GENOMIC ORGANISATION AND TRANSCRIPT IDENTIFICATION IN THE CLASS II REGION OF THE HUMAN MAJOR HISTOCOMPATIBILITY COMPLEX

A thesis submitted for the degree of Doctor of Philosophy by

Amanda Jackson

Human Immunogenetics Laboratory
Imperial Cancer Research Fund
44 Lincoln's Inn Fields
London WC2A 3PX

and

Department of Genetics and Biometry
University College
Gower Street
London

May 1995
ABSTRACT

The aim of the work described in this thesis was to provide a complete picture of the class II region of the human major histocompatibility complex (MHC), identifying all the genes present, their genomic organisation and function.

Completion of the class II cosmid map was achieved by cloning two remaining regions: 1) 40 kb of the DNA between the DMB and LMP2 genes and 2) 24 kb of the DNA between the genes DPA1 and DNA.. This map facilitated the localisation of a retinoid receptor β gene, hRXRβ, 60 kb centromeric of DPB2 and eight cDNAs, one of which encoded a pseudogene with homology to a rabbit muscle phosphatase inhibitor-2 (IPP-2) gene. Exon amplification revealed two exons within 2.2 kb centromeric of DMB which could be alternate splice products of the DMB gene.

Three cDNAs composed of two exons and one intron, mapping between the DMB and LMP2 genes, were isolated from cDNA libraries. The three transcripts were not detectable on northern blots. Four unspliced transcripts contiguous with genomic DNA from within the class II region, were present in cDNA libraries. All seven transcripts probably result from aberrant mRNA transcription. Screening and partial sequencing of cosmids covering the class II region identified no other transcripts between DPB2 and DRB1.

Experiments were designed to investigate the evolutionary origins of the TAP and LMP genes. To this end, LMP-related genes on other chromosomes were investigated for linkage to TAP-related sequences. Three cDNAs were identified, mapping within 12 kb of the LMP7-related gene, MB1, on chromosome 14. No sequence homology between the cDNAs and TAP genes was observed. The MB1 gene consisted of three exons and two introns. This organisation is very different from LMP7 which is composed of six exons and five introns. Hence, it is likely that MB1 and LMP7 are separated by considerable evolutionary time.
ACKNOWLEDGEMENTS

I would like to thank my friends and colleagues who have helped to make my time at the ICRF such an enjoyable experience. Firstly, I gratefully acknowledge my supervisor, John Trowsdale, for his guidance and support. Next, I would like to thank members of the Human Immunogenetics Laboratory, both past and present: Gordon Allan, Mônica Belich, Ian Campbell, Will Foulkes, Richard Glynne, Ulrike Grünberg, Jethro Herberg, Adrian Kelly, Lesley-Anne Kerr, Pat Miller, Ian Mockridge, Beatrice Plougastel, Stephen Powis, Jiannis Ragoussis, Derek McCusker, Sheetal Shah, Frances Sanderson, Claire Thomas and Marieka van Ham. Particular thanks must go to Louise Hosking for her friendship no matter how bad my moods and to Philippe Sanséau who is, without a doubt, My Favourite Frenchman. I am also indebted to Stephan Beck, Fiona Francis, Gabriele Senger, Denise Sheer and my external supervisor Yvonne Edwards. I am particularly grateful to my parents, Roger and Edwina, and my sister Zoë for their continued support and encouragement. Many thanks to my friends and flatmates Jill Amos, Robin Dey, Wendy Douglass, Susan Kettle, Erinn LaFey, Robert Mettler, Ann Smith and Ashley Vaughan for many memorable evenings. Last, but by no means least, I would like to thank William Newell for his constant support and encouragement without which I would probably still be writing.
CONTENTS

Chapter 1
The Human Major Histocompatibility Complex
1.1. Discovery and localisation of the MHC 10
1.1.1. Class I molecules 10
1.1.2. Class II molecules 13
1.1.3. Class III molecules 14
1.2. Function of the MHC class I and class II loci 14
1.3. Structure of MHC class I and class II molecules 17
1.4. Antigen processing 23
1.4.1. The class I antigen processing pathway 23
1.4.2. Function of the transporters associated with antigen processing (TAPs) 25
1.4.3. Alternative class I processing pathway 29
1.4.4. Involvement of the proteasome in class I antigen processing 30
1.4.5. Heat shock proteins and antigen processing 31
1.4.6. Antigen processing by the class II pathway 33
1.4.7. Function of the invariant chain (Ii) 33
1.4.8. Involvement of DM in class II antigen processing 35
1.5. Polymorphism of the MHC 36
1.6. Molecular genetics of the MHC 38
1.6.1. Physical location of the MHC 38
1.6.2. Mapping the MHC region 40
1.6.3. The class II region 43
1.6.4. The class I region 47
1.6.5. The class III region 49
1.7. Evolution of the MHC 51
1.8. The MHC gene cluster 53
1.9. The MHC and disease associations 55
1.10. Approaches to finding novel genes 58

Chapter 2
Materials and Methods
2.1. Solutions 62
2.2. Screening recombinant DNA libraries 65
2.2.1. cDNA libraries 65
2.2.2. Cosmid libraries 66
2.2.3. PI genomic library 67
Chapter 3

Completion of an overlapping cosmid map of the human MHC class II region

3.1. Introduction
3.1.1. Cloning the DMB to LMP2 gap
3.2. Results and discussion
3.2.1. Estimating the size of the region uncloned in cosmids
between the genes DMB and LMP2

3.2.2. Construction of a cosmid library from the YAC 11.2
3.2.3. Flow-sorted chromosome 6 cosmid library
3.2.4. P1 genomic library
3.2.5. Isolation of cosmid clones from the DMB to LMP2 region
3.2.6. Isolation of a P1 clone
3.2.7. Cloning of the gap between DPAl and DNA
3.3. Conclusions

Chapter 4
Genomic organisation and transcript identification in the class II region KE3 to DMB

4.1. Introduction
4.1.1. Screening cDNA libraries with cosmid inserts to identify novel genes
4.1.2. Exon amplification of the MHC class II region
4.1.3. Sequencing the class II region
4.1.4. Localisation of a gene provisionally mapped to the MHC: the human retinoid X receptor β gene (hRXRβ)
4.2. Results and discussion
4.2.1. Exon amplification
4.2.2. Exon amplification of the cosmid O19A
4.2.3. Exon amplification of the cosmid HA14
4.2.4. Localisation of the hRXRβ gene
4.3. Conclusions

Chapter 5
Screening for transcripts in the human class II region DMB to DRB1

5.1. Introduction
5.2. Results and discussion
5.2.1. Mapping a phosphatase inhibitor pseudogene to the class II region
5.2.2. Phosphatase inhibitors
5.2.3. IPP-2 is a pseudogene
5.2.4. Isolation of four cDNAs from the class II region DMB to LMP2
5.2.5. Class I-related sequence in the class II region
5.2.6. Completion of cosmid screening of the class II region

-6-
Chapter 6
Cloning and characterisation of genes linked to the \textit{LMP}-related proteasome gene, \textit{MBI}

6.1. Introduction 158
6.1.1. The proteasome 158
6.1.2. Involvement of the proteasome in the class I antigen processing pathway 159
6.1.3. \textit{LMP2} and \textit{LMP7}: two proteasome subunits localised within the MHC class II region 160
6.1.4. Two proteasome components related to \textit{LMP2} and \textit{LMP7} 162
6.1.5. Evolution of the \textit{LMP/TAP} gene cluster: are the \textit{LMP}-related proteasome genes, \textit{MBI} and \textit{delta}, linked to \textit{TAP}-related genes? 163
6.2. Results and discussion 164
6.2.1. Isolation of cosmids encompassing the \textit{MBI} and \textit{delta} genes 164
6.2.2. Isolation of cDNA clones with MB4 cosmid insert 164
6.2.3. Partial nucleotide sequence of cDNA clones encoded within the cosmid MB4 165
6.2.4. Initial mapping of the cDNA clones M11, M14 and M15 166
6.2.5. Genomic sequencing of cosmid MB4 166
6.2.6. Genomic organisation of \textit{MBI} 172
6.3. Conclusions 174

Chapter 7
Concluding discussion

7.1. Genomic analysis of the human MHC class II region 178
7.2. Isolation of genes linked to the \textit{LMP}-related proteasome gene, \textit{MBI} 187

Chapter 8
References

Literature cited 190
# FIGURES

1.1. Molecular genetic map of the MHC  11
1.2. Schematic representation of the MHC class I and class II molecular structures  18
1.3. Diagrammatic representation of the three-dimensional crystal structure of HLA-A2  20
1.4. Diagrammatic representation of the three-dimensional structure of class II HLA-DR1 superimposed on class I HLA-A2  22
1.5. Schematic model of the class I antigen processing pathway  32
1.6. Schematic model of the class II antigen processing pathway  37
1.7. Position of the MHC on chromosome 6  41

3.1. Map of the class II region showing areas uncloned in cosmids  85
3.2. Map of the MHC class II region showing deletion mutants spanning the area  87
3.3. PFGE mapping of the class II region  90
3.4. Cosmid and P1 map of the class II region between the genes DMB and LMP2  95
3.5. Flow-sorted chromosome 6 cosmid library hybridisation  97
3.6. Southern blot analysis of cosmids A12 and A15  98
3.7. FISH mapping of cosmids A12 and A15  99
3.8. FISH mapping of P1 clone, DP1  101
3.9. Southern blot analysis of cosmids mapping to the region telomeric of DPA1  102
3.10. Cosmid map of the class II region between DPB1 and DNA  104
3.11. Complete YAC and cosmid map of the class II region  106

4.1. Map of the class II region from KE3 to DMB  107
4.2. Schematic representation of exon amplification  113
4.3. Nucleotide and amino acid sequence of exons HA14-2 and HA14-4  116
4.4. Southern blot analysis of exons HA14-2 and HA14-4  118
4.5. Determination of positions of sequences homologous to exons HA14-2 and HA14-4 in the mouse MHC  120
4.6. Zoo blot analysis of exon HA14-4  122
4.7. FISH mapping of cosmid H7  124
4.8. Map of the class II region showing localisation of the hRXRβ gene  125
5.1. Map of the class II region from DMB to DRA
5.2. Cosmid map of the class II region between DMB and LMP2
5.3. Southern blot analysis of cDNA P19
5.4. Nucleotide and amino acid sequence of human IPP-2 cDNA P19
5.5. Genomic Southern blot and Zooblot analysis of cDNA P19
5.6. Northern blot analysis of human IPP-2
5.7. Amino acid sequence comparison of human and rabbit IPP-2 cDNAs and human genomic IPP-2
5.8. FISH mapping of human IPP-2
5.9. Cosmid map of the class II region
5.10. Nucleotide and amino acid sequence of cDNA S1
5.11. Splicing patterns of cDNAs S1, S2, S3 and S4
5.12. Nucleotide and amino acid sequence of cDNA S4
5.13. Nucleotide and amino acid sequence of class I-homologous sequence from cosmid A15
5.14. Map of the class II region showing cosmid inserts screened onto cDNA libraries
5.15. Expressed transcripts of genomic DNA, mapping to the class II region, isolated from cDNA libraries
5.16. Genomic sequence analysis of the class II region

6.1. Southern blot analysis of cDNA clone M14
6.2. Southern blot analysis of cDNA clone M15
6.3. Southern blot analysis of MB1
6.4. Partial nucleotide and amino acid sequence of cDNA clone M14
6.5. Partial nucleotide and amino acid sequence of cDNA clone M15
6.6. Genomic organisation of MB1
CHAPTER 1: INTRODUCTION

1. The Human Major Histocompatibility Complex

The major histocompatibility complex (MHC) is one of the best characterised regions of the human genome. Situated on chromosome 6, region p21.3, it spans 3.8 Mbp of DNA and contains over 100 genes (figure 1.1). These include 2^\text{I} class II genes at the centromeric end of the MHC and 3^\text{I}- associated genes at the telomeric end. Between these two regions is a heterogeneous collection of 45 class III genes including those for complement components C2, C4 (C4A and C4B) and factor B together with three members of the Hsp70 gene family, and the cytokines tumour necrosis factor (TNF) \text{a} and \text{b}.

Several features of the MHC have stimulated the enormous amount of work on this gene cluster: it encodes molecules which play a central role in the regulation of the immune response. Class I and II genes are highly polymorphic and are associated with a large number of diseases, many of the autoimmune type. The MHC represents 1/750th of human genetic material and can be considered as a model for the molecular organisation of the human genome. This chapter will discuss the MHC in three sections: 1) general background including classical studies of the MHC; 2) genomic organisation of the MHC and 3) outline of research carried out during the course of this study.

1.1. Discovery and localisation of the MHC

1.1.1. Class I molecules

The MHC was originally discovered because of the role of its products in the rejection of transplants made between incompatible individuals. The first experiments identifying a phenotype controlled by the MHC were carried out by Little and Tyzzer in 1916. It was shown that the growth of tumours transplanted from one strain of inbred mouse to another were dependent on the strains selected as donor and host (Little and Tyzzer, 1916; Klein, 1986). It was concluded that the trait of susceptibility to a grafted tumour was genetically determined. In 1937 Gorer carried out a series of experiments to investigate the genetic basis of tumour transplantation. Two pure lines of mice (black and albino) were used to make
Figure 1.1. Map of the human major histocompatibility complex, located on chromosome 6p21.3 (from Campbell and Trowsdale, 1993).
various crosses. A transplantable tumour from an albino mouse was transferred to all animals and its growth or regression monitored. The albino mice possessed an antigen (called antigen II) in their erythrocytes, detectable with sera from immunised rabbits, which was lacking in the blacks. Haemagglutination experiments were carried out on the animals to test for the presence of antigen II. A correlation was found between erythrocyte agglutination by rabbit antiserum containing antibody II and tumour rejection, leading to the conclusion that "antigen II must be present in the tissues of the host, otherwise the tumour will regress" (Gorer, 1937). The study of blood groups in man and animals had shown that antigenic differences were determined by dominant genes and, in the light of Gorer's experiments, it was suggested that the genes determining susceptibility to tumour transplantation might be identical with those determining antigenic differences. Snell (1948) suggested that the genes concerned should be called histocompatibility genes, denoted by the symbol H. Since one of these genes was defined by the blood group antigen II, (Gorer, 1937) the locus for the genes was called H-2. Further genetic linkage and serological studies determined that the H-2 genotype carried two components, defined by the series of antigens H-2K and H-2D (Amos et al., 1955). These antigens also played a major role in determining the success of tumour transplantation.

