTGCATG TCCCTGTTGA TGGCAGAGCT
69001 GGGGTCTGGA TGCAGGTCTG CTTCGGGGTA ATCCGCTCTT TTGCCTTCCA
69051 GGGTCCTGCC TCTTACAATA TGAGCTGTCA AGTTAGATGC CTGCACTCAG
69101 TAAACCTACT CTGTTTTAAG TAAAAACAAC AAGAAACAAA TCTGAATATG
69151 CTAGCCTATC TCAGGTACGT TAAAGGAAAT TTTTAAATAG GGGGTTTTTT
69201 GACCATTTGG GGGAGTTTTG GGGGAGGGGC CTTCTGTCTA TACTTGAGCT
69251 GGGGGATGTT AGGGTTGTTC ATCTGGATCT AGAGGTTTTC CTGTAATGTT
69301 CTTACTCCAG AAGGAAATCT CTAGATGGGG AAAGAAGGTT TCAGCTTTTA
69351 TTCTAGTAAG CAGGGCTCTA CCTATAAAGA GCTGCTTCCA CCACTCTTTT
69401 TTTTTTTTTT GAGACGGAGT CTTGCTGTGT TGCCCAGGTT GGAATGCAGT
69451 AGTGCAATCT CGGCTCACCA CAACCTCTGT CTTCCGGGTT CAAGTGATTC
69501 TCCTGCCTCA GCCTCCCAAG TAGCTGAGAC TACAGGTGTG TGCCACCATG
69551 CCTGGCTAAT TTTTGTATTT TTAGTACAGA TGGGGGTTTC ACTATGTTGG
69601 TCAGGCTGGT CTCGAACTCC TGACCTCGTG ATCTGACTGC CTTGGCCTCC
69651 CAAAGTGCTG GGATTACAGG CATGAGCCAC CGCACCTGGC CTCCACCACT
69701 ATTATAATAT CACCAGGTTC CCCATTTGAA TCCTTCAGTG CCAAAGGTTT
69751 TGCAGAATTC AAATGTTTTT GGGACTTGAT AGGGCTGACC TAAAAGTACA
69801 CTCACTCTAT ATTAGGTAGC CCAGGAGGGC CTAGGCAGCC CAAGAACCAA
69851 ACACATGAGT GTTTCTGCAG GGAAATGTAT GAATATTGAC ATCAGTAGGA
69901 TGAAAATAAA TAATAGTCTT ACTTTAGTTC AGATTAGGTT TCTGTCACCA
69951 AATGAATTTT GGTGGCAGCC TGATGAAAAA TGTTGGTTCT CAGAGTGTTT
70001 TTGAGTTTAG AATTGTGGTT AAGGGAGTAT GGACCTGTTG ATAACAAAAA
70051 CAGGAACAAG GCCAGGTGTG GTGGCTCACA CCTGTAATCC CAGCACTTTG
70101 AGAGGCTGAG GTTGGTGGAT CACCTGAGGT CAGGCATTCG AGATCAGCTT
70151 GTCCAACATG GCGAAAACCC ATCTCTTCTA AAAATATAAA AATTAGCTGG
70201 GCGTGGTGGC ATGCGCCTGT AATCCTAGCT ACTTGGGAGG CTGACGCATG
70251 AGAATCACTT GAACCTGGGA GGTAGAGGTT GCAGTGAGCC AGGATCGCAC
70301 CATTGCATTC CAGCCTGGGC AAGAAGAGTG AAACTTCATA AAAAACAAAA
70351 ACAAAAACAA AAAACAGAGA AACAGGAACA ACAATCGCCA GCATATACCA
70401 AGTGCTTATC GTGTGTGCCA GGTACTCTAA TTATGTACTA TGTCAGTTGA
70451 TTCTCAAAAC ATATATGGCA CAACATGGGT ACTCTGAACA TGGGCACAAT
70501 CAATGTACAA TGCTATAATG TATAACACAG GACAATGTAG CTGTTAAAAG
70551 CATGGACACT CTATCTAGTC CATCTGGGTT ATAATCTCTG CTCTACCAGT
70601 GAATAACTGT AACTCTGGCA AATGACTTCT CTATGCCCTG TTTCCTCAGC
70651 TGGGAAATGG GGGATAATAT CAGTACTCAC CTCCTAAGAT TGTTGTGAGG
70701 ATTAAATGTG TTACTTTATA AGAAATGTCT GGCACATAGC AAAGGTTGTT
70751 ATTATTTTAA TTTTTACACA TGGGCAAACT AAGCCTCAAG TAACCTGTCC
70801 AAGAATACGT AGCTATGAAG TGTGGAGCTG GGATTTGGAG CTGGGGTTTG
70851 AATCCAGGCA ATCTAACTCC AGAGCCTACC TTCTATGCTA CTTTTTGGCT
70901 ACGAGCAAAC AATCTGTCAA GAAACAAAGT AGCTACTAAT CTAAACAGAT
70951 GTGAAATTTG AAGACCAGTT GATCTTTGGG GAATGTTGGG TTCTTCAGAC
71001 AATGGTAGCT CAGTAATGTA AAGGGACAAA TGACAGCCAC ATGCCAACTT
71051 GGTTAACTCC TTCCCTAGGT CCTGATGACC AAATAACCAG CTTATTTCTC
71101 AACTATTGGT TGGCTTTCAT TTCAGGTCTG GTCAGCTGCT TATGACCTTG
71151 TTCCCCACTG AGCAGACTCA CCATCTGGGC CCTGGCGGGC AGCAGCATGC
71201 AGTGCCGTGT CCCCATGGCG GTCCTGGTGG GCAGGGTCAG CCCCGAGCCG
71251 AAGCAGCAGG CACAGGGCAG GGGCATCGTG GCGGGCACAG GCCCGGTGCA
71301 GTGGTGGGGG CTGCCCAGCA TCTACATCGA GGCCTGGGTG TCGCTGGAGG

```
71351 AGGGCCTGGG CCCGGACCAG CCGTCCTGCA GACAAGTAAC GACGAAAGCG
71401 ACGTTCTCGG CGTTGGCGGC GGGAAGTGGA GGCCATGGAA CTCTTGGGCT
71451 GGGGAAGGAA AAAAGGCAGC AGTCAGGACT TCAGCCTTGG CTGGTCCTTC
71501 TCCCTCCATC TCTGACATCC CCTGTTGTTT CTCCCTTTGG TTCCGTCTTT
71551 TTTTAATATC TTCAGCAAGA GATGAGGCCT AACCTAACCC TGATCCTTTA
71601 TCAGATGATA GATTTGAAAA AAAATTTTTT TTGAGGGGGG TGATAGGATC
71651 AGAGGTTTAA TTTTTTTAAT GTAAAATTCG AGAAAAGGGT AAATAATTGG
71701 TTTAAGGCTC AGGAGCCCAG GTAAATTTTT AATTTTTAAC AAAGAACTTT
71751 AAAAAAACCC AACAGGGCCG GATGGGGAAA ATTTTTATCA GCAGAAATCT
71801 GAGTTTTAAA AAGTCACAGA TAATCTCCAA TAATGATCTA GAAATTGAAT
71851 ATCATGTACC CGGCAGACAG ATGTGGAGGC TTCTTCCTCT GGAACCTGGG
71901 GGGAGGGGTT ACTCATCAGA CCTGCCCCCG CCCCCCCAAG TACCCCCAGA
71951 GCCGTAGGCC CAAGGCCTGT GTTTAAGAAG CTCGGAGACG GGAGGCGGGA
72001 AGGGCGGAGA CACTCCAGGC TGGAGGAAAT GGCGCAAGCA GAGACGCAGG
72051 TGGAGGACGG AAGTGAACTG TGAGGGGCGT TACCGGATGT CGTTCCGCCC
72101 CGACCGGGTA GTTCTTGGCC AGATCTCCCA GGGGAAACTA GGGAACTTAA
72151 ATTAAAGGGG CCGTCTGAAA CCAGAAGACT GGACTGGAGG CGAGGAAAAG
72201 GAGGCGAGGG GAGGGGAGGG AGAAAAGAGA GTTATTTGGA GGTTTTTTCC
72251 CGCCTCCTCT AACTTGGCAG AGAGAGGAGA TGGTTCAGTG ATGGACGAAA
72301 AGATGAGAAG ACAGAGAAAA TAGAGGAGAT AAAGACAGGA TAAAAATCAC
72351 ATTAAACATG GAAAACAAAA ACAAAAACCA CAGTGGGACA ACAACAGGGA
72401 CAGATCAAAA AAAAAGAAAA AATACAGACA AAAGACGGAA GAAGACTATC
72451 GTAGGATGGG GCAAGTGAGA TGCAAAAATT TGGACTTGAG AAATATGTAG
72501 AAAAAGATGG AGATGTTAAC AACGGGAGGC AGGGGAGGGG GCGGGATGGT
72551 GGAGAGAGAG AGAAAGGTAG AGAGTTAGTT TAGAATTAAG CCCAGAATGC
72601 TCTTTTCCCA ACACAGGTTG CATGATGACA TCCTTACCTT TTCACCGCAT
72651 TCACAACCCT TTATGCTTCC TCTGTTACCA CCAATCAAGT TCTCCTTCTC
72701 TCACCTCAGT ACTCCCCCGT CTCCGCCCCT GCCTCATCCC TAGACCTTTC
72751 CGACTGGGAT GGCTAACCTG TTGTAAGCCC GCAGCTTTGG GCCTGGTCTC
72801 TGCTGCTCCC AGGCGGCCCC TTTGGTACTG CCTGAGCAAG AAGTGCTGGA
72851 GAGGAGGACC AGTCATCAAT AGGAGGATGA GATTGGGAGA GACACTCGGT
72901 GCAGGAGGCT GAGTGAGCAG GGGAGCACTA AGACCCAGGG GTAGTGGAGG
72951 ACTGCAGCAA CGAGCTGGAG GAGGAGAAGT AAGCGGTGGG GGGTGGGAGC
73001 CATCTGGTAC TTTGACAGCA TTCAAAACAG CATCGGCCAT AACAACAGAA
73051 ATGGCCAGTC AGTCCCAAGG TATCCAGCAG CTTCTGCAAG CTGAGAAGCG
73101 GGCAGCTGAG AAGGTGGCAG ATGCCAGAAA GAGTGAGTCT CCTCTTTCCT
73151 CCCTTAGGAG TTTGGAAAGA AAATTGGGGG TGGGGGACAG CAAACATTTT
73201 GGGAAAACCC AAGGCTGGCG GGAAGACAGA CAGCTAGGGT CTGGAGGCTG
73251 GTTAGGAGGG AAGAAATGGA TGGATATTAG AATCTGGCAC CTGGTTGGCT
73301 GAGAGAAGGC TGTATAACTT TCTGGAAGGG ACTGACTCCT GCTATTACAT
73351 TGTGTGTGTG TGGGTCCATC CCCACTCACT GTCCTTTCTT CTGCCTCCAG
73401 GGAAGGCCCG GCGACTGAAG CAGGCAAAGG AGGAGGCACA GATGGAGGTG
73451 GAGCAATACC GCAGAGAGCG AGAGCACGAA TTCCAGAGCA AGCAGCAGGC
73501 GGTGAGTTGA GGCAGAGTCG GGTTGAGACC CCACTGCAAG TTGGTGGGTG
73551 CATCTAGTGA GGTGTGTAAG GGTGACTCAA CAAGAAAATA TGGTGGCAGA
73601 GGGCTGAGGC TGAGGGGACC CTGGCAGGGA CCACAACATT GGTGAAACTT
73651 TGTGATGATA TGTAGGAGAG TCTGGGAGTT TTGAAGGCCA CATAGAGCTT
73701 GTGGGCGGAA TGCCACAGTC TGTGTAAAGT ATAACATCTA TGTGGAGTAT
73751 GATTAACATT TGTGGTGGAG GGTAGAGTTT TATGGTCATG GATGGTGAGG
73801 TGGTGGGGAT ATTACGGTCT GTTTTAGGAT GAAGTTGCAT GTTAGGTCTA
73851 AGGGGAAAGG GGACTGTGTT GATCTCTTTG GTGTTGGGAT ATTTCTGTGG
73901 GATGGGGGTG GTTTCTGAGA GGGCCTTTCT TCTAGGCTTT GTTTCAGGAT
73951 CTTTCCCCTC ATATGCCTGG ACCCTTGTCT GTTTCTGCTT TTCCCTTTCT
74001 CTCTTCCACC CCTCTCCCTA CCCCCCAGGC CATGGGCTCC CAGGGGAACC
74051 TGTCTGCTGA GGTGGAGCAG GCTACAAGGC GCCAGGTGCA GGGCATGCAG
74101 AGCTCCCAGC AGAGAAACCG AGAGCGTGTC CTGGCCCAGC TTCTTGGCAT
74151 GGTCTGCGAC GTCAGGCCCC AGGTCCACCC CAACTACCGG ATTTCTGCCT
74201 AGGGCCACCG TAGGGCCTGA CTCCTTCTGC CAGTTCCCTC CCTCAAAGAA
74251 ATCCTCCAAT CAAAATCACC TCCCACCATA ATCCCTGTCT TCTTTCCATC
74301 CCCTAGAAAT CCTGGGAGGC AGGATCCAAT AATTTTCCTG TGACACTTAT
74351 AAATATCCTG CTCACATCTG AATCTCCTTG TTGTTCTTTA ACCCTCACTG
74401 GGACTTTGTA AACTTCCAAG TCATTCTCAC CTAAACCCTC TGTGAAATTT
74451 GTAATATGGG GAAGTAGGAA TGTGGAAAAC ATCCTGACTT CAGTGTCTGG
74501 CCGATGTGGG TCCCTCTCTT GACCCTGTCA CTTGCTGGCT GTGAAACCAG
74551 GACAAGCTAC TTAACTTGGT AGCCTCGATG TCCTCCTCTG TGAAACTGGG
74601 ATGATAATAA TGCCTACCTT GTGAGGGTTG CTTCAATGAT TAGGAATCAT
74651 TCTGTAAAGT CTAGCACAGT TCCTTGCATG TTGTAGCAGT GATTCAGTAA
74701 GTAGCAACCC TGTGATACTA TTACCACCAC CTGCTCACTG GTCAAAACCT
74751 ACACAGCTGT TTCCTCACGT CCATCACTGG CTCTCTAATT CCACTTGTTC
74801 ATTCTGTGAC CCTAGTTATT TTCTGAAAAA TTGGTTCTTC TCTTTTCCCA
74851 GAGACCTTCT GATCTCCAAA AAGAGGAGAT GACTACATTT AGCCCCTCTC
74901 TTATAATTCC AGGTAGATAA CTGCATTTTG TAGCCTCTCT TTGTTTTTCT
74951 TTTGCTGATC TTTGTCTTTA TTAGATTTTC CTCCTTTCCT ATTTCCCCAA
75001 AGACTTATCA GATGCTCATT GCTTTCTAAG ATCTAAAATG ATACTGTGTT
75051 CCCTCATATG CATGCCCTTC CTTTCTATAT CCTTGACACC TTACTTTCCC
75101 ATTGTAACAA TAAAAAAAGT ATCAATAAAA TAATTATTGG CAAATAAATT
75151 GGTGAGTTGA AGCAGCCTCC TTTTGCCTCA TCATTTCTCA TTTTCAGTCA
75201 CTTTGTTTTT TTTTTTTTTG AGATGGAGTT TTGCTCTTGT TGCCCAGGCT
75251 GGAATACAAT GGCGTGATCT CAGCTCATTG CAACCTCTGC CTCCCAGGTT
75301 CAAGCGATTC TCCTGCCTCA GCCTCCCAAG TTGCTGGAAT TATGGGTGTG
75351 TGCCACCACG CCTGGCTATT TTTTGTATTT TTAGTAGAGA TGGGGTTTCG