The presence of histocompatibility antigens in humans was shown by a series of observations. First, multitransfused patients were found to raise antibodies to the transfused leucocytes of some donors but not others (Dausset, 1958). Second, women who had undergone multiple pregnancies contained leucocyte antigens in their blood which induced fetomaternal leucocyte incompatibility (Payne and Rolfs, 1958). This was demonstrated by the production of leukoagglutinins in maternal sera which were capable of agglutinating the leucocytes of some of their children. Family studies showed that these leucocyte antigens, detected by leukoagglutinins, were present in seven families for two generations and one family for three generations. This suggested that the leucocyte factors were genetically transmitted to the offspring either as dominants or recessives.

Initial attempts to classify leucocyte antigens into allelic systems used antisera obtained from multitransfused patients. These studies were unsuccessful, largely because the reagents were multispecific. Leucocyte agglutinins from women sensitised during pregnancy are less complex and were used in experiments to establish the first leucocyte antigen system, Group 4 (Van Rood
and Van Leeuwen, 1963). This system was based on a statistical analysis of the reactivity patterns of sera from multiparous women with large panels of donors and defined a single locus. The existence of a further leucocyte antigen system, defining a locus independent of Group 4, was described in 1964 using similar methods (Payne et al., 1964). These two loci were named HLA-A and HLA-B for human leucocyte antigen locus A and human leucocyte antigen locus B. It was soon demonstrated by family and population studies that these loci were closely linked (Ceppellini et al., 1967). Later, a third locus, HLA-C, was described which was found to be closely linked to both the HLA-A and HLA-B loci (Svejgaard et al., 1973). Survival of human tissue grafts was found to be determined by matching the HLA-A, -B and -C loci which suggested that the HLA antigens were analogous to the mouse histocompatibility (H-2) loci. These molecules were classified together as the classical transplantation or class I antigens.

1.1.2. Class II molecules

The development of a histocompatibility test in vitro promoted the discovery of a fourth human histocompatibility locus. When lymphocytes from unrelated individuals were cultured together in the same tube they underwent cell division and morphological transformation, a phenomenon called the mixed lymphocyte reaction (MLR: Bach and Hirschhorn, 1964). It was discovered that the MLR reaction occurred with leucocytes matched for products at the HLA-A, -B and -C loci (Yunis and Amos, 1971), leading to the proposal that the HLA and MLR loci were closely linked but genetically separable identities. Interpretation of data from a number of sources, including MLR and antibody typing led to the concept of the class II region, HLA-D. Further antibody typing and serological studies in HLA-D matched individuals led to the discovery that the HLA-D region could be subdivided into three loci known as the HLA-DP, -DQ and -DR loci (Tosi et al., 1978; Shaw et al., 1980). Combining these results led to the concept of the major histocompatibility complex: a gene cluster involved in the survival of tissue transplants.

The development of typing assays in the mouse provided evidence for the existence of the mouse MHC class II region. Synthetic polypeptides were injected into different inbred strains of mice and their ability to mount an immune response tested. This response was found to be controlled by 'immune response' (Ir) genes which mapped to a region of the mouse distinct from the H-2K and -D
loci. These two genes (*I-A* and *I-E*) were found to correspond to the human \*HLA-DQ* and \*DR* loci respectively, thus defining the mouse MHC class II region.

### 1.1.3. Class III molecules

In 1963 a mouse serum protein was described which mapped to the H-2 complex, midway between the \*H-2K* and \*H-2D* loci (Shreffler and Owen, 1963). This was later identified, immunochemically, as being the complement component C4 (Meo et al., 1975). Similarly, in humans, two genes coding for C4 were found to be closely linked to the \*HLA-B* locus (O'Neill et al., 1978). Analysis of a family with complement component C2 deficiency revealed evidence for close linkage between the C2 defect and the HLA loci (Fu et al., 1974). The factor B (Bf) locus was assigned to a position between \*HLA-B* and \*HLA-D* based on family recombinant studies (Lamm et al., 1976). Further studies showed that the C4, C2 and Bf loci were tightly linked and mapped to a region which later became known as the class III region (Weitkamp and Lamm, 1982).

### 1.2. Function of the MHC class I and class II loci

A series of key experiments contributing to an understanding of the function of class I and class II molecules was carried out by Zinkernagel and Doherty (1975; Doherty et al., 1976). Different strains of mice were infected intracerebrally with lymphocytic choriomeningitis virus (LCMV) and killed one week later. Their cytotoxic T lymphocytes (CTLs) were tested for their ability to lyse target cells (LCMV-infected fibroblasts) from mice with varying H-2 haplotypes. It was discovered that the H-2\*^ restricted target cells were lysed only by H-2\*^ compatible virus CTLs. All the mouse strains produced CTLs in response to the viral infection but in each case these lymphocytes were only able to lyse target cells of the same MHC haplotype. This led to the suggestion that viral and H-2 components were present in a common complex which was recognised by T cells.

Studies investigating helper T cell activation arrived at similar conclusions showing that T cell activation by antigens could only occur in an MHC-compatible environment (Sprent, 1978). Purified mouse helper T cells were injected into irradiated syntenic mice together with sheep erythrocytes to act as antigen. Activation of helper T cells by the sheep antigen only occurred if the
donor and host mouse shared H-2 determinants. It was later determined that CTLs recognise and kill infected target cells that share class I molecules of the MHC with the host in which the CTL developed (reviewed by Swain and Dutton, 1980). The products of the class II region were essential for accessory cells such as macrophages to present antigens to helper T cells (Heber-Katz et al., 1983). Thus, the concept of MHC-restricted antigen presentation was evolved in which cytotoxic T cells and helper T cells recognise foreign antigen only in association with self class I or class II molecules respectively.

At first it was not known whether the T cells recognised the antigen and MHC molecule separately with different receptors or together with the same receptor. Over subsequent years it became apparent that T cells had a single receptor which recognised a single complex of MHC molecule in the form of a peptide (Davis and Bjorkman, 1988). Initial experiments carried out on class II molecules showed that an immunogenic peptide derived from chicken ovalbumin (residues 323-329) showed high binding affinity for these molecules (Buss et al., 1986). This observation was complemented by Babbit and coworkers who showed that a 10 amino acid stretch of the hen-egg lysozyme protein could be presented in H-2K mice to T cells, creating the structural determinant recognised by helper T cells (Babbitt et al., 1985). Over subsequent years it became apparent that MHC class II molecules bound peptides of 12-25 amino acids in length (Chicz et al., 1992).

Similar binding studies were carried out on MHC class I molecules. It was demonstrated that epitopes of influenza A viral nucleoprotein, recognised by CTLs in association with MHC class I molecules, could be defined by short synthetic peptides derived from the nucleoprotein sequence (Townsend et al., 1986). MHC class I molecules are now known to bind peptides of 8-10 amino acids in length (Saper et al., 1991).

The interaction between T cells and antigen involves the CD4 and CD8 glycoproteins found on the surfaces of T cells. The majority of CD4+ T cells are of the helper phenotype, whereas most CD8+ cells are cytotoxic (Littman, 1987). In general, class I molecules present peptides derived from endogenous proteins to CD8+ cytotoxic T lymphocytes while class II molecules present peptides derived from exogenous proteins to CD4+ helper T lymphocytes. The interaction between MHC/antigen complex and the T cell receptor results in T cell activation. Stimulated helper T lymphocytes secrete lymphokines which promote antibody
production by B cells and assist in the activation of CTLs. Stimulated CTLs lyse the cell presenting the foreign antigen (Klein, 1986).

A second mechanism by which class I and class II molecules play an important role in determining the specificity of an individual's immune response is through the selection of the T cell repertoire during development. Classical studies on the control of the immune response in mice had indicated that the MHC was responsible for determining whether or not an immune response could be mounted against a given peptide (Benacerraf, 1981). It was shown that the ability to mount an immune response was directly correlated with the affinity of purified MHC molecules for that peptide (Babbitt et al., 1985; Buus et al., 1986). Thus MHC molecules dictate, by their ability to selectively bind peptides from a protein antigen, whether a T cell response can be generated against a given protein antigen. The entire family of antigen-specific receptors on lymphocytes is required to discriminate between antigens that are self and antigens that are foreign. Recognition of non-self antigens activates the immune response to eliminate the foreign molecules. The immune system learns to differentiate between self and non-self antigens during development. This is accomplished by providing each lymphocyte with a unique receptor then selectively deleting or inactivating the population of cells bearing anti-self receptors. This process, the acquisition of self-tolerance, occurs by positive and negative selection of T cells.

Selection of T cells appropriate for antigenic peptides in association with self-MHC, is developed in the thymus gland. Double positive CD4+CD8+ T cells, bearing receptors which recognise self-MHC molecules (on epithelial cells) are positively selected for differentiation into CD4+8- or CD4-8+ single positive cells. Thus, CD4+8+ cells only differentiate into CD4+ or CD8+ mature T cells if they come into contact with thymic epithelial cells of the MHC class II or class I haplotype that are recognised by their receptor. This is known as positive selection (Schwartz, 1989; Marrack and Kappler, 1988).

Negative selection also occurs in the thymus and is responsible for the deletion of T cells with T cell receptors (TCRs) specific for peptides derived from self-proteins bound to self-MHC molecules. These T cells would otherwise be autoreactive (Schwartz, 1989). This event occurs at the CD4+CD8+ stage of T cell development and involves participation of the CD4 or CD8 molecules in the recognition of MHC class I or II molecules, respectively. Evidence for this process came from experiments involving the 1-E molecule which reacts with TCRs bearing the Vβ17a epitope. Mice strains which could not express 1-E due to
a defect in the *Ea* genes possessed mature T cells utilising V\( \beta \)17a, whereas strains which expressed 1-E normally, selectively deleted their V\( \beta \)17a positive T cells (Schwartz, 1989). Negative selection is thought to be the major mechanism for establishing immunologic self tolerance. However, such clonal deletion does not remove all autoreactive T cells.

Many self-T cell epitopes, including all housekeeping gene products, ubiquitous cell surface molecules and constituents of lymphoid and epithelial cells are present within the thymus and can induce negative selection. However, autoreactive T cells not expressed in the thymus, such as tissue specific antigens, are not removed by clonal deletion. It has been postulated that self-antigens with highly specific tissue distribution are controlled in the periphery (reviewed in Nossal, 1994). Transfer of such self-antigens to local lymphatic tissue is thought to enable induction of anergy, a phenomenon whereby lymphocytes can be functionally silenced without being killed. The anti-self cells may receive a nonlethal down-regulatory signal and be induced into a state of anergy. Such anergic cells may remain susceptible to further signals on renewed contact with antigen and may lead to eventual deletion. As yet, this process has not been fully elucidated.

### 1.3. Structure of MHC class I and class II molecules

Class I and class II molecules were initially characterised in 1973 (Cresswell et al., 1973). They are cell surface glycoproteins composed of \( \alpha \beta \) polypeptide chain heterodimers (figure 1.2). Both molecules have four extracellular domains, each approximately 90 amino acids long and encoded on separate exons (Malissen et al., 1982). In class I molecules, three of these domains are contained within the 44kD \( \alpha \)-chain (\( \alpha_1 \), \( \alpha_2 \) and \( \alpha_3 \) domains) which is encoded within the MHC and is often referred to as the heavy chain. The fourth domain is composed of the 12kD, non-covalently bound light chain, \( \beta \)-2 microglobulin (\( \beta \)-2m). \( \beta \)-2m is encoded on human chromosome 15. The \( \alpha_3 \) domain and \( \beta \)-2m are relatively conserved and show amino acid sequence homology to immunoglobulin constant domains (Orr et al., 1979). The polymorphic \( \alpha_1 \) and \( \alpha_2 \) domains show no significant homology to immunoglobulin constant or variable domains, but have been reported to show weak sequence homology to each other (Orr at al., 1979). In the case of class II molecules, two domains are contained within the 34kD \( \alpha \)-chain and two within the 29kD \( \beta \)-chain, both chains being encoded by the MHC. The \( \alpha_1 \) and \( \beta_1 \)
MHC class I

α2

α1

α3

β2-m

TM

Cytoplasm

Expressed by virtually all nucleated cells

MHC class II

β1

α1

β2

α2

TM

Cytoplasm

Restricted expression e.g. macrophages, B-cells and activated T-cells

Figure 1.2. Schematic illustration of the structure of class I and class II molecules. TM, transmembrane region. Glycosylation sites are indicated by open circles. Intra-domain disulphide bridges are indicated by S-S.
domains are furthest away from the cell surface and are highly polymorphic whereas the α2 and β2 domains are membrane proximal and less polymorphic. Class I molecules are expressed by virtually all nucleated cells. Class II molecules are expressed primarily by B lymphocytes, macrophages, dendritic cells and activated T lymphocytes (Cresswell, 1987).

Recently, the crystal structures of both class I and class II molecules have been determined (HLA-A2: Bjorkman et al., 1987a/b; HLA-Aw68: Garrett et al., 1989; HLA-B27: Madden et al., 1991; HLA-DR1: Brown et al., 1993). HLA-A2 consists of two pairs of structurally similar domains (figure 1.3). The α3 and β2-m domains form the base of the molecule and are both β-sandwich structures composed of two anti-parallel β-pleated sheets. The α1 and α2 domains each consist of an anti-parallel β-pleated sheet spanned by a long α-helical region. A deep groove was observed on the top surface of the HLA-A2 molecule composed of eight strands of anti-parallel β-sheets as a floor and two α-helical regions as the sides. This was proposed to form the peptide binding groove of the class I molecule. Additional electron density was observed in the groove, suggesting that a peptide or mixture of peptides had copurified and cocrystallised with HLA-A2. A polyalanine 9-mer was later modelled into the groove (Saper et al., 1991) providing a model for peptide binding. It was suggested that the peptide in the cleft was a nonamer in a largely extended conformation, with side chains protruding in a roughly alternating pattern.

The crystal structure of HLA-Aw68 also showed uninterpreted electron density (Garrett et al., 1989), which seemed to be different to that seen in HLA-A2. Comparison of the three-dimensional structures of HLA-A2 and HLA-Aw68 demonstrated how different pockets within the class I cleft were produced by amino acid differences between the two molecules. This polymorphism resulted in limited local structural changes, the backbone structure generally remaining unperturbed. These structural changes provide a basis for the observed allelic specificity in peptide binding. Since the affinity of antigenic peptides for individual MHC class I molecules could depend on only a few of the peptide's side chains fitting into the binding groove pockets, a wide range of peptides could be accommodated by various combinations of filled and empty pockets.

Peptide binding specificity was further explored by the crystallisation of another class I molecule, HLA-B27 (Madden et al., 1991). The clear electron density in the binding site of HLA-B27 indicated that the bound peptides would be restricted at four positions (P2, P3, P7, P9) to a B27 sequence-specific structural
Figure 1.3. Diagrammatic representation of the three-dimensional crystal structure of HLA-A2. (a) Side on view showing the $\alpha_1$ and $\alpha_2$ domains and the antigen binding groove, located between them, on the top surface of the molecule. The $\alpha_1$ and $\alpha_2$ domains form a platform with a single eight-stranded $\beta$-pleated sheet covered by $\alpha$ helices. (b) The antigen binding groove viewed from above (from Bjorkman et al., 1987).
motif, since these side chains were bound in pockets in the bottom and sides of the binding groove. A TCR could interact directly with the accessible peptide side chains (P1, P4, P5, P6, P8).

Acid elution and sequencing of peptides bound to class I molecules revealed specific consensus sequences. For HLA-B27, the most restricted peptide position was P2 (arginine) followed by P1 and P9 (charged amino acids). These well conserved positions are termed "anchor" positions. Further constraints were found at positions P3 (hydrophobic) and P6 (nonpolar residues). The remaining positions were unrestricted. Other anchor residues occur at different positions dependent on the class I molecule (for example asparagine at position 5 in H-2D\textsuperscript{b} (Falk et al., 1991).

Determination of the class II HLA-DR\textsubscript{1} crystal structure revealed overall similarity with that of class I molecules (figure 1.4; Brown et al., 1993). The two \(\alpha\)-chain domains, \(\alpha\textsubscript{1}\) and \(\alpha\textsubscript{2}\), of DR\textsubscript{1} superimpose closely on the corresponding \(\alpha\textsubscript{1}\) and \(\beta\textsubscript{2-m}\) subunit of HLA class I. The \(\beta\)1 domain of DR\textsubscript{1} superimposes on the \(\alpha\textsubscript{2}\) domain of the HLA class I molecule, and the \(\beta\)2 domain less closely on the class I \(\alpha\)3 domain. The class II peptide binding groove has the same eight stranded anti-parallel \(\beta\)-sheet floor and \(\alpha\)-helical sides as that of the class I groove. However, there are detailed changes in the class II structure so that the peptide binding site is open at both ends. Electron density in the binding groove indicates that long peptides, at least 15 residues, are bound in an extended conformation, projecting out of both ends of the site. This is consistent with the observation that class II peptides vary in length from 12-24 amino acids (Chicz et al., 1992). This contrasts with peptides bound to class I molecules where octamers and nonamers bind with extended conformations, and the N- and C-termini are bound in conserved, peptide terminal binding sites. A dimer of the class II \(\alpha\beta\) heterodimer was discovered in the three HLA-DR\textsubscript{1} crystals studied. It was suggested that class II molecules dimerised during their recognition by helper T cells. This dimer of dimers could lead to an increased affinity for the CD4 receptor and to the crosslinking of T cell receptors to initiate cytoplasmic signalling pathways.

Recently, binding of a decamer peptide to the class I molecule HLA-A2 has been reported showing the carboxy terminal residue positioned outside the peptide binding site (Collins et al., 1994). Rearrangement of several protein side chains at the C-terminus created an opening through which the peptide was able to extend. The decamer peptide studied in these experiments (MLLSVPLLLG) is the N-
Figure 1.4. Diagrammatic representation of the three dimensional crystal structure of class II $HLA-DR1$ (blue) superimposed on class I $HLA-A2$ (red). (a) $DR1 \alpha \beta$ heterodimer and $HLA-A2$ molecule. (b) The antigen binding groove of $HLA-A2$ and $DR1$ viewed from above (from Brown et al. 1993).
a) RED = HLA A2
BLUE = HLA DR1

b)
terminal 10 amino acids of the 17 amino acid signal sequence of calreticulin. It is found bound to HLA-A2, in vivo, whereas the nonamer (MLLSVPLLL) is not. These observations suggest that peptides could extend out of the class I binding site. It will be interesting to see if the novel surfaces presented by peptides extending out of the ends of class I molecules are recognised by T cells.

1.4. Antigen processing

Antigen processing can be defined as the structural modification and trafficking of protein antigens that enable the determinants recognised by T cells to interact with MHC molecules. These antigenic determinants are then transported to the cell surface for T cell recognition (Yewdell and Bennink, 1990). Two distinct pathways are used to process antigens: the endogenous and the exogenous pathways. Viruses, bacteria and protozoan parasites, such as those that cause malaria and sleeping sickness establish their infections inside the host cells, where antibodies cannot reach them. In infected cells, class I MHC molecules bind to and display peptides derived from the parasite. The complexes of parasite peptides and host class I molecules form antigens which can be recognised by cyotoxic T lymphocytes (Engelhard, 1994). In this way, T lymphocytes can identify and kill infected cells, selectively sparing healthy cells. Since class I-associated peptides are produced from proteins that originate from the cytoplasm, this pathway became known as the endogenous or class I pathway.

Specialised antigen processing cells (APCs), such as macrophages, roam the body ingesting extracellular materials, degrading them to produce peptides and presenting the peptides as antigens complexed with MHC class II molecules. When helper T lymphocytes recognise a class II-peptide complex on these antigen presenting cells, they secrete lymphokines that promote differentiation of immune system cells. Class II molecules associate with peptides either derived from proteins in the extra-cellular medium or from proteins located on the outer membrane. This pathway of antigen recognition is known as the exogenous or class II pathway.

1.4.1. The class I antigen processing pathway

The data summarised above led to the suggestion that the proteins recognised by class I-restricted T lymphocytes were degraded in the cytoplasm, and the peptides
derived from them were subsequently transported to the cell surface in association with class I molecules of the MHC (Towsend and Bodmer, 1989). Association of class I molecules with peptide was proposed to occur in the endoplasmic reticulum (ER). Supportive evidence came from a number of sources. Class I antigen processing is blocked by brefeldin A which inhibits exocytosis by interfering with normal vesicular traffic between the ER and Golgi (Yewdell and Bennink, 1989). Similar effects were observed in antigen presenting cells infected with a recombinant vaccinia virus expressing the adenovirus E19 glycoprotein (Cox et al., 1990). E19 specifically binds class I molecules and retains them in the ER. The effects of brefeldin A and E19 on antigen presentation indicate that peptides enter the secretory pathway in a pre-Golgi compartment, presumably the ER.

Association of class I molecules with antigen in the ER is supported by studies using the mutant mouse cell line RMA-S (Townsend et al., 1989). These cells express only 5% of the wild type level of class I molecules on their cell surface, even though heavy chains and β2-m are synthesised in the normal amounts. This results from a deficiency in the assembly of class I heavy chains with β2-m, the chains remaining trapped in the ER. Incubation of these cells with high concentrations of the appropriate peptides enhances class I assembly and the cells express more normal levels of class I molecules on the cell surface. Thus, peptides stabilise the assembly of class I molecules, a process which occurs in the ER.

During assembly, class I molecules are associated with an 88kD molecular chaperone, calnexin, that is released on peptide binding (Jackson et al., 1994). The ability of calnexin to influence the transport of incompletely assembled class I molecules in Drosophila melanogaster cells was assessed. Stably transfected Drosophila cell lines were prepared that expressed murine Kb, Db, or Ld heavy chains with or without β2-m in the presence or absence of calnexin. Expression of calnexin was shown to slow intracellular transport of both peptide-deficient heavy chain-β2-m heterodimers and free heavy chains. Calnexin also impeded the rapid intracellular degradation of free heavy chains. Thus, calnexin might act to facilitate assembly of class I complexes by retaining or protecting assembly intermediates.
1.4.2. Function of the transporters associated with antigen processing (TAPs)

Since class I assembly occurs in the ER, there must be specific mechanisms by which peptides are generated in the cytoplasm and transported across the ER membrane. Insight into these mechanisms came from studies on cell lines with defective class I presentation including the human mutant lymphoblastoid cell line (LCL) 721.174 and the murine mutant RMA-S. These cell lines were derived in a similar manner by γ-irradiation (LCL 721.174) or ethylmethane sulfonate (EMS) mutagenesis (RMA-S) followed by several rounds of selection with anti-MHC antibody plus complement (DeMars et al., 1984; Ljunggren et al., 1989). The mutant cell lines show similar phenotypes. First, they retain most of their class I MHC molecules in the ER in an incompletely assembled state (Salter and Cresswell, 1986; Cerundolo et al., 1990). Treatment of these cell lines with γ-interferon does not result in increased expression of class I MHC molecules. Second, the mutants are unable to present intact protein antigens to class I-restricted CTLs but can present exogenously supplied peptides. This leads to increased cell surface expression of stable class I molecules on the cell surface suggesting that their class I molecules are fully functional (Hosken and Bevan, 1990). Third, the genetic defect in these cell lines is not due to defects in either the class I heavy chains or β2-m (Salter et al., 1985; Ohlen et al., 1990). These observations have led to the hypothesis that these cell lines are defective in the ability to transport antigenic peptides into the lumen of the ER.

A similar phenotype to those defined by the LCL 721.174 and RMA-S mutants, called the cim (class I modification) phenotype, was observed in the rat (Livingstone et al., 1989; Livingstone et al., 1991). Natural alleles at the cim locus modify the specificity of class I antigen presentation and the time of residence of newly synthesised class I molecules in the ER. It was therefore suggested that the cim locus product was also involved in peptide loading or assembly of class I antigens.

The genetic location of the defects in LCL 721.174 was mapped to the class II region of the human MHC on chromosome 6 (Cerundolo et al., 1990). LCL 721.174 has a deletion spanning from the DPB2 gene in the class II region to the complement gene cluster in the class III region, thus mapping the gene responsible for the antigen presentation defect within this interval. Similarly, the genetic defect responsible for the RMA-S phenotype was mapped to the murine chromosome 17, which encodes the MHC (Hosken and Bevan, 1992). The cim
locus was mapped to the rat MHC between the human DP homologue, RT1.H, and the DQA homologue RT1.Bα (Deverson et al., 1990).

Two genes mapping to the region of the MHC known to encompass the gene(s) responsible for the antigen presentation defect were discovered simultaneously in the mouse, rat and human. These are now known as TAP1 and TAP2 for transporter associated with antigen processing. Initially they were known as HAM1/2, PSF1/2 or RING 4/11, and MTP1/2 in the mouse, human and rat respectively (Deverson et al., 1990; Monaco et al., 1990; Trowsdale et al., 1990; Powis et al., 1992; Spies et al., 1990). These genes showed protein sequence homology to the ATP-binding cassette (ABC) superfamily of transporters which includes the human and mouse multidrug-resistance proteins (MDR: Hyde et al., 1990). Typically, eukaryotic ABC transporters consist of two hydrophobic, membrane spanning domains and two ATP-binding domains. All known eukaryotic members of this family (MDR, cystic fibrosis gene, and the yeast STE-6 gene) contain all four domains within a single polypeptide. Thus, the products of the TAP1 and TAP2 genes were the first members of the family that contained only two domains. It therefore seemed likely that the TAP1 and TAP2 gene products functioned as a heterodimer (Monaco et al., 1990). The ABC transporter superfamily includes over 30 members, each specific for a different substrate. Transported substrates include sugars, inorganic ions, amino acids, peptides and proteins (McGrath and Varshavsky, 1989; Felmlee et al., 1985). TAP1 and TAP2 were strong candidates for a specialised transport system that passed the peptide products of proteolysis from the cytosol into the ER (Trowsdale et al., 1990).

Evidence for such a function came from a number of sources. Transfection of TAP1 cDNA into mutant LCL 721.134 cells restored normal levels of surface HLA-A2 and -B5 expression (Spies and DeMars, 1991). However, transfection of the same cDNA could not restore the defect in LCL 721.174 cells. Both 721.174 and 721.134 cell lines contain class I genes HLA-B5 and HLA-A2 of one MHC haplotype (the other haplotype having been deleted). However, surface HLA-B5 levels are not detectable in the mutants and surface HLA-A2 levels are only 50-60% of that of the parental line, as determined by fluorescence activated cell sorting (FACS) analysis. 721.174 has a further deletion over the class II region and 721.134 contains a point mutation in TAP1. Hence, transfection of TAP1 restores class I surface expression in 721.134 but not in 721.174 since both TAP1 and TAP2 genes are deleted in 721.174. This implies that both genes are required
to restore the antigen processing defect. Similarly, the defect in RMA-S could be
restored by transfection of either a rat or mouse TAP2 cDNA, leading to increased
class I assembly, surface expression and antigen presentation (Attaya et al., 1992;
Powis et al., 1991).

An antiserum raised against a peptide representing the C-terminal 15 amino
acids of the TAPI protein sequence immunoprecipitated both TAPI and TAP2
proteins from human cells. This suggested that the two TAP genes encoded
proteins that associated with each other, possibly as a heterodimeric complex
(Kelly et al., 1992). The same antiserum was used to precisely locate the TAPI
protein by high resolution immuno-electron microscopy on ultrathin frozen cell
sections. Immunogold particles were found specifically on the membranes of the
ER and cis-Golgi. Since the antibody was raised against a peptide from the C-
terminus of the TAPI ATP-binding domain and labelling occurred on the
cytosolic side of the membrane, it was suggested that the TAPI protein was
oriented in the ER and Golgi membranes with its ATP-binding domain in the
cytosol (Kleijmeer et al., 1992). Transfection of TAPI and TAP2 cDNAs in the
cell line T2, which is derived from 721.174 and lacks both transporter genes, is
both necessary and sufficient for the rescue of class I surface expression
(Momberg et al., 1992).