75401 CCATGTTGGT CAGGCTGGTC TCAAACTCCT GACCTCAAGT GATCCACCTG
75451 CCTTGGCCTC CCAAAGTGCT GAGATTAGAG GTGTGAGCCA CTGTGCCTGG
75501 CCTTCAGTCA CTTTCTTGTT TTTTGTTTAC ATATTCCCTA AACAGCCCAA
75551 ATGGCTATCC TTTGAAACTT CTTGGAGAAA CAAGAACAAG TAGTACTTTA
75601 TTATTTCTCT AAAGTGAGAA ACATGGTTCC TCATTTGGGA ATCTGAGGAC
75651 TATAGATCGC AACTGTAGAG AAAAGCTGGA GTGTAGGAGC AAGTGCTCTT
75701 TGCCCCTTTA CCTTGCATTT TCTTCATAGC ACTTACTGCT ACTGGTTTTT
75751 TGAGACAAGG TCCTGCTGTG TTGCCCAGGC TGGAGTTCCA GCTCACGGCA
75801 GCCTTGACCC CCTGGACTCA AATGATCCTC CCACTTCAGC CTCCTGAGTA
75851 GCTGGGATTA CGGGCGAGTG CCACTATGCC TTGCTAATTT TAAAATTTTT
75901 TGTAGAGATG GGGTCTCACT TGCCCAGGCT GGTCTGAAAC TCCTGGGCTC
75951 AAGCAATCCT TCGGGCTCGG CTTCCTCAAG GGTTGGGTTA CAGGCCTGAG
76001 CCACTGCACC CTGACCACTT ATCGATACTT GACATTATAT TTGTGTTTAT
76051 GTGTTTTCTT TCCTGTAATG TAAACACTGT GAGAACAGGG CTGTTCACCG
76101 TTGTGTCCCC AGATCCTAGG ACAACATGTG GCACAAGGGA GGCAGTTGAT
76151 AAATACTTTT GAATAAATTA AATGATACTT GGGAAAATAC CTTCTATGAC
76201 ACCATTCTTG AATTAGTTAC TTCATTTGTC ACTGAAGACA AGCTTACTTC
76251 ACCAAGAATT TGAACCAATA AGGTAACCTG CAGTGTATTT ACTAACCAGA
76301 TTCTTTGAGC AGGGAGGCAG AATACAATAG AGAATGAGAG ATGTTTGCAT
76351 CCTGGCTGTA ACCTCACCAG CCGTACTGCT TGAGATATGT TGCTTTGCTT
76401 CGCTTCTGTC AATAAGATGA GAATAACGGT ACCTACTCCT TAGTATTAAA
76451 TGATTAAGTA TGTTAACAGG GAGAGGGCCA AACGTTTGTT GTTTTATTAC
76501 ACAGCAGGAC ATCAGGTCTT ACTTTTGTAG CTCCCCATCT CAAAGACGGG
76551 GATAGCAAAT GTTTCATTCA GGAAAAAAAT CCAGGTTGAA CAATGGGGCT
76601 GTTGGGGCGG GGCCAAGAAC ATTCTGCTCG AATTAACAGT ATTAATGGGC
76651 CGGGCGCGGT GGCTCACGCC TGTAATCCCA GCACTCTGGG AGGCCGAAGT
76701 GGGTGGATCA CCTGAGGTCA TACATGGGTG AAGCCCCGTC TCTACTAAAA
76751 AAACAAAAAT TTGCTGGGCG TGGTGGCGGG CGCCTGTAAT CCTAGCTACT
76801 CGGGAGGCTG AGGCAGGAGA ATCGCTTGAA CCCGGGAAGC AGCGGTTGCA
76851 GTGAGCCGAG ATCAGGACAT TGCACTCCCG CCTGGGCGAC AGGGCGAGAC
76901 TCTGTCTCAA AACAAAAACA AAAACAGTAT TAATGGAATG TAGTATAACC
76951 CTCAAGCCCT ACTATTAACA CTTGGGGCCG AATCCAGACC CCGTCTTCCC
77001 GCTCGGATTC AGAACACCTT CCTGACTCAC TGGCCCTAGG GCATCAGCTA
77051 CCTCGGACAG CATCCTTTTG GGAAAATACC GCCCACCAGC CCCACGACTG
77101 GGAAAGAGTC GGGAAACACC CCCGAGCAAT CCAGTTCCCT GAGACTTCCC
77151 TCCTCCCTCC CCTCAGCTAG GGCCTGCCGG TTCCTAGTGC GTGCCCAGCA
77201 GTCCTCAGGT CACCTTCACT ACCGGGCCAA GGACCCCGTG GGAACTCGCA
77251 GCCTTCGCCA CACTCGTTCC TCGCGCATCC ATGGAGGGGT GCCTACAGAG
77301 AAGACCTGCG TGGCAAAAAC CTAAACGAAG AGATGAGGGG CATGGAGAGG
77351 AGTAGGATAA GAGAATAAAG ATAACAGTGG GGGGGAGACG TTAGTTTCCT
77401 TTATATCTTT TGTTACTGGC GGTAGCAGTG AAGTTAGAAA CGGTTTTAAA
77451 ACAAATTTCA GACAGGCATT TTCCAAAGGC AAGCCTGGAG CGCACGGATC
77501 TGTATAACCG CGGAAGGCCC TGTTTCCGGT CCCTTGCGCC TGCGCTCTTG
77551 CAGCCAAGAA GGCGGGAGGC TGGAGTAGAG GGAAGCCTGC AACCGGAAGT
77601 GAAGGCAGAT TTCCCTCCTT CGTCGCTGTT GCTGCCGCCA TACGCGCTCT
77651 CCCTGTTTAG GTAAGCTTTG GCCTTCGCTA CAATCCGTTT CCATCTGCGC
77701 TTCTCCGCAC CCATCCCGTC ACATGGGTTC CTGATACCCT TTTCACAGGC
77751 GATGGTCTGG TCGCTGGGGC CTAGTTGGTT CGCTATTTCC TTAGCTTGCA
77801 TCCCTTTCGA GAGCAAAGAG CTCCTGGGGG AAGGAAGGGA AGCTAAGGGG
77851 GGACCCAATC CAAGATGGTG TCCTCGGCGC CATTGTGTTC GTTTTGCTCC
77901 CTTCTTCCAA TGGGTTCTTC TCATATTGGA GGCCTCAGCA TCAATGAGAG
77951 GCGGCGCTCG GCGTCCCTTG GTCTTGGTAT TTGCGGAGGG CGGGGCTCTT
78001 CTCACCTTCC TTGTCCTTTC TTGAGCTCTT TCTCGGCCCT CGGTGGGACT
78051 GGGAGGAGGA GCTGGTTTCT GGGCCCAGTT GGATTTTTCT CACCTTGACT
78101 TGCCCAACTT AATTTGGAGT GCCTTCCAAG TGTTTACGAT ACGATTGGTG
78151 TCATTGTATG TTTCTCCAAA AGGAGTCTCA CCTTCGTAGC GTAACAGTGA
78201 TGTGAGACCA CTTGGTAAAG ATCCTGTTAA AGCCTGGGCG GGGATTGCCT
78251 TTCTCTGTCA CCTATTAGCT TTCTTATTGT AGGGTGGAGA CATGAATTTT
78301 GTTTTTTTGT GGCCGAGCCA TTTGTCTTGC ACCGCCCCTC CCCCCCATGC
78351 TAATTACACA AGGCTTGCTT AAACAGCGGA AGGGAGGATA CTGAGAAGTG
78401 GGAGGCTGAG AGCTATGGGA GGTGGACGGC GGCCATATGA TGTTTTCTTT
78451 TCGAAAGGTG AGCGCTTTGC GCAGTGATGA CCCTCATCTA TCACCCTTGA
78501 CTGATGGCTG CTGAGTTAGG CATCCATAAC GGTGGGATTA TAATAGGGAA
78551 AGCGGAGTCT TCCTTTGAGG ACTTTTCAGG ACTCTACTTG TCATCTCCAT
78601 TTTCCACTTT ACTAAGTTAT TAGTCGTATT TTACCTTTTA TTATCTATTC
78651 TATTTCCTCA CTGTTACTTT CAGATCAAGA ATTTATAAGT TGGTCTTCCC
78701 CTTCCAACTT TTCTGGTTTC CGCTACTGTG ATTGCTAATC TTGTTGGGAA
78751 CCTCTGTCCT AACCACTTTC CCTGGTACTG CTTTTTCTGT TCTGTTATAT
78801 TTGCTTTTCG TTTTTATGTT TTGTATCTGT TTTTCTTTCC AGGTAAAAGT
78851 TTCCTGGTTT AGGGAAAGTG GGAACTGGGG ATGGAAAAGA GGTGGTGAAG
78901 GCTGTGCTCG TGATTAAGTC TTGCTTTTTT TTTTCCCCCC TCCAGCTCTT
78951 CTGTTAGAAA TAGTATCTTT GTTTTCCTTT GCTGTTCCTC AATCCCCTAC
79001 TCTTCACCCC TTGTTTTCAC CTATTTTGCG AGAACCCATC CAGATCCCCC
79051 TTCCCTTCTT CCCCTGCCGG CCCAGTTATG GCAGAGAACG ATGTGGACAA
79101 TGAGCTCTTG GACTATGAAG ATGATGAGGT GGAGACAGCA GCTGGGGGAG
79151 ATGGGGCTGA GGCCCCTGCC AAGAAGGATG TCAAGGGCTC CTATGTCTCC
79201 ATCCACAGCT CTGGCTTTCG TGACTTCCTG CTCAAGCCAG AGTTGCTCCG
79251 GGCCATTGTC GACTGTGGCT TTGAGCATCC GTCAGAAGGT AAATTTTCTC
79301 TTGGGCATGT AGTGCTCATT GGGCTCTTTA AGGGTACAAT ACAAAGATGT
79351 GTTTGTCGTT GCTCAGGTGG TGGTAAGGGT TTATACTTAA GGCTAGATCA
79401 GGGCCAGGTG CAGTGGCTCA CGCCTGTAAT CCCAGCACTT TGGGAGGCCG
79451 AGGCAGGAGG GTGGCCACTT GAGCTCAAAA GTGCAAGAGA AGCCTGGGCA
79501 ACACAGCGAG ACTCCTGTCT CTACAAAACG TTCAGAAATT AAGCAGGTGA
79551 AGGTTGAGGC TTCAGTGAGC CGTGATTGCA CCACTGTGCA CCAGCCGGGG
79601 CGACAGTGAG GAAGAAAAAA TCAGGGGATA AGTATCAAAA ACAATTTTGG
79651 ATAGAGGAGG CTTATACAGG CTTATTCTTT CTTTCGTGAT AGCACCAAAG
79701 TGCTAATGAT CCAAAAGTGA CTTCCAGGTC TGCCATTCAT TCTTGTGACT
79751 GGCTTTTCTT GTCTGCTTAT TTTTAATTTT GTCACTTGAC TTCTAATTTT
79801 AAATTTCCAG AAAGGTCCTG CTTGGACCTG TAGTCTCCCT CTGTTGGGCC
79851 AGGCCAACTG TGGTCTCTGG AAACCTCTAT GACTGGTTTA GAGATGACTG
79901 GCTTCTGGGT CAGGTACCAA GTCCTTCATT TTGTCCAGGG TTGTAGTAGT
79951 TACGTGACCC GAAGTATAGC AACCTAAGCA GGAGAAGTGG TCTGTGGCAG
80001 GTATTCAAAT GTCATGAATT GTTACAGATT AAGAAAAATA AGGACAGAGC
80051 TAGGATCATT GAAGGTGAGC GGTTGGTAGA TGCAAGGGGT TTGTTACTAG
80101 GACTGGGAAG GCCTAGATCT GGAGGAGGCT AAAGCTAGGA GGAATTAGGA
80151 GAGTCTGATT TTGAGGTGAA TGTAATTGAG CAGAGAGAGG TAAAATGGGT
80201 CTGGAAGTTG GCAAGAACCA GGTAAATACT AGACTTTGAG AATTGAGTGG
80251 TAAGAAATGG GCTTGGCATG GTGAAAAAGG TAGAGTTATC TGGAGACTGA
80301 AGTCTAATTT ATCTTCCTCC CCCCCCAACT TTTAGTCCAG CATGAGTGCA
80351 TCCCTCAGGC CATTCTGGGA ATGGATGTCC TGTGCCAGGC CAAGTCGGGC
80401 ATGGGAAAGA CAGCAGTGTT TGTCTTGGCC ACACTGCAAC AGCTGGAGCC
80451 AGTTACTGGG CAGGTATATT TGGGGAGAGT GCTGGGGAGG GGATTTTGGT
80501 TAGGACTATA AGGGAAGGGT GTTTTTGTCC TAGCTACATG ATGCTTGCAG
80551 AGCCATGAGC ACATGACCTC TGTTACCCTT GACAACCTGA CAGCTGTGGG
80601 GGATGTTCTG TCGCAAGCGT GGGGTTCATG ATTTAGATCA CATAATTGAA
80651 GTCATTTATT ATCGGCCCAG GTGTGTTTTT GTGACAGTCA CTTCCCTAGA
80701 GGGGATAATG AAGAGCTACA TTTACCATAT GTCTCCGTCT ACTTCCTGCC
80751 TAAGGTGTCT GTACTGGTGA TGTGTCACAC TCGGGAGTTG GCTTTTCAGA
80801 TCAGCAAGGA ATATGAGCGC TTCTCTAAAT ACATGCCCAA TGTCAAGGTA
80851 AGCCAAGGTA AAGAGACCTG AGAGTGAGGG TGTGGCAAGT TGGAGGGATA
80901 AGAAACTTGT AGGCCAATAG TCTCTTTAAT TTTGGAGAAG CTTTAGTTTG
80951 CTGTGGTGTA ACAGAGTGTT GAGTTCCTAT GTAACAGGAG GATTCGTAAT
81001 TGGGCTATGG ATGATGCTTA ACACAAGACC ACCCTTTTCT TACTACTTTA
```

```
81051 TACTGACTTT GAATCATATC AGTTTAATAA TTTTGGGGTA TGTGGCAGAG
81101 AAAGCCGGAA ACTTTAAAAC AGCTCCAGTG GTGTGTGAAT ATTGAGGATT
81151 CTGGCCAAGT GCACAATGGC TTACACCTGT AATCTCAGCA GTTTGGGAGG
81201 CCAAGGCGTG TGGATTATTT GAGGTCAGGA GTTTAAGACC AATGTGGCCA
81251 ACAGGATGAA ACCCTCTCTC TACTAAAAAT GCAAAAATTA GCCGTGCATG
81301 GTGGCACACA CCTATAGTCC CACCTGCTTG GGAGGCTGAG GCAGGGGAAT
81351 CGCTTGAACC CAGGAAGCAG AGGTTGTAGT GAGCTGGGAT TGTGCCATTA
81401 CACTCCAGCC TGGGTGGCAG AGTGAGACTC CCATCTCAAA AAAAAAGAAA
81451 AAATCTGATT GAAGTTAAGC ATTTTTGGCA AGAATCCTTC ATAGGTGATA
81501 CTGTATCTCC TGTTATGCCA CAAATCTGGT CGACTTATGT TAGTTATTTT
81551 ATTTTATTTT TATTTATTTG TTTTGAGATG GAGTCTCGCT GTGTCCCTCA
81601 GGCTGTGAGT GTAGTGGCGC GATCTCAGCT CACTGCAACC TGCGCCTCCC
81651 ACGTTCAAGC GATTCTCCCG CCTCAACCCC CCGAGTAGCT GGGACTACAG
81701 TGTGCCATCA TGCCTGGCTA ATTTTTGTTT TTTTTTAGTA GTGACAGGGT
81751 TTCGCCATGT TGGCCAGGCT GGTCTCGAAC TCCTGACCTC AAGTGATCCC
```