The development of a TAPI-/- knockout mouse has allowed investigation of
the involvement of the TAP complex in selecting the T cell repertoire (Van Kaer
et al., 1992). A deletion was introduced into the TAPI gene of embryonic stem
cells by gene targeting, and mice homozygous for this mutation were produced.
Cell lines from such mice lacked cell surface class I expression and were unable
to present endogenously synthesised proteins to CTLs. Exogenously added
peptide could, however, be presented. An analysis of the T cells from various
organs showed that CD4+CD8+ and CD4+ cells were present at wild type levels.
However, CD8+ T cells were not found in the blood, spleen or lymph nodes of the
mutant mice. This is consistent with a role for class I in the positive selection of
CD8+ cells in the thymus.

Reconstitution of cell surface class I expression using TAP genes in mutant
cell lines, immunoprecipitation and localisation experiments strongly suggest a
role for the TAP complex in class I antigen processing. Direct biochemical
evidence is now available to support this. However, several experiments have
produced results contrary to this hypothesis.

-27-
An assay was developed to study the mechanism of peptide transport across the ER membrane into microsomal vesicles (Koppelman et al., 1992). Peptides were synthesised containing N-linked glycosylation acceptor sequences, which serve as glycosylation substrates. Their transport into microsomes results in the depletion of the pool of dolichol high-mannose oligosaccharides in the lumen of microsomal vesicles. Glycosylation of a marker protein present in the microsomal vesicles would then be inhibited. This assay showed no effect of ATP depletion, nor inhibition of transport by an ATPase inhibitor, oligomycin.

*In vitro* translation experiments also provided evidence suggesting that peptide transport was ATP-independent (Bijlmakers et al., 1993). *In vitro* translation of mouse class I heavy chains and β2-m was performed in rabbit reticulocyte lysate containing dog pancreas microsomes and radio-iodinated peptide. Microsomes were recovered by centrifugation, and their contents analysed on SDS gels. Radio-iodinated peptides, found in association with microsomes, displayed reduced electrophoretic mobility when compared to peptides recovered from the microsomal supernatant. This shift was found to be due to the attachment of an endoglycosidase H-sensitive carbohydrate chain to the peptides. Since the enzyme responsible for carbohydrate attachment was found only in the microsome lumen and the process occurred in an isotonic buffer, it was concluded that peptide translocation was ATP-independent. Furthermore, ATP depletion did not inhibit glycosylation of radio-iodinated peptide.

In both of these experiments the microsome preparation would be of great importance in interpreting the results. If, for example, the microsomes were leaky for small peptides this would mask any ATP- and hence *TAP*-dependent transport. Additional evidence for *TAP*-dependent transport has recently been described.

Peptide translocation was assessed in the *TAP*-deficient cell line T2 and in T2 transfected with rat *TAPI* and *TAP2* (Neefjes et al., 1993). Permeabilised cells were incubated with an 11-mer peptide containing an N-linked glycosylation sequence. N-linked glycosylation of the peptide was used to measure peptide translocation. Glycosylated peptide could be recovered from T2/TAPI+2 cells but not from T2 cells. This glycosylation depended on ATP being present. Nonhydrolysable ATP analogues did not yield glycosylated peptide, suggesting that ATP hydrolysis is required for *TAPI*+2-dependent peptide translocation into the ER.

A further assay demonstrating *TAP*-dependent transport was carried out on *TAP*+/+ and *TAP*-/− mice (Shepherd et al., 1993). Microsomes prepared from
TAP+/+ and TAP-/- mice could both assemble exogenously added peptide. When the reaction temperature was lowered from 37°C to 23°C peptide translocation into TAP+/+ microsomes was 4 fold higher than in TAP-/- microsomes. This difference in rates was ATP-dependent.

These results seem to provide strong evidence for TAP-dependent transport of peptides across the ER membrane. However, conclusive evidence will only be provided by the purification of recombinant TAPs, their incorporation into a synthetic lipid bilayer, and reconstitution of peptide transport in vitro.

1.4.3. Alternative class I antigen processing pathway

Evidence for an alternative class I antigen processing pathway has also been demonstrated (Henderson et al., 1992). Normal levels of HLA-A2 class I molecules, complexed with peptides, were found to be expressed on the surface of mutant T2 cells which lack both the peptide transporters TAP1 and TAP2. Peptide sequencing revealed that the peptides were derived from the signal sequences of cellular proteins.

Signal sequences are features of the amino terminal ends of newly synthesised proteins that are heading to the cell surface or other internal cellular compartments (reviewed in Sabatini et al., 1982). When ribosomes synthesise such proteins, the signal sequences ensure that the ribosomes will attach to the ER before the protein is completed, directing the proteins to their final destination. As the protein extrudes into the ER, an enzyme clips the signal sequence from the proteins, providing a source of peptides that could associate with class I MHC molecules.

Signal sequences are now known to be presented on the surface of normal cells in conjunction with HLA-A2. Further work (Wei and Cresswell, 1992) has confirmed these results and shows that most signal-derived peptides bound to HLA-A2 are longer than the nine residue peptides found in HLA-A2 isolated from non-mutant cells. The majority of HLA-A2 molecules isolated from T2 cells are unstable which may be due to the longer length of bound peptides. Trimming of peptides may occur by specialised enzymes resident in the ER (Falk et al., 1990). Peptides of more than nine residues in length may enter the ER and be trimmed to the stable binding length by such enzymes. Trimming may occur in both the TAP-dependent and signal-dependent pathways but may not be efficient if peptides have entered by the signal-dependent route (Townsend, 1992). As yet no
conclusive evidence exists to support this hypothesis. It has been suggested that most peptides are processed by the established path. However, this new pathway could provide an alternative processing route for specific pathogens.

1.4.4. Involvement of the proteasome in class I antigen processing

A pathway for antigen processing was proposed in 1989 (Townsend and Bodmer, 1989) suggesting that antigenic proteins were degraded in the cytoplasm into peptides which were transported through the ER. These peptides then bound to MHC class I molecules and travelled to the cell surface for presentation to T cells. Biochemical evidence (described in section 1.4.2) has shown that the TAPI and TAP2 genes are responsible for the transport of peptides through the ER and recently strong evidence has been produced to support the hypothesis that the proteasome is responsible for peptide production in this pathway.

The discovery of the proteasome and its involvement in the class I antigen processing pathway will be discussed in detail in chapter 6 and only a brief outline of its function will be described in this chapter.

The eukaryotic proteasome is composed of 14-17 different polypeptide subunits and has a sedimentation coefficient of 20S. It exhibits multiple peptidase activities which cleave specifically after basic, hydrophobic and acidic amino acid residues (Wilk and Orlowski, 1983). cDNAs encoding all the proteasome subunits have been cloned and sequence comparisons show that they can be divided into two classes: α and β. The 20S proteasome forms the proteolytic core of a larger protein complex, the 26S proteasome, which is responsible for protein degradation in the ubiquitin pathway (Peters et al., 1993).

In 1982 a group of 16 proteins were precipitated from mouse cell lines, using anti-MHC alloantisera, which were called Low Molecular Weight Proteins, or LMPs (Monaco and McDevitt, 1982). The LMP and proteasome complexes were subsequently shown to be similar in that they share most, but not all, subunits. Two genes encoding shared subunits, LMP2 and LMP7, have been localised to the class II region of the MHC and are closely associated with the TAP peptide transporter (Glynne et al., 1991; Martinez and Monaco, 1991; Kelly et al., 1991b). Incorporation of LMP2 and LMP7 into the proteasome enhances its capacity to cleave oligopeptides after hydrophobic and basic residues, and suppresses cleavage after acidic residues (Gaczynska et al., 1993). This favours the generation of the kinds of oligopeptides that preferentially bind to class I
molecules. The proteasome is now thought to be responsible for the generation of most peptides presented on MHC class I molecules, confirming its major role in the class I antigen processing pathway (Rock et al., 1994).

1.4.5. Heat shock proteins and antigen processing

Recently, it has been suggested that heat shock proteins (HSPs) have a role in antigen processing and presentation (Manara et al., 1993). Studies carried out on normal antigen presenting cells demonstrated that the determinant recognised by the anti-heatshock protein 72/73 monoclonal antibody (MAb) was constitutively expressed on the cell surface of monocytes as well as of B cells. The ability of monocytes to present an antigen to T cells was significantly decreased when preincubated with an anti-HSP 72/73 MAb. This provides evidence for a role of HSP72/73 in antigen processing. It has been suggested that HSPs constitute a relay line in which the peptides, after generation in the cytosol by proteasomes, are transferred from one HSP to another, until they are accepted by MHC class I molecules in the ER (Srivastava et al., 1994).

Some HSPs are known to bind to a diverse array of molecules, including peptides (Ang et al., 1991). It has been observed that immunisation of mice with the HSP gp96, isolated from a given virus-infected/transformed cell line, elicits an antigen-specific MHC class I-restricted CTL response (Blachere et al., 1993). This response is not elicited by gp96 itself but by virtue of the peptide that it chaperones. It has been proposed that peptides are transported into the lumen of the ER where they bind to gp96. Gp96 then transfers the peptides to MHC class I molecules for antigen presentation to T cells (Li and Srivastava, 1993). As yet the involvement of HSPs in the antigen presentation pathway has not been fully elucidated but by combining all the data discussed in section 1.4 it is possible to provide a model for class I restricted antigen processing (figure 1.5).

Antigenic proteins are degraded in the cytoplasm by the 20S proteasome. Peptide products then traverse the ER membrane via the TAP complex. At least one TAP-independent processing route has been described which achieves translocation of signal sequences across the ER membrane. Within the ER lumen the peptides may then be bound to the HSP gp96 which passes them to the assembled class I heavy chain/β2-m complex. This complex passes to the cell membrane for presentation to T cells.
Figure 1.5. (A). Schematic model of the class I antigen processing pathway (from Engelhard, 1994). Intracellular proteins, such as those from viral infections, are degraded into peptides by proteasomes. These peptides are then carried into the endoplasmic reticulum (ER) by the TAP transporter. At the same time, newly synthesised class I heavy chain and β2-m are deposited in the ER and assemble (with the help of accessory molecules such as p88 and heat shock proteins) to form a class I heavy chain/β2-m heterodimer that binds peptide. The resulting MHC-peptide complex is then transported through the Golgi to the cell surface. In a variation on this pathway (B), peptides are formed from signal sequences which are clipped from cellular proteins in the ER.
1.4.6. Antigen processing by the class II pathway

MHC class II molecules have evolved for the surface display of antigenic fragments derived from proteins in the endocytic pathway. In comparison with class I molecules, a distinct mechanism of peptide binding and interaction with a specialised non-MHC-encoded protein change the focus of class II peptide binding from the ER to the endosomes.

1.4.7. Function of the invariant chain (Ii)

MHC class II molecules assemble in the ER as a complex with a nonpolymorphic, non-MHC-encoded chain termed the invariant chain (Ii). The human Ii is a membrane protein that exposes 30 N-terminal residues on the cytoplasmic side of the membrane, spans the membrane between residues 30 and 60, and has a large C-terminal domain on the exoplasmic side (Claesson et al., 1983). Class II-expressing, Ii-negative fibroblasts show decreased assembly of class II αβ dimers, and reduced stability of these transported class II molecules (Anderson and Miller, 1992). Subsequent transfection of Ii into these cells results in recovery of normal class II conformation and efficient intracellular transport. This indicates that one role of Ii is to function as a class II-specific chaperone, facilitating proper class II folding during assembly in the ER.

Investigation of Ii binding to class II molecules was carried out using a soluble form of the Ii molecule (Teyton et al., 1990). Allelic forms of class II antigen were incubated with radiolabelled peptides corresponding to residues 307-319 of the influenza haemagglutinin peptide in the presence of various concentrations of Ii chain. The Ii chain was found to prevent peptides binding to the class II molecules. Further radiolabelling and gel chromatography experiments demonstrated that Ii chain was not bound to those class II molecules that contained peptide. These analyses suggest that the binding of Ii and antigenic peptides to class II molecules are mutually exclusive.

An explanation for these findings comes from considering that class II/peptide binding occurs in a post-Golgi, acidic, proteolytic intracellular compartment (discussed later). The open design of the class II binding site permits it to interact with segments of intact proteins. If class II molecules leave the ER with their binding sites exposed then they will have to avoid unproductive interactions with intact protein fragments in the ER. Avoidance of such binding is
thus mediated by the Ii (Germain, 1994). The Ii has been shown to possess a short
internal segment, the class II-associated invariant chain peptide (CLIP) region
(Riberdy et al., 1992). This region corresponds to exon 3 of Ii and either occludes
or closes the class II binding site so that peptide binding cannot occur. The
binding of Ii also decreases the potential for class II molecules to bind to free
peptides derived from cytosolic proteins that are the ligands for class I. Interaction
with Ii thus enables the formation of stable class II heterodimers capable of
transport out of the ER without the tight binding of antigenic peptides required by
class I molecules, as well as preserving a functional distinction between the two
pathways.

Transport of secreted proteins is thought to occur by default i.e. no specific
secretion signal is required. Accordingly, proteins that are resident within the
secretory pathway (ER, Golgi) should contain a retention signal, such as the
KDEL sequence which retains soluble proteins within the ER (Munro and
Pelham, 1987). Since Ii is retained at intracellular sites it was proposed to encode
such a signal. Simian CV1 cells, devoid of Ii and class II expression, were
transfected with varying length constructs of Ii and labelled with rabbit anti-Ii
antibody (Bakke and Dobberstein, 1990). A MAb was used to visualise
intracellular distribution of the Ii proteins. CV1 cells expressing wild-type Ii
showed little cell surface expression of the proteins, but intracellular localisation
in the perinuclear area and in vesicular structures more distinct from the nucleus.
However, strong cell surface staining was observed in cells transfected with an Ii
construct which had 15 amino acids deleted from the N-terminus. Two
conclusions were drawn from these data: 1) an intracellular sorting signal that
directs class II/Ii complexes to endosomes is present in the 12-15 N-terminal
amino acids of Ii, and 2) Ii chains are directed to endosomes implying that class II
association with peptide occurs in such vesicles.

Peptide binding to class II molecules occurs in vesicles called MHC class II
compartments (MIIC). These vesicles are thought to be lysosomal in origin as
they contain lamp-1 and CD63, both of which are typical lysosomal-associated
membrane proteins (Peters et al., 1991). However, other groups have reported the
presence of class II molecules in a compartment morphologically distinct from the
ER, trans-Golgi reticulum and lysosomes (Neefjes et al., 1990). Whatever the
exact definition of MIIC, it seems clear that binding of class II molecules and
antigenic peptide occurs in these compartments. No definitive data exist on the
sequence of events involved in exposing the class II binding site and producing
suitable peptides from intact protein antigens although a general scheme has been proposed (Germain, 1994).

During the last stage of transit to, or following entry into the MIIC, a combination of acidic pH and specific proteases cleaves the COOH-terminal portions of Ii. The CLIP-containing segment of Ii is removed, exposing the class II binding site. Some class II alleles have pockets that interact well with the released Ii segment and acquire a significant amount of the cleaved peptide (Hunt et al., 1992). The remaining class II binding regions are free to be occupied either by preformed peptides or exposed segments of endocytosed proteins undergoing proteolysis. The surface-expressed ligands would then be created by digestion of the antigen outside the region protected by the binding groove. The class II peptide complexes then pass to the cell surface for presentation to T cells by an as yet unknown mechanism.

1.4.8. Involvement of DM in class II antigen processing

Evidence for (a) novel gene(s) involved in the class II pathway came from studies on mutant cell lines with deletions in the MHC and concomitant defects in class II antigen processing. Three classes of mutants were isolated from B lymphoblastoid cell lines which are discussed in detail in chapter 3. These mutants allowed the mapping of the gene(s) affecting class II antigen processing to the region between the genes RINGS and DQBl in the MHC class II region (DeMars et al., 1985).

Recently, it has been discovered that the mutant phenotypes are due to defects in either of the linked DMA and DMB genes (Morris et al., 1994; Fling et al., 1994). DM genes were first isolated in 1991 (Kelly et al., 1991a) and were found to be members of the immunoglobulin gene family. The α and β chains encoded by these genes share ~30% identity with both class I and II molecules indicating that they must have diverged at around the same time that the class I and II sequences split from each other. This suggested that DMA and DMB were unconventional class II molecules, possibly with a unique function.

The sequence similarities between DM and the classical class II genes, their location adjacent to one another in the class II region and the identical phenotypes of DMA and DMB mutants suggested that the DMα and β chains associated to form a heterodimer. Direct evidence for this has come from immunoprecipitation experiments (Sanderson et al., 1994). Extracts from the Raji cell line were precipitated with an antiserum to DMA. Immunoprecipitated protein was eluted
and reprecipitated either with DMA or DMB antiserum. Reprecipitation with the antiserum to DMA revealed two DMA-specific bands and Reprecipitation with anti-DMB produced a DMB-specific band. Thus, DMA and DMB associate to form a heterodimer. Subcellular fractionation experiments and immunoelectron microscopy on ultrathin cryosections of Raji cells using anti-DMA and -DMB antibodies localised both DMA and DMB to typical MIICs together with the majority of intracellular HLA class II molecules (Sanderson et al., 1994).

Several possible models for the function of DM have been proposed (Sanderson et al., 1994). 1) DM might act as a CLIP sink, binding CLIP to leave conventional class II molecules free to bind antigenic peptides; 2) DM could act as a shuttle to deliver peptides into the MIICs for interaction with class II molecules; 3) DM could act as a chaperone, retaining other class II molecules in the MIICs until antigenic peptide is bound. Whatever the function of DM, these experiments clearly show that it has a vital role in the class II antigen processing pathway. Figure 1.6 shows a picture of the class II processing and presentation pathway as it is understood at present (from Engelhard, 1994).

1.5. Polymorphism of the MHC

The large number of alleles at each of the functional MHC class I and II loci make the HLA loci the most polymorphic loci known (Klein, 1986). As there are several functional MHC loci on a given chromosome and each locus has many alleles, at least $10^{12}$ allelic combinations are theoretically possible. Only a fraction of these possible combinations actually occurs in the human population but this fraction is large enough so that theoretically, virtually no two randomly chosen, unrelated individuals will have the same MHC alleles at all of their MHC loci. This situation contrasts strongly with that in other genetic systems, in which each locus has either one allele only or a low number of alleles at a locus, of which one is common and the others are rare (Klein et al., 1993).

Alleles at both the class I and II loci also show high nucleotide diversity. In other genetic systems, alleles at a given locus usually differ by a few nucleotide substitutions. In the MHC, some alleles differ by 100 or more substitutions. This polymorphism is found predominantly in the membrane-distal domains of the class II $\alpha$ and $\beta$ chains (Bell et al., 1987a) and in the two extracellular domains ($\alpha 1$ and $\alpha 2$) of the class I heavy chains (Parham et al., 1988). Polymerase chain
Figure 1.6. Schematic model of the class II antigen processing pathway (from Engelhard, 1994). MHC class II αβ-chains and the invariant chain (Ii) are assembled in the endoplasmic reticulum (ER). The class II complexes are transported through the Golgi to a late endocytic compartment with lysosomal characteristics (the MHC class II compartment: MIIC). Inside the MIIC the Ii is reduced to a smaller peptide, class II-associated invariant chain peptide: CLIP. The DM molecule is then thought to bind to CLIP, which frees the MHC molecules to bind to peptides derived from extracellular proteins. This class II-peptide complex then travels to the cell surface for recognition by helper T lymphocytes.
reaction (PCR)-based nucleic acid sequencing of class I and II alleles has provided definitive data on the nature and extent of MHC polymorphism (Bodmer et al., 1990). These data have revealed discrete regions of hypervariability in the membrane-distal α1 and α2 domains of class I molecules and in the membrane-distal α1 and β1 domains of class II molecules.

The three-dimensional structure of both class I (HLA-A2: Bjorkman et al., 1987a) and class II (HLA-DR3: Brown et al., 1993) molecules has been deduced, described in section 1.3. These structures have revealed that most of the polymorphic residues occur in the peptide binding cleft. Since polymorphism is prevalent in those loci encoding the functional antigen presenting molecules, and tends to be concentrated in the peptide binding groove, it is thought to have been mostly established by natural selection rather than random drift (Klein, 1986; Parham et al., 1988).

1.6. Molecular genetics of the MHC

1.6.1. Physical location of the MHC

The location of the HLA gene cluster was first assigned to chromosome 6 using human-Chinese hamster hybrid cells (Jongsma et al., 1973). HLA typing studies of an individual with a t(2;6)(p25;p21) translocation revealed that the breakpoint was between the HLA-B and -D loci, mapping the MHC to region p21 on chromosome 6 (Berger et al., 1979).

Chromosomes can be cytogenetically identified by their individual banding patterns that are revealed by a variety of techniques (Craig and Bickmore, 1994). This type of analysis shows that mammalian chromosomes contain domains called G bands, which are late replicating, AT-rich and highly condensed. These G bands are interspersed with domains having reciprocal properties (less condensed, early replicating and GC-rich). The most widely utilised method of chromosome banding (G banding) involves staining chromosomes with the dye Giemsa after proteolytic digestion. This reveals AT-rich regions as dark bands. A contrasting pattern (R banding) can be obtained either by Giemsa staining after heat denaturation in saline, or by staining with the dye chromomycin D. In this case the GC-rich regions are revealed as dark bands. T banding (R banding at elevated temperatures) identifies a subset of GC-rich R bands, half of which are located at chromosome telomeres.
Caesium chloride density gradient centrifugation enabled the distribution of the different types of DNA to be examined. Chromosome bands are thought to correspond to isochores (Bernardi et al., 1985). These are mosaics of long-range regions homogenous in their G+C content. Five types of isochores with different G+C% were discovered (H3, H2, H1, L2 and L1 in descending order of G+C%). 80% of the mapped human genes map to the GC-rich R bands with over half of these in the T bands (Craig and Bickmore, 1994). Codon usage, DNA replication timing and repeat sequence density are also related to chromosome bands and long range G+C% mosaic structures (Fukagawa et al., 1995). Gene-dense R bands with loose chromatin structures replicate early in S phase and are rich in Alu repeats, while G bands with condensed chromatin structures replicate late and are rich in LINE-1 repeats.

Analysis of sequences compiled in the GenBank database showed that the MHC locus is an example of long-range G+C% mosaic structures (Fukagawa et al., 1995). Class I sequences are GC-rich with the GC-richest isochore being H3 (53% GC). In contrast, class II sequences seem to be rather AT-rich and correspond to L and H1 isochores (40 and 45% GC, respectively). Class III sequences vary in their G+C content, although they are richer in GC than class II sequences. The telomeric portion of the class III region seems to belong to the GC-richest isochore H3, and the centromeric portion to the second GC-richest isochore H2 (49% GC). This difference in the G+C content of the class III region can be explained by the fact that the boundary of long-range G+C% mosaic domains, i.e. the transition between the AT-rich and GC-rich domains, contains the junction between classes II and III. Sequencing of cosmids over the class II/III boundary showed that the transition between the AT-rich and GC-rich domains occurred approximately 180 kb telomeric of DRA and has been named the "L/H transition."

The exact location of the human MHC on the short arm of chromosome 6, in relation to banding patterns, has been difficult to determine. Spring and coworkers γ-irradiated a lymphoma cell line to induce mutations of MHC encoded genes (Spring et al., 1985). Five days after γ-irradiation, loss mutants were selected using MAbs specific for HLA-A2 and HLA-Bw4. The mutants were cloned and typed with monospecific alloantisera showing that two groups of mutated cell lines had been produced. The first had lost the expression of class I antigens from one of the haplotypes, while the other had additionally lost different class II antigens from the same haplotype. The mutants in this second class were
either DQw1\(^+\)DQw2\(^+\), DQw1\(^-\)DQw2\(^+\) or DQw1\(^-\)DQw2\(^-\). Karyotyping of the wild-type and mutant lines in conjunction with HLA typing revealed deletions of various sizes on the short arm of one of the chromosome 6s, which could be correlated with loss of particular MHC gene products at the cell surface, thereby enabling mapping of certain HLA genes. These studies resulted in the mapping of the HLA-A and -B loci to the distal part of 6p21.33 and the class II genes to 6p21.31 on the short arm of chromosome 6.

*In situ* hybridisation of probes from the MHC class II region on prometaphase chromosomes, in parallel with G banding, placed the class II loci in sub-band 6p21.31, a dark G-band (Senger et al., 1993). These data suggest that the class III region is within band 6p21.32, a light G-band, and are consistent with the finding that the contiguous class I and III loci are composed of GC-rich isochores, characterised by G-light bands. Recombination analysis of the MHC loci revealed the gene order to be DP-DQ-DR-[C4,C2,Bf]-B-C-A (Weitkamp and Lamm, 1982). Mapping of the MHC relative to other markers on chromosome 6 revealed that the DP locus was closest to the centromere (figure 1.7).

1.6.2. Mapping the MHC region

The advent of cloning technology greatly advanced molecular genetic characterisation of the MHC. It became possible to isolate probes for the classically defined MHC genes which aided the sequencing, and analysis of gene organisation of individual loci. In addition, by hybridising these probes to total human genomic DNA at reduced stringency, it was revealed that there were many other related sequences in the human genome. For example, class I gene probes detected numerous genomic fragments which were found to correspond to seventeen class I genes, whereas only three class I loci had been identified by serological methods (Orr and DeMars, 1983). Similarly class II \(\alpha\)- and \(\beta\)-chain probes revealed class II genes in addition to those encoding the serologically defined DP, DQ and DR antigens (Long et al., 1983). These class I- and class II-related genes were isolated by screening cDNA libraries with probes for the classical MHC genes.

Initially, somatic cell hybrids containing fragments of chromosome 6 and \(\gamma\)-radiation induced mutants with deletions in chromosome 6 were used to show that these related genes mapped to the MHC region. These methods, however, did not allow fine mapping of the genes. An important advance in the understanding of
Figure 1.7. Position of the classical MHC loci relative to one another and to other markers on human chromosome 6 (from Weitkamp and Lamm, 1982).
the organisation of the region was facilitated by the development of the long-range physical mapping technique of pulsed field gel electrophoresis (PFGE: Schwartz and Cantor, 1984). This technique allows the resolution of DNA fragments < 2 Mb in size and the subsequent construction of long-range restriction maps. Such maps provide a direct measure of physical distance along a chromosome, and at the same time provide landmarks, in the form of sites for rare cutter restriction enzymes. As more markers are isolated across a region, their position on the map can be determined, providing information on the order of markers and the physical distance separating them.

PFGE was used to provide a long-range restriction site map of the human MHC (Dunham et al., 1987). High molecular weight DNA from a HLA-homozygous cell line was digested with a range of rare-cutting restriction enzymes. The resulting DNA fragments were separated by PFGE and transferred to nylon membranes. These filters were hybridised successively with a panel of cDNA and genomic probes specific for the class I, class II, and class III regions of the MHC. Analysis of the digest and hybridisation patterns produced by these methods allowed the construction of a physical map. This long-range restriction map established the orientation of the C2 complement gene and 21-hydroxylase B gene loci relative to the class I and class II loci, the C2 gene being on the telomeric side of the 21-hydroxylase B gene. The distance between the 21-hydroxylase B gene and the DRA locus was calculated to be 300-360 kb, while that between the C2 gene and the HLA-B locus was determined to be ~650 kb. The total length of MHC hybridising fragments was 3.8 Mb which provided the first accurate size of the MHC.