APPENDIX II

Repeat masker output for the TN62/TN82 genomic sequence

Repeat class/family	Repeat	Position in sequence Begin	End	Orientation	Repeat length
SINE/Alu	AluSg	10	235	+	225
SINE/Alu	AluSp	243	541	+	298
SINE/Alu	AluSc	617	924	+	307
SINE/Alu	AluSq/x	938	1063	+	125
LINE/L2	L2	1336	1479	+	143
SINE/Alu	AluSq	1480	1787	C	307
SINE/Alu	AluSp	1847	2165	C	318
SINE/MIR	MIR	2228	2302	C	74
SINE/Alu	AluSc	5465	5760	C	295
LINE/L2	L2	5879	6036	C	157
SINE/Alu	AluSg	6037	6342	+	305
LINE/L2	L2	6343	6381	C	38
SINE/Alu	AluSx	6420	6730	C	310
SINE/Alu	AluSg	6731	7006	C	275
SINE/Alu	FRAM	7040	7204	+	164
LINE/L2	L2	7210	7262	C	52
SINE/Alu	AluSg	7263	7563	C	300
SINE/Alu	AluSc	7879	8173	C	294
Simple_repeat	(GGA)n	8301	8328	+	27
SINE/Alu	AluY	8866	9176	C	310
LTR/MER21-group	MER21B	9186	9805	C	619
LTR/MER21-group	MER21B	9861	9922	C	61
Simple_repeat	(CA)n	9923	9965	+	42
SINE/Alu	AluJb	9984	10265	C	281
SINE/Alu	AluJo	10266	10568	C	302
SINE/Alu	AluY	10583	10881	C	298
SINE/Alu	AluY	10882	11175	C	293
Low_complexity	AT_rich	11180	11202	+	22
LINE/L2	L2	11215	11392	C	177
SINE/Alu	AluSp	11530	11807	+	277
LTR/MER21-group	LTR48	11930	12173	+	243
SINE/Alu	FRAM	12174	12333	C	159
SINE/Alu	AluSx	12339	12642	C	303
SINE/Alu	AluSg/x	12654	12821	C	167
SINE/Alu	AluSg1	12824	13055	C	231
SINE/Alu	AluSc	13061	13359	C	298
Low_complexity	AT_rich	13362	13385	+	23
LTR/Retroviral	LTR26E	13432	13575	C	143
SINE/Alu	AluSx	13636	13927	C	291
Simple_repeat	(TTTTG)n	14111	14135	+	24
SINE/Alu	AluY	14136	14414	C	278
SINE/Alu	AluSg	14515	14799	C	284
Low_complexity	GA-rich	14950	15099	+	149
SINE/Alu	AluSx	15103	15341	+	238
Simple_repeat	(A)n	15342	15363	+	21

APPENDIX II Continued...

LINE/L2	L2	15367	15415	+	48
LINE/L2	L2	15781	16164	C	383
SINE/Alu	FLAM_A	16198	16328	+	130
SINE/Alu	AluSq	16360	16661	+	301
LINE/L1	L1ME2	16923	17193	+	270
SINE/Alu	AluJo	17194	17321	+	127
SINE/Alu	AluSg	17322	17628	C	306
SINE/Alu	AluJo	17629	17807	+	178
LINE/L1	L1ME2	17808	18022	+	214
SINE/Alu	AluY	18033	18334	+	301
SINE/Alu	AluSq	18397	18700	+	303
SINE/Alu	AluSx	19097	19401	+	304
SINE/Alu	AluSg	19517	19815	+	298
SINE/Alu	AluSx	20293	20592	C	299
SINE/Alu	AluJo	21026	21335	+	309
SINE/Alu	AluSx	21541	21820	+	279
Simple_repeat	(CAA)n	21821	21847	+	26
SINE/Alu	AluSg	21870	22151	+	281
Simple_repeat	(TAAA)n	22152	22179	+	27
SINE/Alu	AluJo	22187	22408	+	221
DNA/MER2_type	Tigger1	22568	22771	+	203
SINE/Alu	AluJo	22772	22932	C	160
SINE/Alu	AluSp	22933	23224	C	291
SINE/Alu	AluJo	23225	23256	C	31
SINE/Alu	AluJb	23257	23550	+	293
SINE/Alu	AluJo	23551	23575	C	24
SINE/Alu	FAM	23583	23706	C	123
DNA/MER2_type	Tigger1	23707	23763	+	56
SINE/Alu	AluSx	23764	24045	+	281
Simple_repeat	(TAAA)n	24046	24080	+	34
DNA/MER2_type	Tigger1	24082	24128	+	46
SINE/Alu	AluJb	24133	24370	C	237
SINE/Alu	AluY	24371	24673	C	302
SINE/Alu	AluJb	24674	24760	C	86
SINE/Alu	AluJo	24813	25104	C	291
DNA/MER2_type	Tigger1	25105	25240	+	135
SINE/Alu	AluSp	25241	25551	C	310
DNA/MER2_type	Tigger1	25552	25651	+	99
SINE/Alu	AluJb	25653	25934	+	281
SINE/Alu	AluSx	25939	26219	+	280
Simple_repeat	(CA)n	26229	26285	+	56
SINE/Alu	AluSx	27183	27348	C	165
SINE/Alu	AluSx	27349	27644	C	295
SINE/Alu	AluSx	27645	27775	C	130
SINE/Alu	AluSx	27790	28078	C	288
DNA/MER1_type	MER5A	28161	28293	+	132
SINE/Alu	AluS	28468	28527	+	59
SINE/Alu	AluSx	28529	28774	+	245
SINE/Alu	AluSg	28992	29292	C	300
Low_complexity	AT_rich	30726	30762	+	36

APPENDIX II Continued...

Simple_repeat	(TC)n	31183	31237	+	54
SINE/Alu	FLAM_A	31238	31360	C	122
Low_complexity	CT-rich	31370	31493	+	123
SINE/MIR	MIR	32396	32425	+	29
LINE/L1	L1MC4	34247	34334	C	87
SINE/Alu	AluJo	34335	34623	C	288
LINE/L1	L1MC4	34624	34730	C	106
SINE/Alu	AluSx	34826	35119	C	293
SINE/Alu	AluSq	35128	35434	C	306
SINE/MIR	MIR	35469	35564	+	95
SINE/Alu	AluJb	35594	35886	+	292
SINE/Alu	AluSx	36050	36358	C	308
Low_complexity	GC_rich	38591	38619	+	28
SINE/Alu	AluJo	40019	40323	C	304
SINE/Alu	FLAM_C	40486	40622	C	136
LTR	LTR42	41005	41083	+	78
Simple_repeat	(TAAA)n	41580	41612	+	32
Simple_repeat	(TTCA)n	43422	43448	+	26
LINE/L2	L2	43450	43491	C	41
Low_complexity	CT-rich	43492	43628	+	136
Low_complexity	AT_rich	45525	45548	+	23
Simple_repeat	(GA)n	46883	46907	+	24
Simple_repeat	(CAGAGA)n	46979	47156	+	177
DNA	MER94	47831	47955	C	124
SINE/Alu	AluSx	48253	48550	+	297
LINE/L1	L1ME3	48570	49018	C	448
LTR/MER21-group	MER39	49394	49792	C	398
SINE/Alu	AluSx	49793	50101	C	308
SINE/Alu	AluY	50304	50599	+	295
Simple_repeat	(CAAAA)n	50600	50621	+	21
SINE/Alu	AluY	50622	50634	+	12
Simple_repeat	(CCCTAA)n	51083	51173	+	90
LINE/L2	L2	51365	51911	C	546
Simple_repeat	(TG)n	51912	51966	+	54
LINE/L2	L2	51967	52035	C	68
SINE/Alu	AluJo	52036	52303	C	267
SINE/Alu	AluSq	52303	52598	C	295
LINE/L2	L2	52304	52599	C	295
SINE/Alu	AluJo	52600	52645	C	45
SINE/Alu	AluSx	52651	52951	C	300
SINE/Alu	AluSx	52952	53251	C	299
SINE/Alu	FRAM	53255	53410	C	155
SINE/Alu	AluSg	53421	53713	C	292
LINE/L2	L2	53714	54112	C	398
SINE/Alu	AluJo	54682	54988	C	306
SINE/Alu	AluY	55030	55343	+	313
SINE/Alu	AluSq	55706	55992	+	286
SINE/Alu	AluSg	56439	56731	C	292
SINE/Alu	AluSx	56734	57029	C	295
LINE/L2	L2	57291	57738	+	447

APPENDIX II Continued...

LINE/L2	L2	57785	57906	+	121
SINE/Alu	AluSx	57907	58209	C	302
SINE/Alu	AluY	62204	62504	+	300
SINE/Alu	AluSg	63653	63997	+	344
SINE/Alu	AluSx	64073	64384	+	311
Low_complexity	AT_rich	64990	65010	+	20
LINE/L1	L1M4	65025	65171	C	146
SINE/Alu	AluJo	65184	65486	C	302
SINE/Alu	AluSg1	65651	65956	+	305
SINE/Alu	AluSx	66273	66571	+	298
SINE/Alu	AluY	66608	66928	+	320
SINE/Alu	AluY	67025	67320	+	295
Simple_repeat	(CAAAA)n	67321	67366	+	45
SINE/Alu	AluSx	67383	67514	+	131
SINE/Alu	AluSx	67515	67827	+	312
SINE/Alu	AluSx	67828	68014	+	186
SINE/MIR	MIR	68180	68274	+	94
SINE/Alu	AluJo	68286	68593	+	307
SINE/Alu	AluSp	68626	68933	+	307
SINE/Alu	AluSg	69397	69691	C	294
DNA/MER2_type	MER46C	69718	70037	C	319
SINE/Alu	AluSq	70060	70364	+	304
SINE/MIR	MIR	70380	70461	C	81
SINE/MIR	MIR	70527	70742	+	215
SINE/MIR	MIR	70743	70890	C	147
Low_complexity	A-rich	72362	72433	+	71
SINE/MIR	MIR	74507	74702	+	195
SINE/Alu	AluSp	75202	75502	C	300
LINE/L2	L2	75708	75745	+	37
SINE/Alu	AluJo	75746	76011	C	265
LINE/L2	L2	76012	76171	+	159
SINE/Alu	AluSx	76648	76924	+	276
SINE/Alu	AluJ	79402	79606	+	204
SINE/Alu	AluSx	81161	81453	+	292
SINE/Alu	AluSx	81547	81799	C	252

APPENDIX IIIA

Mutation analysis primer sequences.

Primer	Sequence [a]	Position [b]	Orientation [c]	Product
AIF-1388f	5'-CTGCATTCTTTGAGGTTAAGGC-3'	2387-2408	+	AIF-1 1st round-1
AIF-1m13f1	5'-TGTAAAACGACGGCCAGTCACACTATAGCGTTGTTGATGG-3'	2463-2484	+	AIF-1 2nd-1
AIF-1m13f2	5'-TGTAAAACGACGGCCAGTATGGGGATGTGAGGGTGG-3'	2876-2893	+	AIF-1 2nd-2
AIF-1m13r1	5'-GGAAACAGCTATGACCATGCTACACGTTGCCCATCATCC-3'	3015-2996	C	AIF-1 2nd-1
AIF-11201f2	5'-AGCTCTAGGTGAGTCTTGG-3'	3200-3219	+	AIF-1 1st round-2
AIF-1m13f3	5'-TGTAAAACGACGGCCAGTCACTGTGGTAGGTGGGCC-3'	3345-3363	+	AIF-1 2nd-3
AIF-1m13r2	5'-GGAAACAGCTATGACCATGGGCGATATTGGTGAGAAACG-3'	3469-3449	C	AIF-1 2nd-2
AIF-11605r	5'-CGTTGTCTCCTCCACCTAGC-3'	3604-3585	C	AIF-1 1st round-1
AIF-1m13f4	5'-TGTAAAACGACGGCCAGTAAGGGACCGCAGTGTGGC-3'	3710-3727	+	AIF-1 2nd-4
AIF-1m13r3	5'-GGAAACAGCTATGACCATGGTAGGCGGTGGAGACAGG-3'	3854-3836	C	AIF-1 2nd-3
AIF-1m13f5	5'-TGTAAAACGACGGCCAGTCTGGGCCTTCAGCAGTCC-3'	4052-4069	+	AIF-1 2nd-5
AIF-1m13r4	5'-GGAAACAGCTATGACCATGTGAGATGGACAGACCTGGG-3'	4208-4190	C	AIF-1 2nd-4
AIF-1m13r5	5'-GGAAACAGCTATGACCATGGTCTTCCTGCCCTTCCTGC-3'	4599-4581	C	AIF-1 2nd-5
AIF-12715r	5'-CTATCCACAGCTGCCTGCC-3'	4714-4696	C	AIF-1 1st round-2
1C7211f1	5'-TAGCTTCACCAATCAGCTTGC-3'	26312-26331	+	1C7 1st round-1
1C7m13f1	5'-TGTAAAACGACGGCCAGTGGTCTGGGTACTGGTGAGC-3'	26413-26431	+	1C7 2nd-1
1C7m13r1	5'-GGAAACAGCTATGACCATGAGAGCCGTGTGTTTCAGCC-3'	27048-27030	C	1C7 2nd-1
1C7933r1	5'-GGAGTCCACTGAGTGTCACC-3'	27093-27074	C	1C7 1st round-1
1C7359f2	5'-CCCAAAGTATCTATTCCCATGC-3'	29359-29380	+	1C7 1st round-2
1C7m13f2	5'-TGTAAAACGACGGCCAGTACCCTATGCCTGGAGAAAGC-3'	29381-29400	+	1C7 2nd-2
1C7m13f3	5'-TGTAAAACGACGGCCAGTGGTGGAGAAAGGTGAGATGC-3'	29820-29839	+	1C7 2nd-3
1C7m13r2	5'-GGAAACAGCTATGACCATGGCTGAGGATGTTCTGCATGG-3'	29990-29971	C	1C7 2nd-2
1C7m13f4	5'-TGTAAAACGACGGCCAGTTCACATGGGAACACACTGC-3'	30231-30249	+	1C7 2nd-4
1C7m13r3	5'-GGAAACAGCTATGACCATGCTGCCTTCTTGGACCTTTCC-3'	30464-30445	C	1C7 2nd-3
1C7m13r4	5'-GGAAACAGCTATGACCATGGTCCTCTCAGTCCATCTCG-3'	30841-30823	C	1C7 2nd-4
1C71926r2	5'-GACACCTTCCTCAACCCAGG-3'	30927-30908	C	1C7 1st round-2
ATP130f	5'-ACAGTGGGACAACAACAGGG-3'	72380-72399	+	ATP6G 1st round-1
ATPm13f1	5'-TGTAAAACGACGGCCAGTTTCACCGCATTCACAACC-3'	72641-72659	+	ATP6G 2nd-1
ATPm13f2	5'-TGTAAAACGACGGCCAGTCTGCAAGCTGAGAAGCGG-3'	73084-73101	+	ATP6G 2nd-2
ATPm13r1	5'-GGAAACAGCTATGACCATGCCTAACCAGCCTCCAGACC-3'	73256-73238	C	ATP6G 2nd-1
ATP1374f2	5'-GCAGGGACCACAACATTGG-3'	73624-73642	+	ATP6G 1st round-2
ATPm13r2	5'-GGAAACAGCTATGACCATGATGTGGCCTTCAAAACTCCC-3'	73693-73674	C	ATP6G 2nd-2
ATP1472r2	5'-CAGACTGTGGCATTCCGCC-3'	73722-73704	C	ATP6G 1st round-1
ATPm13f3	5'-TGTAAAACGACGGCCAGTAGGTGGTGGGGATATTACGG-3'	73798-73817	+	ATP6G 2nd-3
ATPm13r3	5'-GGAAACAGCTATGACCATGGGTTTAGGTGAGAATGACTTGG-3'	74437-74416	C	ATP6G 2nd-3
ATP2337r	5'-AGGAGGACATCGAGGCTACC-3'	74587-74568	C	ATP6G 1st round-2
BAT1-249f	5'-TGTGATTGCTAATCTTGTTGGG-3'	78727-78748	+	BAT1-1st round
BAT1-m13f1	5'-TGTAAAACGACGGCCAGTTTGGGAACCTCTGTCCTAACC-3'	78744-78764	+	BAT1-2nd
BAT1-m13r1	5'-GGAAACAGCTATGACCATGGGATGGAGACATAGGAGC-3'	79207-79187	C	BAT1-2nd
BAT1-768r	5'-TCTGACGGATGCTCAAAGCC-3'	79286-79267	C	BAT1-1st round

[a] Underlined sequences represent the M13 tails

[b] Positions are given as co-ordinates in the TN62/TN82 sequence (HSY14768)

[c] + refers to primers in the forward orientation and C refers to primers in the reverse orientation

APPENDIX IIIB

PCR-SSP primer sequences.