Once the region of interest, such as the MHC, has been mapped it can then be cloned using the previously mapped probes for the isolation of genomic DNA fragments. Yeast artificial chromosomes (YACs) are commonly used for cloning large fragments of contiguous DNA. The YAC cloning system is based on the introduction of exogenous linear DNA molecules into a yeast host where they are replicated as 'artificial' chromosomes. Cloning vectors supply the units that are essential for an autonomously replicating stable chromosome in yeast, along with appropriate yeast selectable markers (reviewed in Schlessinger, 1990). There are now several human YAC libraries with average insert sizes of 350-1000 kb, including one of the human MHC class II region where the average insert size is 620 kb (Ragoussis et al., 1991). The YAC library (10000 clones) was constructed using partially digested DNA from the cell line GM1416B. Clones
were isolated by hybridising cDNAs for MHC class II molecules to filters of the YAC library. These clones were mapped by restriction enzyme digestion and hybridisation studies.

Genomic libraries in cosmid vectors are useful in the analysis of smaller regions of the human genome since the average insert size is about 40 kb which makes the clones easier to manipulate. Most of the human class II region (Blanck and Strominger, 1988; Blanck and Strominger 1990; Hanson et al., 1989), all of the class III region (Spies et al., 1989a/b; Sargent et al., 1989a/b; Kendall et al., 1990) and part of the class I region (Shukla et al., 1991; S. Powis, personal communication) have been cloned in cosmids. All cosmid libraries were prepared in essentially the same manner. Genomic DNA, either isolated from lymphoblastoid cell lines or in the form of YAC clones, was partially digested with rare-cutting restriction enzymes to obtain fragments in the size range of 35-45 kb. These fragments were ligated into the appropriate cosmid vector and transformed into bacterial cells. The resulting colonies were fixed onto nylon membranes and screened with probes from the region of interest. Cosmid clones, isolated in this manner, were mapped by a combination of restriction enzyme digestion and hybridisation studies. Walking probes, isolated from cosmid end fragments, were used to rescreen the cosmid libraries to complete the maps. As cosmid clones are easy to manipulate they are ideal in searching for transcribed sequences. Methods used to obtain the current map of the genomic organisation of the human MHC will be described in detail with particular reference to the class II region.

1.6.3. The class II region

The class II genes are found in a region at the centromeric end of the MHC which, from PFGE mapping, spans approximately 1 Mb (Hardy et al., 1986; Dunham et al., 1989). The genes which encode the α and β chains of the class II antigens DP, DQ, and DR are localised in discrete genetic subregions. The DP subregion is the most centromeric class II subregion and is separated from the DQ loci by approximately 250 kb, which in turn is separated from the DRB genes by ~75 kb. There is a gap of ~100 kb between the DRB and DRA loci. In addition, the class II region contains class II pseudogenes and genes which are apparently intact but for which no protein product has yet been identified (DNA and DOB).
The first class II cDNA clones isolated were for the DRA gene. Polysomes from a B-lymphoblastoid cell line were selected with a monoclonal antibody against the DRα chain, and the mRNA from the antibody-bound polysomes was used as a template to make cDNA (Lee et al., 1982). The cloning of DRA facilitated the cloning of other class II α-chain genes by hybridisation of the DRA cDNA to genomic or cDNA libraries at low stringency (Spielman et al., 1984; Auffray et al., 1984). Similar approaches were taken to isolate DRB and DQB1 gene clones (Long et al., 1983). These were used to screen libraries at reduced stringency to obtain clones for the other class II β-chain genes. To date the class II region contains 6 α-chain genes and between 7 and 11 β-chain genes, the exact number depending on the haplotype.

Recently, two genes related to classical class I and class II genes and nine non-classical class II genes have been discovered in the class II region. A cosmid walk extending centromeric from the DPA1 locus led to the mapping of the COL11A2 gene which encodes the α-2 chain of type XI collagen (Hanson et al., 1989). This gene is located 45 kb centromeric of the HLA-DPB2 locus. Following this, a detailed molecular map of the class II region was established to determine the presence of HTF (Hpa II Tiny Fragment) islands (Hanson et al., 1991). HTF islands (also called CpG islands) are discussed in detail in section 1.8. Briefly, these islands are composed of tracts of non-methylated CpG-rich regions which colocalise with expressed sequences (Bird, 1987). The presence of HTF-islands is characterised by clusters of sites for methylation sensitive, rare-cutter restriction enzymes with CpG dinucleotides in their recognition sequences. DNA fragments encompassing such islands are usually a good starting point in the search for expressed transcripts. Four clusters of rare-cutter sites were identified in the class II region: one centromeric of the DP locus; two between DNA and DOB; and one between DQB1 and DQB3.

Three probes associated with HTF islands, two from the cluster centromeric of DP and one from the cluster between DNA and DOB, were used to screen cDNA libraries. cDNAs corresponding to four novel genes (RING1-RING4) were isolated (Hanson et al., 1991). RING1 and RING2 are located ~90 kb centromeric of DPB2, RING3 lies 110 kb telomeric of DPA1 and RING4 is located 25 kb centromeric of DOB. Further screening of cDNA libraries with cosmid fragments around the HTF islands led to the identification of additional new genes, RING5-RING7 and RING9-RING12 (Kelly et al., 1991a/b; Glynne et al., 1991; Trowsdale et al., 1990; Spies et al., 1990). Nucleotide sequencing of RING6 and RING7
showed that they were both related to class II histocompatibility antigens (Kelly et al., 1991a). *RING6* shows similarity to class II α chain genes whereas *RING7* is related to both class I and class II genes. Subsequent studies have shown that *RING6* and *RING7* (now renamed *DMA* and *DMB*) are involved in the class II antigen processing pathway (Morris et al., 1994; Fling et al., 1994; Sanderson et al., 1994; section 1.4.9).

*RING4* and *RING11*, now called *TAP1* and *TAP2*, encode highly related protein products that are members of the ATP-binding superfamily of transmembrane transporters (Trowsdale et al., 1990; Kelly et al., 1992). These genes are involved in the transport of antigenic peptides from the cytosol into the ER in the class I antigen processing pathway (discussed in section 1.4.2). The *RING10* and *RING12* genes, renamed *LMP7* and *LMP2*, show sequence homology with proteasome components (Glynne et al., 1991; Kelly et al., 1991b). *LMP2* and *LMP7* are now known to compose two subunits of the multicatalytic proteasome complex. Their expression alters the catalytic activities of the proteasome resulting in the generation of oligopeptides that preferentially bind to MHC class I molecules in the class I antigen processing pathway (Gaczynska et al., 1994).

The other genes in the class II region have no obvious association with the immune system. Sequence analysis of the entire *RING1* cDNA sequence showed weak identity with a number of sequences in the EMBL database, all of which were related to *RING1* solely on the basis of their high glycine content. However, sequence analysis using only the 139 N-terminal amino acids showed matches with a distinct cysteine-histidine motif in several proteins (Hanson et al., 1991). It has been suggested that some cysteine-histidine motifs are responsible for a zinc-dependent nucleic acid-binding function (Berg, 1990) and it is tempting to speculate that the *RING1* motif may also be involved in metal-dependent nucleic acid binding. However, the precise function of *RING1* remains as yet undetermined.

Probes from non-class II genes in the mouse *H-2K* region have been used to map human equivalents of the murine *KE3-KE5* genes to the human class II region. *KE4* corresponds to *RING5*, whilst *KE3* and *KE5* represent previously unidentified genes in the class II region centromeric of *COL11A2* (Hanson and Trowsdale, 1991). The murine homologue of *RING5* (*KE4*) encodes a putative transmembrane protein which is expressed at high levels in embryonic tissue, suggesting a regulatory role. An additional mouse gene, *KE6*, has recently been
isolated and found to map to the mouse $H-2K$ region proximal to $KE4$ (Aziz et al., 1993). Comparison of the complete $KE6$ sequence with the partial sequence of $RING2$ showed that $RING2$ was the human homologue of $KE6$. The $KE6$ gene gives rise to two transcripts: a 1 kb mRNA which is abundant in kidney and liver tissue and a 1.4 kb mRNA which is found in spleen tissue. Further sequence analysis of $KE6$ showed that the predicted protein product is highly homologous to several bacterial proteins which are members of a superfamily of short chain alcohol dehydrogenases. Members of this family either have oxidoreductase or dehydrogenase function and it was postulated that the $KE6$ protein would act in a similar manner. Since $KE6$ is highly expressed in kidney and liver, both tissues which are affected in polycystic kidney disease (PKD), the expression of the $KE6$ gene was investigated in rodent models of PKD using northern blot analysis. $KE6$ was found to be downregulated in two distinct murine models of heritable PKD. This disease is characterised by progressive enlargement of the kidneys due to numerous expanding cysts ultimately leading to renal failure. The observation that $KE6$ is downregulated in both kidney and liver tissues in mouse models of PKD suggested that $KE6$ was involved in the manifestation of this disease. As yet, the precise function of $KE6$ in the pathogenesis of PKD is unknown. By analogy, $RING2$ may be involved in human PKD but this hypothesis remains to be investigated.

The $RING3$ protein product is homologous to the Drosophila female sterile homeotic gene, $fsh$ (Hanson et al., 1991). The Drosophila $fsh$ locus is thought to be involved in development as $fsh$ expression is required maternally during oogenesis and then later in embryogenesis for normal embryonic pattern formation to occur (Haynes et al., 1989). Mammals, including humans, are known to have embryonically-expressed homologues of many Drosophila developmental genes and it is possible that a similar system of interacting gene products is involved in establishing the body pattern of mammals (Dressler and Gruss, 1988). This raises the possibility that $RING3$ is involved in human embryonic development.

Recently a novel human gene ($HSET$) has been localised to the class II region, mapping 40 kb centromeric of $KE3$ (Ando et al., 1994). Analysis of the nucleotide sequence of the $HSET$ cDNA revealed significant similarity to kinesin-related proteins in yeast, Drosophila and humans. Its predicted amino acid sequence contains a domain with strong sequence similarity to the ATP-binding and motor domains of a microtubule motor protein, kinesin. This protein is thought to be
involved in mitotic chromosome segregation, suggesting that the \textit{HSET} gene encodes a novel kinesin-related protein.

1.6.4. The class I region

The class I region spans approximately 1.8 Mb at the telomeric end of the MHC. The \textit{HLA-A}, \textit{-B} and \textit{-C} genes encode the \( \alpha \) chains of the class I cell surface glycoproteins and are involved in antigen presentation to cytotoxic T lymphocytes. One of the major difficulties in cloning the class I region was its large size and the presence of a large number of related sequences in the region. However, the use of PFGE mapping showed that \textit{HLA-C} is 130 kb distal to \textit{HLA-B} and just over 1000 kb proximal to \textit{HLA-A} (Dunham et al., 1987).

Recently, a number of class I-related genes have been isolated by screening cosmid and cDNA libraries with class I probes at low stringency. From sequencing studies it was shown that three of these were intact genes while the others were pseudogenes (Koller et al., 1989). The three intact genes, designated \textit{HLA-E} (Koller et al., 1988), \textit{HLA-F} (Geraghty et al., 1990) and \textit{HLA-G} (Kovats et al., 1990), encode proteins that are known to be expressed and give protein products. The function of these non-classical class I genes is poorly understood.

In the mouse, other highly divergent class I-related genes telomeric of the complex have been identified (Shawar et al., 1990). These genes appear to have a specialised role in presenting bacterial formyl-methionyl peptides to the immune system and have not yet been identified in other species. Two more genes \textit{HLA-H} and \textit{cdal2} have also been described (Shukla et al., 1991; Ragoussis et al., 1989). There is no evidence for expression of \textit{HLA-H} at the protein level and it is now known to encode a pseudogene. The \textit{cdal2} gene (now renamed \textit{HLA-J}) is located \(-50\) kb from \textit{HLA-A} and its function is unknown.

The myelin/oligodendrocyte glycoprotein gene (MOG) has been localised to the class I region mapping \(-80\) kb telomeric of \textit{HLA-F} (Pham-Dinh et al., 1993). The gene B30.2 is located \(-250\) kb centromeric of \textit{HLA-A} (El Kahiou et al., 1992). Ret finger protein (RFP) has been localised telomeric of the MHC by \textit{in situ} hybridisation (Amadou et al., 1995). It maps to chromosome band 6p21.3-p22 along with the butyrophilin (BT) gene which is telomeric of RFP (Amadou et al., 1995). MOG, B30.2, RFP and BT are all structurally related genes (Vernet et al., 1993a). The putative B30.2 and RFP peptides share similarities over their whole lengths, whereas the NH\(_2\)-terminal part of BT is homologous to the Ig-like
domain of MOG and B-G antigens of the chicken MHC. It is possible that these related genes have resulted from gene duplication and exon shuffling (Amadou et al., 1995).

The \( S \) gene is located 160 kb telomeric of \( HLA-C \) (Zhou and Chaplin, 1993). \textit{In situ} hybridisation showed that \( S \) gene expression was restricted to the differentiating keratinocytes in the granular layer of the epidermis. This suggests a role of the \( S \) gene in the development of differentiating keratinocytes. A combination of PFGE and YAC mapping along with the use of recombinant families localised the octamer transcription factor 3 \((OCT3)\) gene to the MHC class I region, mapping \( \sim 100 \) kb telomeric of \( HLA-C \) (Crouau-Roy et al., 1994). \( OCT3 \) is a transcription factor containing a homeodomain which may play a role in early development.

The \( P5-1 \) gene is a further example of a gene which is not structurally related to class I genes and is located within the class I region (Vernet et al., 1993b). The \( P5-1 \) locus maps between \( HLA-B \) and \( HLA-E \). Its pattern of transcription is restricted to lymphoid tissues which suggests that its protein product has an immunological function. Low-stringency hybridisation of \( P5-1 \) onto YAC clones covering the class I region showed that three copies (\( P5-2 \) to \( P5-4 \)) are located around the \( HLA-A \) gene and one copy (\( P5-5 \)) is located close to the \( HLA-G \) gene. This demonstrates the presence of a multiple-copy gene family, all members of which map within the MHC.

Recently, the putative GTP-binding protein gene \( HSR1 \) has been located close to \( HLA-E \) (Vernet et al., 1994). Several murine probes have been found to be conserved within the human MHC class I region (Amadou et al., 1995). The mouse \( Tctex-5 \) gene maps at a locus \( \sim 250 \) kb telomeric of \( HLA-E \) and the murine \( Tu42 \) gene maps telomeric of MOG. A gene (\( TUBB \)) thought to be involved in immotile cilia syndrome (ICS) has also been located to the class I region (Volz et al., 1994). This is a \( \beta \)-tubulin gene and maps to a segment 170-370 kb telomeric of \( HLA-C \).

A distinct family of five related sequences has been identified in the class I region. They are distantly related to classical class I chains (Bahram et al., 1994). These \( MIC \) (MHC class I chain-related) genes are conserved in primates and various other mammalian species. These results indicate that a primordial \( MIC \) gene originated from a class I gene before the evolution of the class I gene family. The \( MICA \) gene in this family is located 5 kb centromeric of \( HLA-B \) and is the most divergent mammalian class I gene known. Taken together, these results
define a second lineage of evolutionary-conserved MHC class I genes. This implies that MICA and possibly other family members have been selected for specialised functions that are either ancient or derived from those of typical MHC class I genes.

Recently, several reports have described cDNA clones mapping to the MHC class I region (Goei et al., 1994; Wei et al., 1993). These include three transcripts which map centromeric of HLA-B (Marshall et al., 1994; Venditti et al., 1994). As yet, the function of these transcripts is unknown.

1.6.5. The class III region

The class III region, which is bounded at the centromeric end by the class II region and at the telomeric end by the class I region, has been shown by PFGE mapping to span about 1.1 Mb (Dunham et al., 1987; Carroll et al., 1987). Molecular mapping of the class III region by PFGE combined with Southern blot hybridisation, established the organisation of the class III gene cluster with respect to the HLA-DR subregion and the HLA-B locus. The tumour necrosis factor genes lie ~250 kb centromeric of the HLA-B locus and ~390 kb telomeric of the C2 gene. The C2 gene is on the telomeric side of the steroid 21-hydroxylase B gene which lies about 300 kb telomeric of the DRA gene. In the last seven years over 900 kb of DNA from the class III region has been cloned in overlapping cosmids (Spies et al., 1989a/b; Sargent et al., 1989a/b; Kendall et al., 1990; Spies et al., 1990) and a detailed molecular map of the region has been established (Milner and Campbell, 1992). The remaining interval between the proximal end of the cosmid contig has been covered by YAC clones and, recently, by overlapping cosmids (Fukagawa et al., 1995).

Analysis of the cloned regions for coding sequences, using positional cloning, led to the discovery of at least 31 novel genes in the class III region (reviewed in Milner and Campbell, 1992; Sugaya et al., 1994). Three of the novel genes were shown by nucleotide sequencing to be members of the heat shock protein hsp70 multigene family (Sargent et al., 1989b; Milner and Campbell, 1990). Hsp70-1 and hsp70-2 are very closely related, encoding identical protein products. Hsp70-hom shares 90% identity with hsp70-1 but unlike hsp70-1 and -2, expression of this gene is not heat inducible.

Screening cDNA libraries with probes from the region between C4 and TNFα resulted in the identification of thirteen novel genes (Sargent et al., 1989b). One
of these genes corresponds to the human homologue of the murine B144 gene (Tsuge et al., 1987) and is located ~10 kb centromeric of TNFα. A second gene corresponds to the RD gene which encodes a predicted protein product of 42kD (Lévi-Strauss et al., 1988). The remaining genes were designated G1-G11.

Cosmid inserts from the region centromeric of the CYP21B gene were used to screen a cDNA library from a monocytic cell line resulting in the identification of the genes G12-G18 (Kendall et al., 1990). Further screening of this cDNA library led to the isolation of the additional novel genes G9a and G7a which are located centromeric and telomeric of the hsp70 gene cluster, respectively. The derived amino acid sequence of G7a showed 48% identity with the Saccharomyces cerevisiae valyl-tRNA synthetase and thus, it is likely that the G7a gene encodes the human valyl-tRNA synthetase (Hsieh and Campbell, 1991). An additional gene, designated G7b, has been discovered using similar methods. The G7b gene maps directly centromeric of G7a (Olavesen et al., 1993). The 'opposite strand gene', OSG, was identified by Morel et al. (1989). It overlaps the CYP21B gene and is transcribed in the opposite direction. This suggests a functional or regulatory relationship between CYP21B and OSG, but this has not been confirmed. BAT1 (~40 kb telomeric of TNFβ) and BAT5 (~100 kb telomeric of the hsp70 gene cluster) were identified by Spies et al. (1989a/b) by screening cDNA libraries with cosmid inserts.

Cosmid walking from the CYP21B gene in the class III region to the class II region has recently resulted in the isolation of three novel genes (Sugaya et al., 1994). The first gene, located approximately 150 kb centromeric of CYP21B, corresponds to the gene for the receptor of advanced glycosylation end products of proteins (RAGE) which is a member of the immunoglobulin superfamily. The second gene is located directly centromeric of RAGE and codes for the PBX2 homeobox gene, designated HOX12. PBX2 is a homeodomain-containing protein and is closely related to PBX1 which is involved in t(1;19) chromosomal translocations in acute pre-B-cell leukaemia. Recently, sequence analysis has revealed that G17 (discovered by Kendall et al., 1990) encodes HOX12 (Aguado and Campbell, 1995). The third gene, located directly centromeric of HOX12, is the human homologue of the mouse mammary tumour gene int-3. Sequence analysis of this gene showed nucleotide sequence identity with various Notch homologues in Drosophila, zebrafish, rat and mouse. Notch was first found as a Drosophila neurogenic gene required in development (Fortini and Artavanis-Tsakonas, 1993). Notch homologues have been isolated from a variety of
vertebrates, including two from humans. Due to its homology with Notch homologues, this MHC-located gene was termed NOTCH3.

Genomic sequence analysis of a 90 kb DNA segment containing the TNFα and β genes predicted the presence of three potential exons mapping between the BAT1 and TNFβ genes (Iris et al., 1993). One of the exons was proposed to encode one and a half ANK repeats that were suggested to belong to a potential new member of the NF kappa B family of proteins. A gene, named ikbl (for Ik B-like protein), encoding this exon has recently been isolated which maps ~12 kb telomeric of TNFβ (Albertella and Campbell, 1994). Two pairs of duplicated loci have been mapped to the class III, C4A/C4B region (Morel et al., 1989; Matsumoto et al., 1992; Gitelman et al., 1992; Bristow et al., 1993). The XA and TN-X genes are encoded on the opposite DNA strand to the steroid 21-hydroxylase B gene (Morel et al., 1989; Gitelman et al., 1992). The XA gene encodes a mRNA of unknown function, the expression of which is limited to the adrenal gland (Gitelman et al., 1992). The TN-X gene encodes a 400 kDa protein that resembles the extracellular matrix protein, tenascin (Matsumoto et al., 1992). The second pair of genes, YA and YB, are transcribed on the same DNA strand as the steroid 21-hydroxylase B gene and encode proteins of unknown function (Bristow et al., 1993).

1.7. Evolution of the MHC

The common ancestry of class I and class II genes is reflected in their related functions and structures. It is not clear when in evolutionary time MHC molecules arose. However, class I and II genes have been described in amphibians, reptiles and fish as well as mammals, which implies that class I and class II genes evolved in their present form before the radiation of vertebrates which took place about 400 million years ago (Kaufman et al., 1990). The ancestral class I and II genes have been duplicated many times to give the multiple loci seen in the mammalian MHC. The organisation of the human class II region suggests that the subregions were generated from a primordial α/β gene pair through a series of duplication and divergence events (Klein, 1986).

Comparisons of the nucleotide and amino acid sequences of the class II α-chain genes and their protein products revealed that DPA1, DQA1, DRA and DNA are equally diverged from one another, suggesting that they arose by duplication at roughly the same time (Auffray et al., 1984). A similar analysis of the β-chain
genes reveals that **DPB1**, **DQB1** and **DRB** are equally diverged from one another but that **DOB** is more distantly related. This suggests that the **DOB** gene may have arisen before the other β-chain genes in evolutionary time (Tonnelle et al., 1985). Sequence comparisons between genes within subregions reveal that additional duplications have probably occurred more recently.

As the sequences of the class II genes of mouse and man were deduced it became apparent that the subregions were homologous. Thus, the **DRA** and **Ea** genes are more closely related to one another than to the other loci, as are the **DRB** and **Eb** genes (Kaufman et al., 1984). In the same way, the **DQA** and **DQB** genes are related to the mouse **Aa** and **Ab** genes, respectively. The mouse β-chain gene, **Ob**, is homologous to the human **DOB** gene and the mouse β-chain pseudogene was found to be most related to β-chain sequences in the human **DP** subregion (Widera and Flavell, 1985).

When the organisation of the genes in the human and mouse class II regions was deduced it was revealed that the relative positions of the homologous subregions are conserved between the two species (Steinmetz et al., 1986). This shows that the overall organisation of the class II region was established before the radiation of rodents and primates about 135 million years ago (Klein, 1986). Differences between the two class II regions, such as the duplication of the human **DQA1/DQB1** gene pair to give **DQA2** and **DQB2**, and the deletion from the mouse class II region of sequences homologous to **DPA1** or **DPA2**, have probably occurred more recently. In contrast, the class II genes of the chicken MHC are more closely related to one another than they are to **DR**, **DQ** or **DO**. Thus, the chicken class II genes probably arose from a primordial class II pair after the separation of the mammalian and avian lineages about 300 million years ago (Kroemer et al., 1990).

The chicken MHC has a different organisation to the mammalian MHCs in that the class I and II genes are interspersed rather than being grouped in distinct regions (Kroemer et al., 1990). The chicken MHC also contains multiple genes encoding a distinct family of cell surface antigens, the B-G molecules, whose function is unknown. There is no mammalian equivalent of the B-G gene cluster, although a mammalian gene in the class I region, myelin/oligodendrocyte glycoprotein: **MOG**, has similarities to B-G antigens (Steinman, 1993). It has been speculated that **MOG** may be a factor in MHC-linked susceptibility to the demyelinating autoimmune condition multiple sclerosis.
As in the human MHC, the mouse class I and II regions are separated by a class III region. The class III regions of mouse and man seem to be conserved, although many of the most recently discovered genes in the human class III region have yet to be mapped in the mouse. Both class III regions contain the genes for steroid 21-hydroxylase and the complement components C2, C4 and factor B in a similar organisation (Chaplin, 1985). There is no distinct class II region in the chicken MHC, but there are a number of interspersed non-class I/ class II/ B-G genes, as yet uncharacterised which may turn out to be homologues of genes in the mammalian class III region (Kroemer et al., 1990). Investigation into the association of chicken class III genes with the MHC may help to resolve the controversy of whether these genes have become associated with the MHC by chance (Klein, 1986) or whether the association of class III genes with the MHC is evolutionarily advantageous and has been maintained by selection (Bodmer et al., 1996).

1.8. The MHC gene cluster

Almost all the genes in the MHC class II region are involved in the antigen processing and presentation pathways. This is in contrast to most other gene families in mammals, whose members are scattered throughout the genome. Interspersed with the class I and II genes are loci that are not involved in either antigen presentation or the immune system. This is true of the class III region which may have originated as an insertion between the class I and II regions. Evidence to support this hypothesis has come from a study of the G+C content of the MHC (Fukagawa et al., 1995). The class III region consists of DNA with a higher G+C% than the class II region. The transition between the AT-rich class II region and GC-rich class III region occurs approximately 180 kb telomeric from DRA in an area containing large blocks of repetitive sequences (Alu and LINE repeats). Iris et al., (1993) also found dense Alu clusters of several tens of kilobases in the telomeric portion of the class III region and detected all the major Alu families classified by Jurka and Smith (1988). These regions may have played a role in a transposition event resulting in the insertion of the class III region into the human MHC. Ring3, mapping between DNA and DOB, is an example of one of the few non-immune genes in the class II region. This gene, as discussed in section 1.6.3, is highly related to the Drosophila gene fsh (female sterile homeotic).
Clustering of genes with related functions is most evident within the class II region where genes involved in the class I antigen processing pathway reside. The protein products of these genes are involved in the degradation of antigenic proteins into peptides (LMPs) and the subsequent transport of the peptides into the ER for assembly with class I molecules (TAPs). Considering these facts, the MHC could be regarded as an organised cluster of genes with related functions. If this is the case, what advantages may be gained by keeping the MHC genes together?

The MHC genes may be co-ordinately regulated. Little is known about large scale gene regulation but it would seem sensible that co-ordinately-regulated genes were located in similar areas of chromatin so that transcription could occur at the same time. Bacterial cistrons are known to be clustered together as operons, allowing co-ordinate regulation. However, transfection experiments have shown that many eukaryotic genes function normally in different genomic positions. One of the best studied gene clusters, the β-globin cluster, is conserved in mammals and is under the control of regulatory elements some distance from the locus. It is thought that individual regulatory elements may act on more than one gene in this locus (Dillon and Grosveld, 1993). The result of this is a situation where expression of one gene can affect the timing and level of expression of another gene. The order of genes within the Hox cluster has been maintained from Drosophila to humans, suggesting that the regulatory principle acting on the β-globin gene cluster may also act in these loci (Krumlauf, 1993). On a smaller scale, the transporter and proteasome-related sequences are expressed in similar circumstances and are all inducible by interferon.

The MHC genes may be maintained together to allow co-evolution of function. Since many MHC genes are involved in aspects of antigen processing and presentation, linkage of the genes might maintain advantageous combinations of alleles at different loci (Trowsdale, 1993). Keeping the MHC loci together also means that they are in a position to exchange sequences by gene conversion or non-homologous recombination. There is evidence for this occurring in the mouse, and circumstantial evidence that sequence exchanges have participated in generating the polymorphism in some class I genes (Belich et al., 1992).

Although the concept of the MHC as a gene cluster encompassing functionally related genes is highly attractive, there is no direct evidence that their linkage is significant. There may not be any particular reason why MHC genes have stayed together. The mouse surfet locus contains six housekeeping genes and has been conserved as a gene cluster for over 600 million years (Colombo et al., 1992).
this case the genes are so tightly linked that there may not have been enough time for them to become separate.