Primer No.[a]	Sequence (5' → 3')	Type	Position [b]	Orientation [c]	Loci
63	TGCCAAGTGGAGCACCCAA	control	517-535	+	DRB1 (M17378)
64	GCATCTTGCTCTGTGCAGAT	control	598-579	C	DRB1 (M17378)
2795	AGCTTCACCAATCAGCTTGC	Consensus	26312-26331	+	-
2796	CTTTTACCCAGAACAAGCCTC	Allele	26499-26479	C	1C7*-412
2797	CTTTTACCCAGAACAAGCCTG	Allele	26499-26479	C	1C7*-412
2872	TGGGATGTTCTACTCCAAGC	Consensus	29402-29421	+	-
2873	ATCTCGGAACCACGTGACG	Allele	29617-29599	C	1C7*2708
2874	ATCTCGGAACCACGTGACA	Allele	29617-29599	C	1C7*2708
2875	TGCACACGTAGATGCTGGC	Allele	29768-29750	C	1C7*2859
2876	TGCACACGTAGATGCTGGA	Allele	29768-29750	C	1C7*2859
2877	CAATGTCCTTGGGAGGCAG	Allele	29881-29899	+	1C7*3008
2878	CAATGTCCTTGGGAGGCAA	Allele	29881-29899	+	1C7*3008
2882	CCGGAGAGAGTAGATTTGGC	Consensus	30403-30384	C	-
2881	CAGCACCGTCTATTACCAGG	Consensus	30062-30081	+	-
2883	ACCACAGCCGGC/TAGCTGC	Allele	30479-30462	C	1C7*3571
2884	ACCACAGCCGGC/TAGCTGA	Allele	30479-30462	C	1C7*3571
2885	CTGGGACCACAGCCGGC	Allele	30484-30468	C	1C7*3577
2886	CTGGGACCACAGCCGGT	Allele	30484-30468	C	1C7*3577
2887	CTCTGGGGTCAAATAGTAATTTA	Allele	30703-30681	C	1C7*3790
2888	CTCTGGGGTCAAATAGTAATTTG	Allele	30703-30681	C	1C7*3790
2889	GGGTGAGCCATGTGTTTTTTTT	Allele	30713-30734	+	1C7*3843
2890	GGGTGAGCCATGTGTTTTTTTA	Allele	30713-30734	+	1C7*3843
2893	ACATAAGTTTTGATGATGATCAA	Allele	30787-30809	+	1C7*3918
2894	ACATAAGTTTTGATGATGATCAT	Allele	30787-30809	+	1C7*3918
2895	GAGCAGGAACTCCACTATGC	Consensus	31080-31061	C	-
2934	TCTTCATAGTCCAAGAGCTC	Consensus	79121-79102	C	-
2935	AGAGGTG/CGTGAAGGCTGTG	Allele	78888-78906	+	BAT1*-180/-65
2936	AAAAGAGGTG/CGTGAATTAAGTC	Allele	78885-78920	+	BAT1*-180/-65
2937	GCTCTAGGTGAGTCTTGGG	Consensus	3201-3219	+	-
2938	ATGGCGATATTGGTGAGAAAC	Allele	3470-3450	C	AIF1*825
2938	ATGGCGATATTGGTGAGAAAT	Allele	3470-3450	C	AIF1*825
2940	AGCATCTGCTGAGCTATGAG	Consensus	4289-4270	C	-
2941	CTGGGCCTTCAGCAGTCC	Allele	4052-4069	+	AIF1*206
2942	CTGGGCCTTCAGCAGTCT	Allele	4052-4069	+	AIF1*206

[a] Allele specific primers are grouped with their corresponding consensus primer

[b] Positions are given as co-ordinates in the TN62/TN82 sequence (HSY14768)

[c] + refers to primers in the forward orientation and C refers to primers in the reverse orientation

References

REFERENCES

Abderrahim, H., Sambucy, J. L., Iris, F., Ougen, P., Billault, A., Chumakov, I. M., Dausset, J., Cohen, D. & Le Paslier, D. (1994). Cloning the human major histocompatibility complex in YACs. *Genomics* **23** (3), 520-7.

Abraham, L. J., Du, D. C., Zahedi, K., Dawkins, R. L. & Whitehead, A. S. (1991). Haplotypic polymorphisms of the TNFB gene. *Immunogenetics* **33** (1), 50-3.

Abraham, L. J., Marley, J. V., Nedospasov, S. A., Cambon-Thomsen, A., Crouau-Roy, B., Dawkins, R. L. & Giphart, M. J. (1993). Microsatellite, restriction fragment-length polymorphism, and sequence-specific oligonucleotide typing of the tumor necrosis factor region. Comparisons of the 4AOHW cell panel. *Hum. Immunol.* **38** (1), 17-23.

Aguado, B. (1997). Sequence analysis and functional characterisation of three novel genes located in the human major histocompatibility complex. D.Phil. Department of Biochem. J., University of Oxford.

Aguado, B. & Campbell, R. D. (1995). The novel gene G17, located in the human major histocompatibility complex, encodes PBX2, a homeodomain-containing protein. *Genomics* **25** (3), 650-9.

Aguado, B. & Campbell, R. D. (1998). Characterization of a human lysophosphatidic acid acyltransferase that is encoded by a gene located in the class III region of the human major histocompatibility complex. *J. Biol. Chem.* **273** (7), 4096-105.

Aguado, B. & Campbell, R. D. (1999). Characterization of a human MHC class III region gene product with S-thioesterase activity. *Biochem. J.* **341** (Pt 3), 679-89.

Aguado, B., Milner, C. M. & Campbell, R. D. (1996). Genes of the MHC class III region and the functions of the proteins they encode. HLA and MHC: genes, molecules and function. Browning, M. &McMichael, A. Oxford, Bios Scientific Publishers, 39-75.

Albertella, M. R. & Campbell, R. D. (1994). Characterization of a novel gene in the human major histocompatibility complex that encodes a potential new member of the I kappa B family of proteins. *Hum. Mol. Genet.* **3** (5), 793-9.

Albertella, M. R., Jones, H., Thomson, W., Olavesen, M. & Campbell, R. D. (1996). Localization of Eight Additional Genes in the Human Major Histocompatibility Complex, Including the Gene Encoding the Casein Kinase II beta Subunit (CSNK2B). *Genomics* **36** , 240-251.

Alderton, R. P., Eccleston, L. M., Howe, R. P., Read, C. A., Reeve, M. A. & Beck, S. (1992). Magnetic bead purification of M13 DNA sequencing templates. *Analyt. Biochem. J.* **201** (1), 166-169.

Alonso, M. C., Navarrete, C., Solana, R., Torres, A., Pena, J. & Festenstein, H. (1985). Differential expression of HLA-DR and HLA-DQ antigens on normal cells of the myelomonocytic lineage. *Tissue Antigens* **26** , 310-317.

Amadou, C., Ribouchon, M. T., Mattei, M. G., Jenkins, N. A., Gilbert, D. J., Copeland, N. G., Avoustin, P. & Pontarotti, P. (1995). Localization of new genes and markers to the distal part of the human major histocompatibility complex (MHC) region and comparison with the mouse: new insights into the evolution of mammalian genomes. *Genomics* **26** (1), 9-20.

Andersson, G., Andersson, L., Larhammar, D., Rask, L. & Sigurdardottir, S. (1994). Simplifying genetic locus assignment of HLA-DRB genes. *Immunol. Today* **15** (2), 58-62.

Ando, H., Mizuki, N., Ota, M., Yamazaki, M., Ohno, S., Goto, K., Miyata, Y., Wakisaka, K., Bahram, S. & Inoko, H. (1997). Allelic variants of the human MHC class I chain-related B gene (MICB). *Immunogenetics* **46** (6), 499-508.

Apple, R. J. & Erlich, H. A. (1996). HLA class II genes: structure and diversity. HLA and MHC: genes, molecules and function. Browning, M. &McMichael, A. Oxford, Bios Scientific Publishers.

Arnett, F. C. & Reveille, J. D. (1992). Genetics of systemic lupus erythematosus. *Rheum. Dis. Clin. N. Am.* **18** , 865-892.

Arrick, B. A., Lee, A. L., Gendell, R. L. & Derynck, R. (1991). Inhibition of translation of transforming growth factor-beta-3 messenger RNA by its 5' untranslated region. *Molecular And Cellular Biology* **11** (9), 4306-4313.

Autieri, M. V. & Agrawal, N. (1998). IRT-1, a novel interferon-gamma-responsive transcript encoding a growth-suppressing basic leucine zipper protein. *J. Biol. Chem.* **273** (24), 14731-7.

Autieri, M. V., Feuerstein, G. Z., Yue, T.-L., Ohlstein, E. H. & Douglas, S. A. (1996a). cDNA cloning of human Allograft Inflammatory Factor-1: Tissue distribution, cytokine induction, and mRNA expression in injured rat carotid arteries. *Biochemical and Biophysical Research Communications* **228** (1), 29-37.

Autieri, M. V., Prystowsky, M. B. & Ohlstein, E. H. (1996b). Isolation and characterization of BART-1: A novel balloon angioplasty responsive transcript in rat carotid arteries. *DNA Cell Biol.* **15** (4), 297-304.

Avoustin, P., Ribouchon, M. T., Vernet, C., N'Guyen, B., Crouau Roy, B. & Pontarotti, P. (1994). Non-homologous recombination within the major histocompatibility complex creates a transcribed hybrid sequence. *Mamm. Genome.* **5** (12), 771-6.

Bach, F. H. & Van-Rood, J. J. (1976). The major histocompatibility complex - genetics and biology. *New England J. Med.* **295** (15), 806-813.

Baech, J., Schmidt Olsen, S., Steffensen, R., Varming, K., Grunnet, N. & Jersild, C. (1997). Frequency of HLA-B27 subtypes in a Danish population and in Danish patients with ankylosing spondylitis. *Tissue Antigens* **49** (5), 499-502.

Baeuerle, P. A. & Henkel, T. (1994). Function and activation of NFκB in the immune system. *Annu. Rev. Immunol.* **12** , 141-179.

Bahram, S., Bresnahan, M., Geraghty, D. E. & Spies, T. (1994). A second lineage of mammalian major histocompatibility complex class I genes [see comments]. *Proc. Natl. Acad. Sci. USA* **91** (14), 6259-63.

Balducci Silano, P. L. & Layrisse, Z. E. (1995). HLA-DP and susceptibility to insulin-dependent diabetes mellitus in an ethnically mixed population. Associations with other HLA-alleles. *J. Autoimmun.* **8** (3), 425-37.

Bankier, A. T., Weston, K. M. & Barrell, B. G. (1987). Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. *Methods Enzymol.* **155** , 51-93.

Barclay, A. N., Brown, M. H., Law, S. K. A., McKnight, A. J., Tomlinson, M. G. & van-der-Merwe, P. A., Eds. (1997). The Leucocyte Antigen Facts Book. Facts Book Series. London, Academic Press Limited.

Bateman, A. (1997). PhD Thesis. , Cambridge.

Bateman, A. & Chothia, C. (1995). Outline Structures for the Extracellular domains of the Fibroblast Growth FActor Receptors. *Nature Structural Biology* **2** (12), 1068-1074.

Beck, S. & Alderton, R. P. (1993). A strategy for the amplification, purification, and selection of M13 templates for large-scale DNA sequencing. *Analyt. Biochem. J.* **212** (2), 498-505.

Beck, S., Hanson, I., Kelly, A., Pappin, D. J. C. & Trowsdale, J. (1992). A homologue of the *Drosophila* female sterile homeotic (*fsh*) gene in the class II region of the human MHC. *DNA Sequence* **2** , 203-210.

Beck, S. & Trowsdale, J. (1999). Sequence organisation of the class II region of the human MHC. *Immunol. Rev.* **167** (February), 201-210.

Bernardi, G. (1995). The human genome: organization and evolutionary history. *Annu. Rev. Genet.* **29** , 445-76.

Binguilac, J., Postlethwait, J. H., Sato, A., Talbot, W. S., Johnson, S. L., Gates, M., Figueroa, F. & Klein, J. (1997). Mapping of the *Mhc* class I and class II regions to different chromosomes in the zebrafish, *Danio rerio*. *Immunogenetics* **46** , 129.

Bjorkman, P. J. & Parham, P. (1990). Structure, function, and diversity of class I major histocompatibility complex molecules. *Ann. Rev. Biochem.* **59** , 253-288.

Blanchard, H., Grochulski, P., Li, Y., Arthur, J. S., Davies, P. L., Elce, J. S. & Cygler, M. (1997). Structure of a calpain Ca^{2+}-binding domain reveals a novel EF-hand and Ca^{2+}-induced conformational changes. *Nat. Struct. Biol.* **4** (7), 532-8.

Bodmer, J. G., Marsh, S. G. & Albert, E. (1990). Nomenclature for factors of the HLA system. *Immunol. Today* **11** (1), 3-10.

Bodmer, J. G., Marsh, S. G. E., Albert, E. D., Bodmer, W. F., Bontrop, R. E., Charron, D., Dupont, B., Erlich, H. A., Fauchet, R., Mach, B., Mayr, W. R., Parham, P., Sasazuki, T., Schreuder, G. M. T., Strominger, J. L., Svejgaard, A. & Terasaki, P. I. (1997). Nomenclature for factors of the HLA system, 1996. *Tissue Antigens* **49** , 297-321.

Bourgois, A. (1975). Evidence for an ancestral immunoglobulin gene coding for half a domain. *Immunochemistry* **12** (11), 873-6.

Boyhan, A., Casimir, C. M., French, J. K., Teahan, C. G. & Segal, A. W. (1992). Molecular cloning and characterization of grancalcin, a novel EF-hand calcium-binding protein abundant in neutrophils and monocytes. *J. Biol. Chem.* **267** (5), 2928-2933.

Bradbury, A. F., Finnie, M. D. & Smyth, D. G. (1982). Mechanism of C-terminal amide formation by pituitary enzymes. *Nature* **298** (5875), 686-8.

Braud, V. M., Allan, D. S., O'Callaghan, C. A., Soderstrom, K., D'Andrea, A., Ogg, G. S., Lazetic, S., Young, N. T., Bell, J. I., Phillips, J. H., Lanier, L. L. & McMichael, A. J. (1998). HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature* **391** , 795-9.

Brennan, F. M., Maini, R. N. & Feldmann, M. (1992). TNF alpha--a pivotal role in rheumatoid arthritis? *Br. J. Rheumatol.* **31** (5), 293-8.

Bresnahan, P. A., Leduc, R., Thomas, L., Thorner, J., Gibson, H. L., Brake, A. J., Barr, P. J. & Thomas, G. (1990). Human fur gene encodes a yeast KEX2-like endoprotease that cleaves pro-beta-NGF in vivo. *J. Cell Biol.* **111** (6 Pt 2), 2851-9.

Brisseau, G. F., Grinstein, S., Hackman, D. J., Nordstrom, T., Manolson, M. F., Khine, A. A. & Rotstein, O. D. (1996). Interleukin-1 increases vacuolar-type H+-ATPase activity in murine peritoneal macrophages. *J. Biol. Chem.* **271** (4), 2005-2011.

Bristow, J., Gitelman, S. E., Tee, M. K., Staels, B. & Miller, W. L. (1993). Abundant adrenal-specific transcription of the human P450c21A "pseudogene". *J. Biol. Chem.* **268**, 12919-12924.

Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L. & Wiley, D. C. (1993). Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* **364** , 33-39.

Browning, J. L., Ngam-ek, A., Lawton, P., Demarinis, J., Tizard, R., Chow, E. P., Hession, C., O'brine-greco, B., Foley, S. F. & Ware, C. F. (1993). Lymphotoxin beta, a novel member of the TNF family that forms a heteromeric complex with lymphotoxin on the cell surface. *Cell* **72** (6), 847-856.

Bunce, M., Barnardo, M. C., Procter, J., Marsh, S. G., Vilches, C. & Welsh, K. I. (1996). High resolution HLA-C typing by PCR-SSP: identification of allelic frequencies and linkage disequilibria in 604 unrelated random UK Caucasoids and a comparison with serology [corrected and republished in Tissue Antigens 1997 Jul;50(1):100-11]. *Tissue Antigens* **48** (6), 680-91.

Bunce, M., Fanning, G. C. & Welsh, K. I. (1995). Comprehensive, serologically equivalent DNA typing for HLA-B by PCR using sequence-specific primers (PCR-SSP). *Tissue Antigens* **45** (2), 81-90.

Burfoot, M. S. & Campbell, R. D. (1994). Improved method of gene detection using exon amplification. *Nucl. Acids Res.* **22** , 5510-5511.

Busch, R., Doebele, R. C., von Scheven, E., Fahrni, J. & Mellins, E. D. (1998). Aberrant intermolecular disulfide bonding in a mutant HLA-DM molecule: implications for assembly, maturation, and function. *J. Immunol.* **160** (2), 734-43.

Busch, S. J. & Sassone Corsi, P. (1990). Dimers, leucine zippers and DNA-binding domains. *Trends Genet.* **6** (2), 36-40.

Campbell, R. D. (1993). The human Major Histocompatibility Complex: a 4000kb segment of the human genome replete with genes. Regional physical mapping. Davies, K. E. &Tilghman, S. M. New York, Cold Spring Harbor Laboratory Press. **5**, 1-33.

Campbell, R. D., Dunham, I., Kendall, E. & Sargent, C. A. (1990). Polymorphism of the human complement component C4. *Exp. clin. Immunogenet.* **7** (1), 69-84.

Campbell, R. D. & Law, S. K. A. (1992). C4 Complement protein. In: Human protein data. Weinheim, VCH.

Campbell, R. D. & Trowsdale, J. (1997). A Map Of The Human Major Histocompatibility Complex. *Immunol. Today* **18** (1), Centrefold.

Carosella, E. D., Dausset, J. & Kirszenbaum, M. (1996). HLA-G revisited. *Immunol. Today* **9** , 407-409.

Carrillo, M. B., Milner, C. M., Ball, S. T., Snoek, M. & Campbell, R. D. (1997). Cloning and characterization of a sialidase from the murine histocompatibility-2 complex: low levels of mRNA and a single amino acid mutation are responsible for reduced sialidase activity in mice carrying the Neu1a allele. *Glycobiology* **7** (7), 975-86.

Carroll, M. C., Katzman, P., Alicot, E. M., Koller, B. H., Geraghty, D. E., Orr, H. T., Strominger, J. L. & Spies, T. (1987). Linkage map of the human major histocompatibility complex including the tumor necrosis factor genes. *Proc. Natl. Acad. Sci. USA* **84** (23), 8535-9.

Cavener, D. R. (1987). Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. *Nucleic Acids Res.*. **15** (4), 1353-61.

Chang, H. Y., Nishitoh, H., Yang, X., Ichijo, H. & Baltimore, D. (1998). Activation of apoptosis signal-regulating kinase 1 (ASK1) by the adapter protein Daxx. *Science* **281** (5384), 1860-3.

Chen, X.-P., Enioutina, E. Y. & Daynes, R. A. (1997b). The control of IL-4 gene expression in activated murine T lymphocytes. *J. Immunol.* **158** , 3070-3080.

Chen, Z. W., Ahren, B., Ostenson, C. G., Cintra, A., Bergman, T., Moller, C., Fuxe, K., Mutt, V., Jornvall, H. & Efendic, S. (1997). Identification, isolation, and characterization of daintain (Allograft Inflammatory Factor-1), a macrophage polypeptide with effects on insulin secretion and abundantly present in the pancreas of prediabetic BB rats. *Prod. Natl. Acad. Sci. USA* **94** (25), 13879-13884.

Chiquet Ehrismann, R., Hagios, C. & Matsumoto, K. (1994). The tenascin gene family. *Perspect. Dev. Neurobiol.* **2** (1), 3-7.

Chiu, H. C., Hsieh, R. P., Hsieh, K. H. & Hung, T. P. (1987). Association of HLA-DRw9 with myasthenia gravis in Chinese. *J. Immunogenet.* **14** (4-5), 203-7.

Christiansen, F. T., Zhang, W. J., Griffiths, M., Mallal, S. A. & Dawkins, R. L. (1991). Major histocompatibility complex (MHC) complement deficiency, ancestral haplotypes and systemic lupus erythematosus (SLE): C4 deficiency explains some but not all of the influence of the MHC. *J. Rheumatol.* **18** (9), 1350-8.

Cohen, R. B., Boal, T. R. & Safer, B. (1990). Increased eIF-2-alpha expression in mitogen-activated primary T lymphocytes. *EMBO J.*. **9** (12), 3831-3838.

Collins, F. S., Brooks, L. D. & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8** (12), 1229-31.

Crider, B. P., Andersen, P., White, A. E., Zhou, Z., Li, X., Mattsson, J. P., Lundberg, L., Keeling, D. J., Xie, X. S., Stone, D. K. & Peng, S. B. (1997). Subunit G of the vacuolar proton pump. Molecular characterization and functional expression. *J. Biol. Chem.* **272** (16), 10721-8.

Cross, S. J. & Thomson, W. (1990). DNA polymorphism of the C2 and factor B gene. *Exp. Clin. Immunogenetics* **7** (1), 53-63.

Cucca, F. & Todd, T. A. (1996). HLA susceptibility to type I diabetes: methods and mechanisms. HLA and MHC genes, molecules and function. Browning, M. &McMichael, A. Oxford, Bios Scientific Publishers Ltd.

D'Alfonso, S. M. P. (1996). A New polymorphism in the TNFA gene (Abstract). *Hum. Immunol.* **47** , 65.

Davies, J. L., Kawaguchi, Y., Bennett, S. T., Copeman, J. B., Cordell, H. J., Pritchard, L. E., Reed, P. W., Gough, S. C., Jenkins, S. C., Palmer, S. M. & et al. (1994). A genome-wide search for human type 1 diabetes susceptibility genes [see comments]. *Nature* **371** (6493), 130-6.

Dawkins, R., Leelayuwat, C., Gaudiari, S., Tay, G., Hui, J., Cattley, S., Martinez, P. & Kulski, J. (1999). Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol. Rev.* **167** (February), 275-304.

de Baey, A., Fellerhoff, B., Maier, S., Martinozzi, S., Weidle, U. & Weiss, E. H. (1997). Complex expression pattern of the TNF region gene LST1 through differential regulation, initiation, and alternative splicing. *Genomics* **45** (3), 591-600.

de Baey, A., Fellerhoff, B., Maier, S. & Weiss, E. H. (1996). Tissue specific expression and alternative splicing of the LST1 gene of the TNF region. *Hum. Immunol.* **47** (1-2), 65.

de Baey, A., Holzinger, I., Scholz, S., Keller, E., Weiss, E. H. & Albert, E. (1995). Pvu II Polymorphism in the Primate Homologue of the Mouse B144 (LST-1): A Novel Marker Gene Within the Tumor Necrosis Factor Region. *Hum. Immunol.* **42** (1), 9-14.

Degli Esposti, M. A., Andreas, A., Christiansen, F. T., Schalke, B., Albert, E. & Dawkins, R. L. (1992b). An approach to the localization of the susceptibility genes for generalized myasthenia gravis by mapping recombinant ancestral haplotypes. *Immunogenetics* **35** (6), 355-64.

Degli Esposti, M. A., Leaver, A. L., Christiansen, F. T., Witt, C. S., Abraham, L. J. & Dawkins, R. L. (1992a). Ancestral haplotypes: conserved population MHC haplotypes. *Hum. Immunol.* **34** (4), 242-52.

Denzin, L. K., Hammond, C. & Cresswell, P. (1996). HLA-DM interactions with intermediates in HLA-DR maturation and a role for HLA-DM in stabilizing empty HLA-DR molecules. *J. Exp. Med.* **184** (6), 2153-65.

Denzin, L. K., Sant'Angelo, D. B., Hammond, C., Surman, M. J. & Cresswell, P. (1997). Negative regulation by HLA-DO of MHC class II-restricted antigen processing. *Science* **278** , 106-109.

Dixon, M., Deed, R., Acland, P., Moore, R., Whyte, A., Peters, G. & Dickson, C. (1989). Detection and characterization of the fibroblast growth factor-related oncoprotein INT-2. *Mol. Cell Biol.* **9** (11), 4896-902.

Dodds, A. W., Law, S. K. & Porter, R. R. (1985). The origin of the very variable haemolytic activities of the common human complement component C4 allotypes including C4-A6. *EMBO J.* **4** (9), 2239-44.

Douek, D. C. & Altmann, D. M. (1997). HLA-DO is an intracellular class II molecule with distinctive thymic expression. *Int. Immunol.* **9** (3), 355-64.

Dunham, I., Sargent, C. A., Trowsdale, J. & Campbell, R. D. (1987). Molecular mapping of the human major histocompatibility complex by pulsed-field gel electrophoresis. *Proc. Natl. Acad. Sci. USA* **84** (20), 7237-41.

Eberhardt, C., Gray, P. W. & Tjoelker, L. W. (1997). Human lysophosphatidic acid acyltransferase. cDNA cloning, expression, and localization to chromosome 9q34.3. *J. Biol. Chem.* **272** , 20299-20305.

Edelman, G. M. (1970). The covalent structure of a human gamma G-immunoglobulin. XI. Functional implications. *Biochem. J.* **9** (16), 3197-205.

Edmundson, A. B., Ely, K. R., Abola, E. E., Schiffer, M., Panagiotopoulos, N. & Deutsch, H. F. (1976). Conformational isomerism, rotational allomerism, and divergent evolution in immunoglobulin light chains. *Fed. Proc.* **35** (10), 2119-23.

Edwards, S. V., Hess, M. C., Gasper, J. & Garrigan, D. (1999). Towards an evolutionary genomics of the avian MHC. *Immunol. Rev.* **167** (February), 119-132.

Endo, T., Imanishi, T., Gojobori, T. & Inoko, H. (1997). Evolutionary significance of intra-genome duplications on human chromosomes. *Gene* **205** , 19-27.

Fan, W., Liu, Y.-C., Parimoo, S. & Weissman, S. M. (1995). Olfactory receptor-like genes are located in the human major histocompatibility complex. *Genomics* **27** , 119-123.

Fan, W. F., Wei, X., Shukla, H., Parimoo, S., Xu, H., Sankhavaram, P., Li, Z. & Weissman, S. M. (1993). Application of cDNA selection techniques to regions of the human MHC. *Genomics.* **17** (3), 575-581.

Feder, J. N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D. A., Basava, A., Dormishian, F., Domingo, R., Jr., Ellis, M. C., Fullan, A., Hinton, L. M., Jones, N. L., Kimmel, B. E., Kronmal, G. S., Lauer, P., Lee, V. K., Loeb, D. B., Mapa, F. A., McClelland, E., Meyer, N. C., Mintier, G. A., Moeller, N., Moore, T., Morikang, E., Wolff, R. K. & et al. (1996). A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat. Genet.* **13** (4), 399-408.

Feinberg, A. P. & Vogelstein, B. (1984). "A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity". Addendum. *Anal. Biochem.* **137** (1), 266-7.

Fortini, M. E. & Artavanis-Tsakonas, S. (1993). Notch: neurogenesis is only part of the picture. *Cell* **75** , 1245-1247.

Fourney, R., M., Miyakoshi, J., Day, R., S. & Paterson, M., C. (1988). *Focus (Bethesda Research Laboratories)* **10** , 5-7.

Froussard, P., Chastagner, P., Somme, G., Abadie, A., Greene, W., Theze, J. & Longacre, S. (1988). P55 IL-2 receptor messenger RNA precursors in murine T lymphocyte nuclei. *Journal Of Immunology* **141** (4), 1358-1364.

Fukagawa, T., Sugaya, K., Matsumoto, K., Okumura, K., Ando, A., Inoko, H. & Ikemura, T. (1995). A boundary of long-range G+C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary. *Genomics.* **25** , 184-191.

Geraghty, D. E. (1993). Structure of the HLA class I region and expression of its resident genes. *Curr. Opin. Immunol,* **5** (1), 3-7.

Geraghty, D. E., Koller, B. H., Handsen, J. A. & Orr, H. T. (1992). The HLA class I gene family includes at least six genes and twelve pseudogenes and gene fragments. *J. Immunol.* **149** , 1934-1946.

Germain, R. N. (1994). MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation. *Cell* **76** (2), 287-99.

Giffon, T., Lepourcelet, M., Pichon, L., Jezequel, P., Bouric, P., Carn, G., Pontarotti, P., Gall, J. Y. L. & David, V. (1996). Cloning of a human homologue of the mouse Tctex-5 gene within the MHC class I region. *Immunogenetics* **44** (5), 331-9.

Glynne, R., Kerr, L. A., Mockridge, I., Beck, S., Kelly, A. & Trowsdale, J. (1993). The MHC-encoded proteosome component LMP7: alternative first exons and post-translational processing. *Eur. J. Immunol.* **23** , 860-866.

Goldberg, A. (1995). Functions of the proteosome: the lysis at the end of the tunnel. *Science* **268** , 522-523.

Goldsworthy, M., Sampath, A., Wilkinson, J. M. & Powis, S. H. (1997). Identification of new HLA class I region genes by sample sequencing. *Immunogenetics* **46** (3), 206-12.

Gomez Escobar, N., Chou, C. F., Lin, W. W., Hsieh, S. L. & Campbell, R. D. (1998). The G11 gene located in the major histocompatibility complex encodes a novel nuclear serine/threonine protein kinase. *J. Biol. Chem.* **273** (47), 30954-60.

Gongora, R., Figueroa, F. & Klein, J. (1997). Complex origin of the HLA-DR10 haplotype. *J. Immunol.* **159** (12), 6044-51.

Gorer, P. A. (1937). *J. Path. Bact.* **44** , 691-697.

Gorer, P. A., Lyman, S. & Snell, G. D. (1948). *Proc. Royal Soc. London* **135** , 499-505.

Gout, I., Dhand, R., Hiles, I. D., Fry, M. J., Panayotou, G., Das, P., Truong, O., Totty, N. F. & Hsuan, J. (1993). The GTPase dynamin binds to and is activated by a subset of SH3 domains. *Cell* **75** (1), 25-36.

Green, S., Walter, P., Kumar, V., Krust, A., Bornert, J. M., Argos, P. & Chambon, P. (1986). Human oestrogen receptor cDNA: sequence, expression and homology to v-erb-A. *Nature* **320** (6058), 134-9.

Griffith, E., Ramsburg, E. & Hayday, A. (1998). Recognition by human gut gamma delta cells of stress inducible major histocompatibility molecules on enterocytes. *Gut* **43** (2), 166-7.

Grimaldi, M. C., Clayton, J., Pontarotti, P., Cambon-Thomsen, A. & Crouau-Roy, B. (1996). New highly polymorphic microsatellite marker in linkage disequilibrium with HLA-B. *Hum. Immunol.* **51** (2), 89-94.

Grinberg, D., Thurlow, J., Watson, R., Smith, R., Peters, G. & Dickson, C. (1991). Transcriptional regulation of the int-2 gene in embryonal carcinoma cells. *Cell Growth Differ.* **2** (3), 137-43.

Groh, V., Bahram, S., Bauer, S., Herman, A., Beauchamp, M. & Spies, T. (1996). Cell stress-regulated human major histocompatibility complex class I gene expressed in gastrointestinal epithelium. *Proc. Natl. Acad. Sci. USA* **93** (22), 12445-50.

Gruen, J. R. & Weissman, S. M. (1997). Evolving views of the major histocompatibility complex. *Blood* **90** (11), 4252-65.

Guillaudeux, T., Janer, M., Wong, G. K., Spies, T. & Geraghty, D. E. (1998). The complete genomic sequence of 424,015 bp at the centromeric end of the HLA class I region: gene content and polymorphism. *Proc. Natl. Acad. Sci. USA* **95** (16), 9494-9.

Guss, H.-J. & Dower, S. (1995). Tumor necrosis factor ligand superfamily: Involvement in the pathology of malignant lymphomas. *Blood.* **85** (12), 3378-3404.

Hall, M. A., Lanchbury, J. S. & Ciclitira, P. J. (1996). HLA class II region genes and susceptibility to dermatitis herpetiformis: DPB1 and TAP2 associations are secondary to those of the DQ subregion. *Eur. J. Immunogenet.* **23** (4), 285-96.

Hanifi Moghaddam, P., de Knijf, P., Roep, B. O., Van der Auwera, B., Naipal, A., Gorus, F., Schuit, F. & Giphart, M. J. (1998). Genetic structure of IDDM1: two separate regions in the major histocompatibility complex contribute to susceptibility or protection. Belgian Diabetes Registry. *Diabetes* **47** (2), 263-9.

Harpaz, Y. & Chothia, C. (1994). Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J. Mol. Biol.* **238** (4), 528-39.

Harvey, W. R. (1992). Physiology of V-ATPases. *J. Exp. Biol* **172** , 1-17.

Hauptmann, G., Tappeiner, G. & Schifferli, J. A. (1993). Inherited deficieny of the fourth component of human complement. Immunodeficiencies. Rosen, F. S. &Seligmann, M., Harwood Academic Publishers, 253-266.

Henderson, J., Salzberg, S. & Fasman, K. H. (1997). Finding genes in DNA with a Hidden Markov Model. *J. Comput. Biol.* **4** (2), 127-41.

Herberg, J. A., Beck, S. & Trowsdale, J. (1998). TAPASIN, DAXX, RGL2, HKE2 and four new genes (BING 1, 3 to 5) form a dense cluster at the centromeric end of the MHC. *J. Mol. Biol.* **277** (4), 839-857.

Higashi, Y., Yoshioka, H., Yamane, M., Gotoh, O. & Fujii Kuriyama, Y. (1986). Complete nucleotide sequence of two steroid 21-hydroxylase genes tandemly arranged in human chromosome: a pseudogene and a genuine gene. *Proc. Natl. Acad. Sci. USA* **83** (9), 2841-5.

Hill, R. L., Delaney, R., Fellows, R. E. & Lebovitz, H. E. (1966). The Evolutionary Origins of the Immunoglobulins. *Biochem. J.* **56** , 1762-1769.

Holzinger, I., de Baey, A., Messer, G., Kick, G., Zwierzina, H. & Weiss, E. H. (1995). Cloning and genomic characterization of LST1: a new gene in the human TNF region. *Immunogenetics* **42** (5), 315-322.

Hoopes, C. W., Taketo, M., Ozato, K., Liu, Q., Howard, T. A., Linney, E. & Seldin, M. F. (1992). Mapping of the mouse Rxr loci encoding nuclear retinoid X receptors RXR alpha, RXR beta, and RXR gamma. *Genomics* **14** (3), 611-7.

Horton, R., Niblett, D., Milne, S., Palmer, S., Tubby, B., Trowsdale, J. & Beck, S. (1998). Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. *J. Mol. Biol.* **282** (1), 71-97.

Hsieh, S. L. & Campbell, R. D. (1991). Evidence that gene G7a in the human major histocompatibility complex encodes valyl-tRNA synthetase [published erratum appears in Biochem. J. 1992 Feb 1;281(Pt 3):879]. *Biochem. J.* **278** (Pt 3), 809-16.

Hsieh, S. L., March, R. E., Khanna, A., Cross, S. J. & Campbell, R. D. (1997). Mapping of 10 novel microsatellites in the MHC class III region: application to the study of autoimmune disease. *J. Rheumatol.* **24** (1), 220-2.

Hughes, A. L. (1998). Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Mol. Biol. Evol.* **15** (7), 854-70.

Huppi, K., Duncan, R. & Potter, M. (1988). Myc-1 is centromeric to the linkage group Ly-6--Sis--Gdc-1 on mouse chromosome 15. *Immunogenetics* **27** (3), 215-9.

Iida, Y. & Masuda, T. (1996). Strength of translation initiation signal sequence of mRNA as studied by quantification method: effect of nucleotide substitutions upon translation efficiency in rat preproinsulin mRNA. *Nucleic Acids Res.* **24** (17), 3313-6.

Imai, Y., Ibata, I., Ito, D., Ohsawa, K. & Kohsaka, S. (1996). A novel gene iba1 in the major histocompatibility complex class III region encoding an EF hand protein expressed in a monocytic lineage. *Biochem. Biophys. Res. Commun.* **224** (3), 855-62.

Iris, F. J., Bougueleret, L., Prieur, S., Caterina, D., Primas, G., Perrot, V., Jurka, J., Rodriguez-Tome, P., Claverie, J. M. & Dausset, J. (1993). Dense Alu clustering and a potential new member of the NF kappa B family within a 90 kilobase HLA class III segment. *Nat. Genet.* **3** (2), 137-45.

Irminger, J. C., Rosen, K. M., Humbel, R. E. & Villa Komaroff, L. (1987). Tissue-specific expression of insulin-like growth factor II mRNAs with distinct 5' untranslated regions [published erratum appears in Proc. Natl. Acad. Sci. USA 1988 Feb;85(4):1070]. *Proc. Natl. Acad. Sci. USA* **84** (18), 6330-4.

Israel, A. & Kourilsky, P. (1996). Regulation of MHC class I gene expression. *HLA and MHC: genes, molecules and function.* Browning, M. &McMichael, A. Oxford, Bios Scientific Publishers.

Jenisch, S., Henseler, T., Nair, R. P., Guo, S. W., Westphal, E., Stuart, P., Kronke, M., Voorhees, J. J., Christophers, E. & Elder, J. T. (1998). Linkage analysis of human leukocyte antigen (HLA) markers in familial psoriasis: strong disequilibrium effects provide evidence for a major determinant in the HLA-B/-C region. *Am. J. Hum. Genet.* **63** (1), 191-9.

Juji, T., Satake, M., Honda, Y. & Doi, Y. (1984). HLA antigens in Japanese patients with narcolepsy. All patients were DR2 positive. *Tissue Antigens* **24** , 316-319.

Jurka, J. & Smith, T. (1988). A fundamental division in the Alu family of repeated sequences. *Proc. Natl. Acad. Sci. USA* **85** (13), 4775-8.

Kallenberg, C. G., Van der Voort Beelen, J. M., D'Amaro, J. & The, T. H. (1981). Increased frequency of B8/DR3 in scleroderma and association of the haplotype with impaired cellular immune response. *Clin. Exp. Immunol.* **43** (3), 478-85.

Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87** (6), 2264-8.

Kasahara, M. (1997). New insights into the genomic organization and origin of the major histocompatibility complex: role of chromosomal (genome) duplication in the emergence of the adaptive immune system. *Hereditas* **127** (1-2), 59-65.

Kasahara, M. (1999). The chromosomal duplication model of the major histocompatibility complex. *Immunol. Rev.* **167** (February), 17-32.

Kaufman, J., Jacob, J., Shaw, I., Walker, B., Milne, S., Beck, S. & Salomonsen, J. (1999). Gene organisation determines evolution of function in the chicken MHC. *Immunol. Rev.* **167** (February), 101-118.

Keet, I. P., Klein, M. R., Just, J. J. & Kaslow, R. A. (1996). The role of host genetics in the natural history of HIV-1 infection: the needles in the haystack. *AIDS* **10 Suppl A** , S59-67.

Kelly, A. P., Monaco, J. J., Cho, S. & Trowsdale, J. (1991). A new human HLA class II-related locus, DM. *Nature* **353** , 571-573.

Kelly, H., McCann, V. J., Kay, P. H. & Dawkins, R. L. (1985). Susceptibility to IDDM is marked by MHC supratypes rather than individual alleles. *Immunogenetics* **22** (6), 643-51.

Kendall, E., Sargent, C. A. & Campbell, R. D. (1990). Human major histocompatibility complex contains a new cluster of genes between the HLA-D and complement C4 loci. *Nucleic Acids Research* **18** (24), 7251-7258.

Klein, J., Figueroa, F. & Nagy, Z. A. (1983). Genetics of the major histocompatibility complex: the final act. *Annu. Rev. Immunol.* **1** , 119-42.

Klein, J. & Sato, A. (1998). Birth of the major histocompatibility complex. *Scand. J. Immunol.* **47** (3), 199-209.

Kobilka, B. K., MacGregor, C., Daniel, K., Kobilka, T. S., Caron, M. G. & Lefkowitz, R. J. (1987). Functional activity and regulation of human beta 2-adrenergic receptors expressed in Xenopus oocytes. *J. Biol. Chem.* **262** (32), 15796-802.

Koldin, B., Suckow, M., Seydel, A., von Wilcken Bergmann, B. & Muller Hill, B. (1995). A comparison of the different DNA binding specificities of the bZip proteins C/EBP and GCN4. *Nucleic Acids Res.*. **23** (20), 4162-9.

Korenberg, J. R. & Rykowski, M. C. (1988). Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* **53** (3), 391-400.

Kozak, M. (1984). Selection of initiation sites by eucaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin. *Nucleic Acids Res.*. **12** (9), 3873-93.

Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44** (2), 283-92.

Kozak, M. (1991). An analysis of vertebrate mRNA sequences: intimations of translational control. *J. Cell Biol.* **115** (4), 887-903.

Krebs, J., Quadroni, M. & Van-Eldik, L. J. (1995). Dance of the dimers. *Nat. Struct. Biol.* **2** (9), 711-4.

Kronenberg, M., Brines, R. & Kaufman, J. (1994). MHC evolution: a long term investment in defense [see comments]. *Immunol. Today* **15** (1), 4-6.

Kropshofer, H., Vogt, A. B., Moldenhauer, G., Hammer, J., Blum, J. S. & Hammerling, G. J. (1996). Editing of the HLA-DR-peptide repertoire by HLA-DM. *EMBO J.* **15** (22), 6144-54.

Kulski, J. K. & Dawkins, R. L. (1999). The P5 multicopy gene family in the MHC is related in sequence to human endogenous retroviruses HERV-L and HERV-16. *Immunogenetics* **49** (5), 404-12.

Landschulz, W. H., Johnson, P. F. & McKnight, S. L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science* **240** (4860), 1759-64.

Law, S. K. E. & Reid, K. B. M. (1995). Complement. Oxford, IRL Press.

Lechler, R. (1994). HLA and Disease. London, Academic Press.

LeClair, K. P., Rabin, M., Nesbitt, M. N., Pravtcheva, D., Ruddle, F. H., Palfree, R. G. & Bothwell, A. (1987). Murine Ly-6 multigene family is located on chromosome 15. *Proc. Natl. Acad. Sci. USA* **84** (6), 1638-43.

Lee, K. H. & Cotanche, D. A. (1995). Detection of beta-actin mRNA by RT-PCR in normal and regenerating chicken cochleae. *Hear. Res.* **87** (1-2), 9-15.

Lepier, A., Graf, R., Azuma, M., Merzendorfer, H., Harvey, W. R. & Wieczorek, H. (1996). The peripheral complex of the tobacco hornworm V-ATPase contains a novel 13-kDa subunit G. *J. Biol. Chem.* **271** (14), 8502-8508.

Leung, S. O., Whitelaw, E. & Proudfoot, N. J. (1989). Transcriptional and translational analysis of the human theta globin gene. *Nucleic Acids Res.*. **17** (20), 8283-300.

Li, S., Paulsson, K. M., Sjogren, H. O. & Wang, P. (1999). Peptide-bound major histocompatibility complex class I molecules associate with tapasin before dissociation from transporter associated with antigen processing. *J. Biol. Chem.* **274** (13), 8649-54.

Liang, S., Hitomi, M., Hu, Y. H., Liu, Y. & Tartakoff, A. M. (1996). A DEAD-box-family protein is required for nucleocytoplasmic transport of yeast mRNA. *Mol. Cell Biol.* **16** (9), 5139-46.

Lie, B. A., Akselsen, H. E., Joner, G., Dahl Jorgensen, K., Ronningen, K. S., Thorsby, E. & Undlien, D. E. (1997). HLA associations in insulin-dependent diabetes mellitus: no independent association to particular DP genes. *Hum. Immunol.* **55** (2), 170-5.

Lui, V. C., Ng, L. J., Sat, E. W. & Cheah, K. S. (1996). The human alpha 2(XI) collagen gene (COL11A2): completion of coding information, identification of the promoter sequence, and precise localization within the major histocompatibility complex reveal overlap with the KE5 gene. *Genomics* **32** (3), 401-12.

Mackay, F., Majeau, G. R., Hochman, P. S. & Browning, J. L. (1996). Lymphotoxin beta receptor triggering induces activation of the nuclear factor kappaB transcription factor in some cell types. *J. Biol. Chem.* **271** (40), 24934-8.

Maini, R. N., Elliott, M. J., Brennan, F. M., Williams, R. O., Chu, C. Q., Paleolog, E., Charles, P. J., Taylor, P. C. & Feldmann, M. (1995). Monoclonal anti-TNF alpha antibody as a probe of pathogenesis and therapy of rheumatoid disease. *Immunol. Rev.* **144** , 195-223.

Maldonado, F. & Hanks, S. K. (1988). A cDNA clone encoding human cAMP-dependent protein kinase catalytic subunit C alpha. *Nucleic Acids Res.*. **16** (16), 8189-90.

Manzella, J. M. & Blackshear, P. J. (1990). Regulation of rat ornithine decarboxylase mRNA translation by its 5'-untranslated region. *J. Biol. Chem.* **265** (20), 11817-22.

McGuire, W., Hill, A. V. S., Allsopp, C. E. M., Greenwood, B. M. & Kwiatkowski, D. (1994). Variation in the TNF-a promoter region associated with susceptibility to cerebral malaria. *Nature* **371** , 508-511.

McGuire, W., Knight, J. C., Hill, A. V., Allsopp, C. E., Greenwood, B. M. & Kwiatkowski, D. (1999). Severe malarial anemia and cerebral malaria are associated with different tumor necrosis factor promoter alleles. *J. Infect. Dis.* **179** (1), 287-90.

Messer, G., Spengler, U., Jung, M. C., Honold, G., Blomer, K., Pape, G. R., Riethmuller, G. & Weiss, E. H. (1991). Polymorphic structure of the tumor necrosis factor (TNF) locus: an NcoI polymorphism in the first intron of the human TNF-beta gene correlates with a variant amino acid in position 26 and a reduced level of TNF-beta production. *J. Exp. Med.* **173** (1), 209-19.

Messer, G., Zemmour, J., Orr, H. T., Parham, P., Weiss, E. H. & Girdlestone, J. (1992). HLA-J, a second inactivated class I HLA gene related to HLA-G and HLA-A. Implications for the evolution of the HLA-A-related genes. *J. Immunol.* **148** (12), 4043-53.

Meyer, R., Hatada, E. N., Hohmann, H. P., Haiker, M., Bartsch, C., Rothlisberger, U., Lahm, H. W., Schlaeger, E. J., van Loon, A. P. & Scheidereit, C. (1991). Cloning of the DNA-binding subunit of human nuclear factor kappa B: the level of its mRNA is strongly regulated by phorbol ester or tumor necrosis factor alpha. *Proc. Natl. Acad. Sci. USA* **88** (3), 966-70.

Mikzuki, N., Ando, H., Kimura, M., Ohno, S., Miyata, S., Yamazaki, M., Tashiro, H., Watanabe, K., Ono, A., Taguchi, S., Sugawara, C., Fukuzumi, Y., Okumura, K., Goto, K., Ishihara, M., Nakamura, S., Yonemoto, J., Kikuti, Y. Y., Shiina, T., Chen, L., Ando, A., Ikemura, T. & Inoko, H. (1997). Nucleotide sequence analysis of the HLA class I region spanning the 237 kb segment around the HLA-B and -C genes. *Genomics* **42** (1), 55-66.

Miller, W. L. (1988). Molecular biology of steroid hormone synthesis. *Endocr. Rev.* **9** (3), 295-318.

Milner, C. M. & Campbell, R. D. (1990). Structure and expression of the three MHC-linked HSP70 genes. *Immunogenetics* **32** (4), 242-251.

Milner, C. M. & Campbell, R. D. (1993). The G9a gene in the human major histocompatibility complex encodes a novel protein containing ankyrin-like repeats [published erratum appears in Biochem. J. 1993 Jun 15;292(Pt 3):952]. *Biochem. J.* **290** (Pt 3), 811-8.

Milner, C. M., Smith, S. V., Carrillo, M. B., Taylor, G. L., Hollinshead, M. & Campbell, R. D. (1997). Identification of a sialidase encoded in the human major histocompatibility complex. *J. Biol. Chem.* **272** (7), 4549-58.

Mizuki, N., Ohno, S., Sato, T., Ishihara, M., Miyata, S., Nakamura, S., Naruse, T., Mizuki, H., Tsuji, K. & Inoko, H. (1995). Microsatellite polymorphism between the tumor necrosis factor and HLA-B genes in Behcet's disease. *Hum. Immunol.* **43** (2), 129-35.

Moghaddam, P. H., Zwinderman, A. H., de Knijff, P., Roep, B. O., Schipper, R. F., Van der Auwera, B., Naipal, A., Gorus, F., Schuit, F. & Giphart, M. J. (1997). TNFa microsatellite polymorphism modulates the risk of IDDM in Caucasians with the high-risk genotype HLA DQA1*0501-DQB1*0201/DQA1*0301-DQB1*0302. Belgian Diabetes Registry. *Diabetes* **46** (9), 1514-5.

Morel, Y. & Miller, W. L. (1991). Clinical and molecular genetics of congenital adrenal hyperplasia due to 21-hydroxylase deficiency. *Adv. Hum. Genet.* **20** , 1-68.

Moseley, P. L., Wallen, E. S., McCafferty, J. D., Flanagan, S. & Kern, J. A. (1993). Heat stress regulates the human 70 kDa heat shock gene through the 3' untranslated region. *Am. J. Physiol.* **264** , L533-L537.

Mueller, C. R., Maire, P. & Schibler, U. (1990). DBP, a liver-enriched transcriptional activator, is expressed late in ontogeny and its tissue specificity is determined posttranscriptionally. *Cell* **61** (2), 279-291.

Muller, A. J. & Witte, O. N. (1989). The 5' noncoding region of the human leukemia-associated oncogene BCR/ABL is a potent inhibitor of in vitro translation. *Mol. Cell Biol.* **9** (11), 5234-8.

Mullighan, C. G., Marshall, S. E., Bunce, M. & Welsh, K. I. (1999). Variation in immunoregulatory genes determeines the clinical phenotype of common variable immuodeficiency. *Genes and Immunity* **1** , 137-148.

Murray, A. J., Lewis, S. J., Barclay, A. N. & Brady, R. L. (1995). One sequence, two folds: a metastable structure of CD2. *Proc. Natl. Acad. Sci. USA* **92** (16), 7337-41.

Nair, S., Dey, R., Sanford, J. P. & Doyle, D. (1992). Molecular cloning and analysis of an eIF-4A-related rat liver nuclear protein. *J. Biol. Chem.* **267** (18), 12928-35.

Nalabolu, S. R., Shukla, H., Nallur, G., Parimoo, S. & Weissman, S. M. (1996). Genes in a 220-kb region spanning the TNF cluster in human MHC. *Genomics* **31** (2), 215-222.

Namen, A. E., Lupton, S., Hjerrild, K., Wignall, J., Mochizuki, D. Y., Schmierer, A., Mosley, B., March, C. J., Urdal, D. & Gillis, S. (1988). Stimulation of B-cell progenitors by cloned murine interleukin-7. *Nature* **333** (6173), 571-573.

Nedwin, G. E., Naylor, S. L., Sakaguchi, A. Y., Smith, D., Jarrett-Nedwin, J., Pennica, D., Goeddel, D. V. & Gray, P. W. (1985). Human lymphotoxin and tumor necrosis factor genes: structure, homology and chromosomal localization. *Nucleic Acids Res.*. **13** (17), 6361-73.

Neeper, M., Schmidt, A. M., Brett, J., Yan, S. D., Wang, F., Pan, Y. C., Elliston, K., Stern, D. & Shaw, A. (1992). Cloning and expression of a cell surface receptor for advanced glycosylation end products of proteins. *J. Biol. Chem.* **267** , 14998-15004.

Neville, M. J. & Campbell, R. D. (1997). Alternative Splicing of the LST-1 Gene Located in the Major Histocompatibility Complex on Human Chromosome 6. *DNA Sequencing-The Journal of Sequencing and Mapping.* **8** (3), 155-160.

Neville, M. J. & Campbell, R. D. (1999). A new member of the Ig superfamily and a V-ATPase G-subunit are among the predicted products of novel genes close to the TNF locus in the human MHC. *J. Immunology* **162** , 4745-4754.

Novotny, V., Prieschl, E. E., Csonga, R., Fabjani, G. & Baumruker, T. (1998). Nrf1 in a complex with fosB, c-jun, junD and ATF2 forms the AP1 component at the TNF alpha promoter in stimulated mast cells. *Nucleic Acids Res.*. **26** (23), 5480-5.

Olavesen, M., Thomson, W., Cheng, J. & Campbell, R. D. (1991). Characterisation of a novel gene G1 in the human MHC class III region. Proceedings of the Eleventh International Histocompatibility Workshop and Conference., Yokohama, Japan, Oxford University Press. **2,** 190-3.

Olavesen, M. G., Snoek, M. & Campbell, R. D. (1993). Localization of a new gene adjacent to the HSP70 genes in human and mouse MHCs. *Immunogenetics* **37** (5), 394-6.

Olerup, O., Smith, C. I. & Hammarstrom, L. (1990). Different amino acids at position 57 of the HLA-DQ beta chain associated with susceptibility and resistance to IgA deficiency. *Nature* **347** (6290), 289-90.

Orkin, S. H., Cheng, T. C., Antonarakis, S. E. & Kazazian, H. H., Jr. (1985). Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. *EMBO J.* **4** (2), 453-6.

Ortmann, B., Androlewicz, M. J. & Cresswell, P. (1994). MHC class I/beta2-microglobulin complexes associate with TAP transporters before peptide binding. *Nature* **368** , 864-867.

Ott, J. (1991). Analysis of human genetic linkage, The John Hopkins University Press.

Pangburn, M. K. (1992). Factor B. In: Human Protein Data. Weinheim, VCH.

Parham, P. (1990). Antigen processing. Transporters of delight [news; comment]. *Nature* **348** (6303), 674-5.

Parham, P. (1999). Virtual reality in the MHC. *Immunol. Rev.* **167** (February), 5-15.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85** (8), 2444-8.

Peelman, L. J., Chardon, P., Nunes, M., Renard, C., Geffrotin, C., Vaiman, M., Van Zeveren, A., Coppieters, W., Van de Weghe, A., Bouquet, Y., Choy, W. W., Strominger, J. L. & Spies, T. (1995). The BAT1 gene in the MHC encodes an evolutionarily conserved putative nuclear RNA helicase of the DEAD family. *Genomics.* **26** , 210-218.

Pende, D., Parolini, S., Pessino, A., Sivori, S., Augugliaro, R., Morelli, L., Marcenaro, E., Accame, L., Malaspina, A., Biassoni, R., Bottini, C., Moretta, L. & Moretta, A. (1999). Identification and Molecular Characterisation of NKp30, a Novel Triggering Receptor Involved in Natural Cytotoxicity Mediated by Human Natural Killer Cells. *J. Exp. Med.* **190** (10), 1505-1516.

Pichon, L., Hampe, A., Giffon, T., Varn, G., Legall, J. Y. & David, V. (1996). A new non-HLA multigene family associated with the PERB11 family within the MHC class I region. *Immunogenetics* **44** , 259-267.

Pichon, L., Venditti, C., Harris, J., Elshof Amy, T., Pinelli, M. & Chorney, M. (1997). Studies of the 6.7 family of dispersed genomic fragments within the MHC class I region. *Exp. Clin. Immunogen.. Oct.* **14** (2), 131-140.

Pierce, S. K. (1994). Molecular chaperones in the processing and presentation of antigen to helper T cells. *Experientia* **50** (11-12), 1026-30.

Plevy, S. E., Targan, S. R., Yang, H., Fernandez, D., Rotter, J. I. & Toyoda, H. (1996). Tumor necrosis factor microsatellites define a Crohn's disease-associated haplotype on chromosome 6. *Gastroenterology* **110** (4), 1053-60.

Poljak, R. J., Amzel, L. M., Avey, H. P., Chen, B. L., Phizackerley, R. P. & Saul, F. (1973). Three-dimensional structure of the Fab' fragment of a human immunoglobulin at 2,8-A resolution. *Proc. Natl. Acad. Sci. USA* **70** (12), 3305-10.

Price, P., Witt, C., Allcock, R., Sayer, D., Garlepp, M., Kok, C. C., French, M., Mallal, S. & Christiansen, F. (1999). The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunol. Rev.* **167** (February), 257-274.

Pshezhetsky, A. V., Richard, C., Michaud, L., Igdoura, S., Wang, S., Elsliger, M. A., Qu, J., Leclerc, D., Gravel, R., Dallaire, L. & Potier, M. (1997). Cloning, expression and chromosomal mapping of human lysosomal sialidase and characterization of mutations in sialidosis. *Nat. Genet.* **15** (3), 316-20.

Pu, W. T. & Struhl, K. (1991). The leucine zipper symmetrically positions the adjacent basic regions for specific DNA binding. *Proc. Natl. Acad. Sci. USA* **88** (16), 6901-5.

Quentin, Y. (1992). Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements. *Nucleic Acids Res.*. **20** (13), 3397-401.

Reinsmoen, N. L., Friend, P. S., miller, W. V., Burgdorf, A., Giblett, E. R. & Yunnis, E. J. (1977). Inheritance of recombinant HLA-GLO haplotype suggesting the gene sequence. *Nature* **267** , 276-278.

Ribas, G., Neville, M., Wixon, J. L., Cheng, J. & Campbell, R. D. (1999). Genes encoding three new members of the leukocyte antigen 6 superfamily and a novel member of the Ig superfamily, together with genes encoding the regulatory nuclear chloride ion channel protein (hRNCC) and a N-N-dimethylarginine Dimethylaminohydrolase homologue, are found in a 30-kb segment of the MHC class III region. *J. Immunology* **163** , 278-287.

Riordan, J. R., Rommens, J. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J. L. & et al. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA [published erratum appears in Science 1989 Sep 29;245(4925):1437]. *Science* **245** (4922), 1066-73.

Rock, K. L., Reiser, H., Bamezai, A., McGrew, J. & Benacerraf, B. (1989). The LY-6 locus: a multigene family encoding phosphatidylinositol-anchored membrane proteins concerned with T-cell activation. *Immunol. Rev.* **111** , 195-224.

Russnak, R. & Ganem, D. (1990). Sequences 5' to the polyadenylation signal mediate differential poly(A) site use in hepatitis B viruses. *Genes Dev.* **4** (5), 764-76.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74** (12), 5463-7.

Sargan, D. R., Gregory, S. P. & Butterworth, P. H. (1982). A possible novel interaction between the 3'-end of 18 S ribosomal RNA and the 5'-leader sequence of many eukaryotic messenger RNAs. *FEBS Lett.* **147** (2), 133-6.

Sargent, C. A., Anderson, M. J., Hsieh, S. L., Kendall, E., Gomez-Escobar, N. & Campbell, R. D. (1994). Characterisation of the novel gene G11 lying adjacent to the complement C4A gene in the human major histocompatibility complex. *Hum. Mol. Genet.* **3** (3), 481-488.

Sargent, C. A., Dunham, I. & Campbell, R. D. (1989). Identification of multiple HTF-island associated genes in the human major histocompatibility complex class III region. *EMBO J.* **8** (8), 2305-2312.

Scaife, R., Gout, I., Waterfield, M. D. & Margolis, R. L. (1994). Growth factor-induced binding of dynamin to signal transduction proteins involves sorting to distinct and separate proline-rich dynamin sequences. *EMBO J.* **13** (11), 2574-82.

Schaffer, F. M., Palermos, J., Zhu, Z. B., Barger, B. O., Cooper, M. D. & Volanakis, J. E. (1989). Individuals with IgA deficiency and common variable immunodeficiency share polymorphisms of major histocompatibility complex class III genes. *Proc. Natl. Acad. Sci. USA* **86** (20), 8015-9.

Schroeder, H., W, Jr., Zhu, Z. B., March, R. E., Campbell, R. D., Berney, S. M., Nedospasov, S. A., Turetskaya, R. L., Atkinson, T. P., Go, R. C., Cooper, M. D. & Volanakis, J. E. (1998). Susceptibility locus for IgA deficiency and common variable immunodeficiency in the HLA-DR3, -B8, -A1 haplotypes. *Mol. Med.* **4** (2), 72-86.

Schuchmann, M., Hess, S., Bufler, P., Brakebusch, C., Wallach, D., Porter, A., Riethmueller, G. & Engelmann, H. (1995). Functional discrepancies between tumor necrosis factor and lymphotoxin alpha explained by trimer stability and distinct receptor interaction. *Euro. J. Immunol.* **25** (8), 2183-2189.

Sharp, P. A. (1981). Speculations on RNA splicing. *Cell* **23** (3), 643-6.

Sheets, M. D., Ogg, S. C. & Wickens, M. P. (1990). Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.*. **18** (19), 5799-805.

Shevach, E. M. & Korty, P. E. (1989). Ly-6: a multigene family in search of a function. *Immunol. Today* **10** (6), 195-200.

Shiina, T., Tamiya, G., Oka, A., Takishima, N. & Inoko, H. (1999). Genome sequence analysis of the 1.8 Mb entire human MHC class I region. *Immunol. Rev.* **167** (February), 193-199.

Shiina, T., Tamiya, G., Oka, A., Yamagata, T., Yamagata, N., Kikkawa, E., Goto, K., Mizuki, N., Watanabe, K., Fukuzumi, Y., Taguchi, S., Sugawara, C., Ono, A., Chen, L., Yamazaki, M., Tashiro, H., Ando, A., Ikemura, T., Kimura, M. & Inoko, H. (1998). Nucleotide sequencing analysis of the 146-kilobase segment around the IkBL and MICA genes at the centromeric end of the HLA class I region. *Genomics* **47** (3), 372-82.

Singal, D. P., Li, J., Ye, M. & Lei, K. (1998). D6S273 microsatellite polymorphism and susceptibility to rheumatoid arthritis. *Tissue Antigens* **52** (4), 353-8.

Singer, S. J. & Doolittle, R. F. (1966). Antibody active sites and immunoglobulin molecules. *Science* **153** (731), 13-25.

Slagel, V., Flemington, E., Traina, D. V., Bradshaw, H. & Deininger, P. (1987). Clustering and subfamily relationships of the Alu family in the human genome. *Molecular Biology And Evolution* **4** (1), 19-29.

Smit, A. F. (1996). The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6** (6), 743-8.
Smit, A. F. & Riggs, A. D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.*. **23** (1), 98-102.

Smit, A. F. & Riggs, A. D. (1996). Tiggers and DNA transposon fossils in the human genome. *Proc. Natl. Acad. Sci. USA* **93** (4), 1443-8.

Smith, D. K. & Xue, H. (1997). Sequence profiles of immunoglobulin and immunoglobulin-like domains. *J. Mol. Biol.* **274** (4), 530-545.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98** (3), 503-17.

Soyombo, A. A., Yi, W. & Hofmann, S. L. (1999). Structure of the human palmitoyl-protein thioesterase-2 gene (PPT2) in the major histocompatibility complex on chromosome 6p21.3. *Genomics* **56** (2), 208-16.

Speek, M., Barry, F. & Miller, W. L. (1996). Alternate promoters and alternate splicing of human tenascin-X, a gene with 5' and 3' ends buried in other genes. *Hum Mol Genet* **5** (11), 1749-58.

Spies, T., Blanck, G., Bresnahan, M., Sands, J. & Strominger, J. L. (1989a). A new cluster of genes within the human major histocompatibility complex. *Science* **243** (4888), 214-7.

Spies, T., Bresnahan, M. & Strominger, J. L. (1989b). Human major histocompatibility complex contains a minimum of 19 genes between the complement cluster and HLA-B. *Proc. Natl. Acad. Sci. USA* **86** (22), 8955-8.

Srivastava, P. K., Udono, H., Blachere, N. E. & Li, Z. (1994). Heat shock proteins transfer peptides during antigen processing and CTL priming. *Immunogenetics* **39** , 93-98.

Steinle, A., Groh, V. & Spies, T. (1998). Diversification, expression, and gamma delta T cell recognition of evolutionarily distant members of the MIC family of major histocompatibility complex class I-related molecules. *Proc. Natl. Acad. Sci. USA* **95** (21), 12510-5.

Stephen E. L, M. J. D., and E S. Lander. (1991). Primer, 0.5. MIT Center for Genome Research and Whitehead Institute for Biomedical Research, Nine Cambridge Center

Suankratay, C., Zhang, X. H., Zhang, Y., Lint, T. F. & Gewurz, H. (1998). Requirement for the alternative pathway as well as C4 and C2 in complement-dependent hemolysis via the lectin pathway. *J. Immunol.* **160** (6), 3006-13.

Sugawara, A., Uruno, A., Nagata, T., Taketo, M. M., Takeuchi, K. & Ito, S. (1998). Characterization of mouse retinoid X receptor (RXR)-beta gene promoter: negative regulation by tumor necrosis factor (TNF)-alpha. *Endocrinology* **139** (6), 3030-3.

Suh, W., Doyle, K. M. C., Fruh, F. K., Wang, K., Peterson, P. & Williams, D. B. (1994). Interaction of MHC class I molecules with the transporter associated with antigen processing. *Science* **264** , 1322-1326.

Superkova, L., Supek, F. & Nelson, N. (1995). The Saccharomyces cerevisiae VMA10 is an intron-containing gene encoding a novel 13-kDa subunit of vacuolar H+-ATPase. *J. Biol. Chem.* **270** (23), 13726-13732.

Swallow, C. J., Grinstein, S., Sudsbury, R. A. & Rotstein, O. D. (1990a). Modulation of the macrophage respiratory burst by an acidic environment: the critical role of cytoplasmic pH regulation by proton extrusion pumps. *Surgery* **108** (2), 363-8.

Tee, M. K., Thomson, A. A., Bristow, J. & Miller, W. (1995). Sequences promoting the transcription of the human XA gene overlapping P450c21A correctly predict the presence of a novel, adrenal-specific, truncated form of Tenascin-X. *Genomics* **28** , 171-178.

Thanos, D. & Maniatis, T. (1995). NF-kappa B: a lesson in family values. *Cell* **80** (4), 529-32.

The MHC sequencing consortium (1999). Complete sequence and gene map of the human major histocompatibility complex. *Nature* **401** , 921-923.

Tisch, R. & McDevitt, H. (1996). Insulin-Dependent Diabetes Mellitus. *Cell* **85** (May 3), 291-297.

Tiwari, J. L. & Terasaki, P. I. (1985). HLA and Disease Association, Springer-Verlag.

Trachtulec, Z., Hamvas, R. M., Forejt, J., Lehrach, H. R., Vincek, V. & Klein, J. (1997). Linkage of TATA-binding protein and proteasome subunit C5 genes in mice and humans reveals synteny conserved between mammals and invertebrates. *Genomics* **44** (1), 1-7.

Tripodis, N., Mason, R., Humphray, S. J., Davies, A. F., Herberg, J. A., Trowsdale, J., Nizetic, D., Senger, G. & Ragoussis, J. (1998). Physical map of human 6p21.2-6p21.3: region flanking the centromeric end of the major histocompatibility complex. *Genome Res.* **8** (6), 631-43.

Trowsdale, J. (1994). 'Both man & bird & beast': comparative organisation of MHC genes. *Immunogenetics* **41** , 1-17.

Tsuge, I., Shen, F. W., Steinmetz, M. & Boyse, E. A. (1987). A gene in the H-2S:H-2D interval of the major histocompatibility complex which is transcribed in B cells and macrophages. *Immunogenetics* **26** (6), 378-80.

Tuckwell, D. S., Brass, A. & Humphries, M. J. (1992). Homology modelling of integrin EF-hands. Evidence for widespread use of a conserved cation-binding site. *Biochem. J.* **285** (Pt 1), 325-31.

Uberbacher, E. C. & Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA.* **88** (24), 11261-11265.

Utans, U., Arceci, R. J., Yamashita, Y. & Russell, M. E. (1995). Cloning and characterization of Allograft Inflammatory Factor-1: A novel macrophage factor identified in rat cardiac allografts with chronic rejection. *J. Clin. Invest.* **95** (6), 2954-2962.

Utans, U., Quist, W. C., McManus, B. M., Wilson, J. E., Arceci, R. J., Wallace, A. F. & Russell, M. E. (1996). Allograft Inflammatory Factor-1: A cytokine-responsive macrophage molecule expressed in transplanted human hearts. *Transplantation* **61** (9), 1387-1392.

van-Ham, S. M., Tjin, E. P. M., Lillemeier, B. F., Gruneberg, U., van Meijgaarden, K. E., Pastoors, L., Verwoerd, D., Tulp, A., Canas, B., Rahman, D., Ottenhoff, T. H., Pappin, D. J., Trowsdale, J. & Neefjes, J. (1997). HLA-DO is a negative modulator of HLA-DM-mediated MHC class II peptide loading. *Curr. Biol.* **7** (12), 950-7.

Venditti, C. P., Harris, J. M., Geraghty, D. E. & Chorney, M. J. (1994). Mapping and characterization of non-HLA multigene assemblages in the human MHC class I region. *Genomics* **22** (2), 257-66.

Vernet, C., Ribouchon, M. T., Chimini, G., Jouanolle, A. M., Sidibe, I. & Pontarotti, P. (1993). A novel coding sequence belonging to a new multicopy gene family mapping within the human MHC class I region. *Immunogenetics* **38** (1), 47-53.

Vernet, C., Ribouchon, M. T., Chimini, G. & Pontarotti, P. (1994). Structure and evolution of a member of a new subfamily of putative GTP-binding proteins mapping to the human MHC class I region. *Mamm. Genome* **5** , 100-105.

Visser, C. J., Tilanus, M. G., Schaeffer, V., Tatari, Z., Tamouza, R., Janin, A. & Charron, D. (1998). Sequencing-based typing reveals six novel MHC class I chain-related gene B (MICB) alleles. *Tissue Antigens* **51** (6), 649-52.

Vogt, A. B., Moldenhauer, G., Hammerling, G. J. & Kropshofer, H. (1997). HLA-DM stabilizes empty HLA-DR molecules in a chaperone-like fashion. *Immunol. Lett.* **57** (1-3), 209-11.

Volanakis, J. E., Zhu, Z. B., Schaffer, F. M., Macon, K. J., Palermos, J., Barger, B. O., Go, R., Campbell, R. D., Schroeder, H., Jr. & Cooper, M. D. (1992). Major histocompatibility complex class III genes and susceptibility to immunoglobulin A deficiency and common variable immunodeficiency. *J. Clin. Invest.* **89** (6), 1914-22.

Volanakis, J. E., Zhu, Z. B., Schaffer, F. M., Macon, K. J., Palermos, J., Barger, B. O., Go, R., Campbell, R. D., Schroeder, H. W., Jr. & Cooper, M. D. (1992). Major histocompatibility complex class III genes and susceptibility to immunoglobulin A deficiency and common variable immunodeficiency. *J. Clin. Invest.* **89** (6), 1914-22.

Volz, A., Weiss, E., Trowsdale, J. & Ziegler, A. (1994). Presence of an expressed beta-tubulin gene (TUBB) in the HLA class I region may provide the genetic basis for HLA-linked microtubule dysfunction. *Hum. Genet.* **93** (1), 42-6.

Wahle, E. & Keller, W. (1992). The Biochemistry of 3'-end cleavage and polyadenylation of messenger RNA precursors. *Annu. Rev. Biochem.* **61** , 419-40.

Wahle, E. & Keller, W. (1997). Capping, methylation, and 3'-end formation of pre-mRNA. Eukaryotic mRNA Processing. Krainer, A. R., IRL Press, 280-309.

Walter, L. & Gunther, E. (1998). Identification of a novel highly conserved gene in the centromeric part of the major histocompatibility complex. *Genomics* **52** (3), 298-304.

Waterhouse, P., Khokha, R. & Denhardt, D. T. (1990). Modulation of translation by the 5' leader sequence of the mRNA encoding murine tissue inhibitor of metalloproteinases. *J. Biol. Chem.* **265** (10), 5585-5589.

Wattavidanage, J., Carter, R., Perera, K. L., Munasingha, A., Bandara, S., McGuinness, D., Wickramasinghe, A. R., Alles, H. K., Mendis, K. N. & Premawansa, S. (1999). TNFalpha*2 marks high risk of severe disease during Plasmodium falciparum malaria and other infections in Sri Lankans. *Clin. Exp. Immunol.* **115** (2), 350-5.

White, P. C., New, M. I. & Dupont, B. (1986). Structure of human steroid 21-hydroxylase genes. *Proc. Natl. Acad. Sci. USA* **83** (14), 5111-5.

Wickens, M. & Stephenson, P. (1984). Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation. *Science* **226** (4678), 1045-51.

Widner, W. R. & Wickner, R. B. (1993). Evidence that the SKI antiviral system of Saccharomyces cerevisiae acts by blocking expression of viral mRNA. *Mol. Cell Biol.* **13** (7), 4331-41.
Williams, A. F. & Barclay, A. N. (1988). The immunoglobulin superfamily - domains for cell surface recognition. *Annu. Rev. Immunol.* **6** , 381-405.

Wilson, A. G., Clay, F. E., Crane, A. M., Cork, M. J. & Duff, G. W. (1995). Comparative genetic association of human leukocyte antigen class II and tumor necrosis factor-alpha with dermatitis herpetiformis. *J. Invest. Dermatol.* **104** (5), 856-8.

Wilson, A. G., Symons, J. A., McDowell, T. L., McDevitt, H. O. & Duff, G. W. (1997). Effects of a polymorphism in the human tumor necrosis factor alpha promoter on transcriptional activation. *Proc. Natl. Acad. Sci. USA* **94** (7), 3195-9.

Wilusz, J., Pettine, S. M. & Shenk, T. (1989). Functional analysis of point mutations in the AAUAAA motif of the SV40 late polyadenylation signal. *Nucleic Acids Res.*. **17** (10), 3899-908.

Yang, X., Khosravi Far, R., Chang, H. Y. & Baltimore, D. (1997). Daxx, a novel Fas-binding protein that activates JNK and apoptosis. *Cell* **89** (7), 1067-76.

Yu, C. Y. (1991). The complete exon-intron structure of a human complement component C4A gene. DNA sequences, polymorphism, and linkage to the 21-hydroxylase gene. *J. Immunol.* **146** (3), 1057-66.

Yu, C. Y., Belt, K. T., Giles, C. M., Campbell, R. D. & Porter, R. R. (1986). Structural basis of the polymorphism of human complement components C4A and C4B: gene size, reactivity and antigenicity. *EMBO J.* **5** (11), 2873-81.

Yu, H., Chen, J. K., Feng, S., Dalgarno, D. C., Brauer, A. W. & Schreiber, S. L. (1994). Structural basis for the binding of proline-rich peptides to SH3 domains. *Cell* **76** (5), 933-45.

Zimmerman, P. A., Guderian, R. H. & Nutman, T. B. (1996). A new TNFA promoter allele identified in South American Blacks. *Immunogenetics* **44** (6), 485-6.