1.9. The MHC and disease associations

The extreme polymorphism of MHC genes has provided a powerful system of markers to study disease association with this locus. In population studies, the frequency of a particular MHC allele is compared, in a group of unrelated, afflicted patients with its frequency in a group of healthy, unrelated individuals. Statistical tests are carried out to determine whether there is a significantly increased or decreased frequency of the allele between the two groups. If a significant difference is found, then that allele is considered to be associated with the disease. In family studies, it is determined whether a disease is inherited with a particular MHC gene more often than expected by chance alone.

Interpreting the results of disease association studies must take into account the phenomenon of linkage disequilibrium. Linkage disequilibrium is defined as the tendency for alleles at different loci to occur on the same chromosome more often than expected by chance and is caused by infrequent recombination between loci (Klein, 1986). Whether linkage disequilibrium is due to selection, genomic structure, or a combination of these factors is hard to establish. Alleles at the class II loci DR and DQ are particularly strongly associated. A consequence of linkage disequilibrium is that if a disease association is found with a particular MHC allele, any gene in linkage disequilibrium with that allele is a candidate disease gene.

Many of the diseases associated with the MHC are of the autoimmune type. Of the MHC-associated diseases more serological data is available on susceptibility haplotypes in insulin-dependent diabetes mellitus (IDDM) than any other disease. The predominant susceptibility haplotypes are HLA-DR3 or HLA-DR4 with a high proportion of DR3/4 heterozygotes (Svejgaard et al., 1983). It has been proposed that the susceptibility gene of DR3 haplotypes is linked to the DQ region (Todd et al., 1987). DQB alleles encoding aspartic acid at position 57 of the DQB molecule are protected against IDDM whereas, non-Asp57 alleles confer susceptibility. This model was extended to include DQA since susceptibility to IDDM in Caucasians was shown to correlate with DQA (Heimberg et al., 1992). It was proposed that non-Asp57 DQB chains and arginine 52 DQA chains in DQαβ dimers also played a role in the pathogenesis of the
disease. The \textit{HLA-DQ-IDDM} association may be caused by the presentation of a peptide derived from pancreatic β cells, the inefficiency to eliminate autoreactive T cells during the process of thymic deletion or the level of \textit{DQ} expression in antigen presenting cells (Bell and Todd, 1989).

Rheumatoid arthritis (RA) is also known to be associated with the MHC (Bell and Todd, 1989). The major serological determinant seen in RA is \textit{DR4}. Unlike diabetes, there are no protective haplotypes in this disease and susceptibility, not protection, seems to be dominant on the \textit{DR4} haplotype. A sequence motif in \textit{DRB1} has been identified which is present in all the disease susceptibility haplotypes. \textit{DR4Dw10}, which confers no susceptibility to disease, differs from other \textit{DR4} haplotypes at \textit{DR} by only three residues in the \textit{DRB1} molecule. Significantly, these involve substantial charge changes.

Coeliac disease (CD) is another autoimmune disease which is linked to the MHC. The etiology of the disease is not fully understood, but it is thought that gluten induces immunologically mediated intestinal damage in genetically predisposed individuals (Rosenberg et al., 1989). Initially, linkage was found between CD and \textit{HLA-DQw2} (Sollid et al., 1989). Results have also been documented linking CD with \textit{DP}. The apparent associations of \textit{DP} with CD can be accounted for by linkage disequilibrium of the \textit{DQw2} and \textit{DR} haplotypes (Rosenberg et al., 1989).

The MHC may also be relevant in controlling cancer. The antigen processing efficiency in various tumour lines was studied using a recombinant vaccinia virus (Vac) to transiently express the K\textsuperscript{d} molecule (Restifo et al., 1993). Three human small cell lung carcinoma cell lines failed to process endogenously synthesised proteins for presentation to K\textsuperscript{d}-restricted, Vac-specific T cells. Pulse-chase experiments showed that MHC class I molecules were not being transported to the cell surface in these cell lines. Northern blot analysis revealed low levels of mRNA for the \textit{TAPs} and \textit{LMP} components of the proteasome. Treatment with γ-interferon enhanced expression of the mRNAs and reversed the functional defects. It was suggested that downregulation of antigen processing could be one of the strategies used by tumours to escape immune surveillance.

Non-autoimmune diseases have also been shown to be associated with the MHC. For example Hodgkin's lymphoma, chronic lymphocytic leukaemia and acute non-lymphocytic leukaemia, all of which are associated with \textit{DP} alleles (Pawelec et al., 1989). Almost 100\% of individuals with the sleep disorder narcolepsy type as \textit{DR2 Dw2 DQw1} (Aldrich, 1990). These findings are
consistent with the tight association of a sleep gene with DR or DQ, or with an effect of the DR and DQ products on sleep induction. The DR and DQ genes from narcoleptics are identical in sequence to those from normal individuals.

In these cases the association with the class II region may be explained by the presence of novel genes with non-immunological functions. This has found to be the case in the association of congenital adrenal hyperplasia with HLA-Bw47 (White et al., 1985). The molecular basis for this disease was found to be due to a deletion in the steroid 21-hydroxylase gene which was in linkage disequilibrium with HLA-Bw47. Recently, an expressed β-tubulin gene (TUBB) has been mapped to the HLA class I region between the HLA-C and HLA-E genes (Volz et al., 1994). Family studies show association of the HLA region with the disease immotile cilia syndrome (ICS). A common cause of the disease is attributable to structural alterations of microtubules (Sturgess et al., 1980) which suggests that TUBB is a candidate gene for ICS.

A third class of diseases associated with the MHC is that of infectious diseases. HLA-B53 has been found to be associated with protection from severe malaria (Hill et al., 1991). It was hypothesised that this protection may be mediated by HLA class I-restricted CTL acting during the liver stage of the parasite's life cycle. A sequence motif for bound peptides was identified by eluting self-peptides from HLA-B53. Peptides bearing this motif have been synthesised from the pre-erythrocytic stage Plasmodium falciparum antigens. These peptides were bound to HLA-B53 and analysed for the ability to elicit CTL responses.

A nonamer peptide from liver-stage-specific antigen-1 (LSA-1) elicited secondary CTL responses in HLA-B53 positive individuals and sequence analysis showed this to be a conserved epitope in the local parasite population. In contrast, similar experiments carried out on the HLA-B35 molecule identified responses to polymorphic epitopes, variants which escape CTL recognition. The association of HLA-B53 with malaria resistance and the observation that HLA-B53-restricted CTL only recognise the Plasmodium falciparum antigen LSA-1 implicates this antigen in provision of protective immunity to malaria.

Although this data provides strong evidence that host genetic susceptibility influences the clinical outcome of malarial infection, a recent report has indicated that the situation is more complicated (Gupta et al., 1994). There may be considerable variation in virulence of the P. falciparum parasite so that the variability in the clinical outcome of P. falciparum infection may be a
consequence of heterogeneity in parasite phenotypes. Disease severity may also depend on the size of the initial parasite inoculum. Thus, it is apparent that the pathogenesis of the disease depends on a complex combination of host and parasite factors and does not depend solely on the immunological status of the host.

As demonstrated here, the MHC is associated with numerous diseases, many of the autoimmune type. Evidence for the linkage of non-autoimmune diseases to the region has been presented, these diseases possibly being caused by novel genes in the region. Thus, a complete understanding of the region will depend on characterising all the genes present and investigating their function.

### 1.10. Approaches to finding novel genes

There is strong circumstantial evidence to support the hypothesis, on which the work described in this thesis is based, that there are previously undiscovered genes in the class II region of the human MHC. Susceptibility to over 40 diseases has been associated with MHC haplotypes and although class I and class II alleles may account for some of these associations, it is unlikely that they will account for all of them. Any new expressed gene in this region is, therefore, a candidate gene for susceptibility to any MHC-associated disorders.

Genomic analysis and gene identification have been essential steps in understanding the MHC. A complete transcription map and sequence of the region not only contributes to a further understanding of the antigen processing pathways and MHC disease associations but also acts as an example for the complete molecular organisation of the human genome.

The analysis of a particular region of the genome for genes controlling a known phenotype without prior knowledge of their functions is termed 'positional cloning' (Collins, 1992). This particular approach has been extremely successful in identifying a number of disease genes, such as Duchenne muscular dystrophy (Monaco et al., 1987), cystic fibrosis (Rommens et al., 1989) and more recently the breast/ovarian cancer gene (Miki et al., 1994). Over the last several years, this approach has also proved highly successful in the discovery of novel genes in the human MHC, as described in section 1.6.2. Some of the methods which can be used to map a specific area have already been discussed and so this section will concentrate on gene identification techniques once the area of interest has been mapped and cloned.
Many methods are available for the identification and isolation of genes in a region-specific manner and several of the more common methods are described in this section. One way to tackle the problem of gene identification has been to find CpG islands, described in section 1.6. CpG islands have also been called Hpa II tiny fragment (HTF) islands. This nomenclature refers to their characterisation by clusters of sites for methylation sensitive, rare-cutter restriction enzymes with CpG dinucleotides in their recognition sequences. Here they will be referred to as CpG islands. The function of these regions is not yet known, but they are found to colocalise with the 5' ends of expressed sequences (Bird, 1987). In humans, about 60% of genes are associated with CpG islands, including all housekeeping genes so far analysed and about 40% of tissue-restricted genes (Larsen et al., 1992). In most cases, the island contains the promoter and one or more exons of its associated gene.

A representative library of human CpG islands has recently been described (Cross et al., 1994). This provides a collection of DNA segments corresponding to 60% of human genes allowing the selection of appropriate cDNA clones and, since CpG islands contain most of the sites for rare-cutting restriction enzymes they are thus landmarks for physical mapping of the genome. The procedure made use of the rat chromosomal protein, MeCP2, which binds DNA that is methylated at CpG. A DNA binding column was made from this protein which was used to separate DNA solely on the basis of its degree of methylation. The resulting CpG islands were then purified and cloned.

Another method for gene isolation is exon amplification (Buckler et al., 1991). This approach, based on the selection of functional splice sites in genomic DNA, has been used recently to identify the neurofibromatosis type 2 tumour suppressor gene (Trofatter et al., 1993), amongst others. Exon amplification is based on the in vivo selection for splice sites flanking exon sequences in genomic DNA. The great advantage of this technique is that it is independent of tissue expression patterns and can lead to the isolation of rarely expressed transcripts, transcripts with a limited expression pattern, and developmentally expressed transcripts. Mammalian genomic DNA segments are inserted into an intron of the HIV-1 tat gene which is flanked by splice sites and exons of the viral gene, contained within the vector pSPL1. Upon transfection into COS cells, the reporter gene, with inserted genomic fragment is transcribed using an SV40 early promoter. When a fragment containing an entire exon with flanking intron sequence is present, the exon is retained in the mature poly A+ cytoplasmic RNA. RNA-based PCR
amplification using oligodeoxynucleotide primers specific for the plasmid exons is then used to detect the presence of the exon. Recently, several improvements have been made to the original vector decreasing the number of false positives obtained and increasing the sensitivity of the procedure (Church et al., 1994). It is now estimated that an exon can be identified for every 25-85 kb of genomic sequence analysed, which is dependant on the complexity of the target DNA which can range from $3 \times 10^6$-6 $\times 10^7$ bp per experiment.

A method has been recently developed that allows the rapid isolation of transcripts defined by overlapping clone libraries and is termed cDNA selection (Korn et al., 1992). DNA from cosmids or YACs is sheared to an average size of 500-800 bp, photobiotinylated and depleted of repetitive sequences by hybridisation to total human DNA covalently bound to cellulose. The resulting material is then hybridised, in liquid, with inserts of amplified cDNA libraries. The resulting hybrids are bound to streptavidin coated magnetic beads. After a series of stringent washes, the cDNAs are eluted from the beads and reamplified. The selected cDNAs are cloned in Bluescript phagemid resulting in a region and tissue specific library. The selection procedure is advantageous as it is insensitive to the number and size of introns and cryptic splice sites. The method detects 5' and 3' exons and genes that lack CpG islands (Parimoo et al., 1991).

Other strategies for gene identification involve cross-hybridisation of fragments from genomic clones to genomic DNA from other species (Monaco et al., 1986), hybridising fragments from genomic clones to cDNA libraries (Elvin et al., 1990), island rescue PCR (Valdes et al., 1994) and sequence analysis of large segments of genomic DNA in search of open reading frames. This final approach has been initiated in the class II region of the MHC aiming to sequence the entire region (Beck et al., 1992b). Such sequence data can be used to analyse the intron/exon structure of genes and can give an indication of their function and origin. Sequence of non-coding regions can provide information concerning evolution of stretches of sequence by analysing repeat sequences, specifically Alu repeats, which can be classified according to their evolutionary origin. This is discussed in more detail in chapter 4. Recently, a database of the human MHC has been compiled which allows access, retrieval and display of physical and genetic data relating to the human MHC (Newell et al., 1994).

These approaches are applicable to any part of the human genome and have been implemented in the characterisation of the class II region of the human MHC. The
following chapters of this thesis describe the application of molecular mapping and cloning techniques in the class II region to completely clone the region in overlapping genomic clones, as when this thesis was started gaps remained in the class II cosmid map, to identify genes within these regions and to characterise any novel genes found. To complement the project on the genomic organisation of the MHC a collaboration between our laboratory and the laboratory of Dr. Stephan Beck (DNA Sequencing Laboratory, ICRF) has been initiated, aiming to sequence the entire class II region of the MHC. The ultimate aim of this research is to provide the first complete picture of a region, identifying all the genes present, their genomic organisation and function.
CHAPTER 2: MATERIALS AND METHODS

2.1. Solutions

L-broth (LB)  
10g bactotryptone  
5g bacto yeast extract  
10g NaCl  

to 1 litre with distilled water

L-agar plates  
1.5% w/v bacto agar in LB  
Sterilised by autoclaving  
Solid media were melted by microwaving and cooled to 50°C before adding antibiotics

GTE  
50mM glucose  
25mM Tris.Cl pH 8.0  
10mM EDTA pH 8.0

PCIA  
25 parts phenol pH 8.0  
24 parts chloroform  
1 part isoamyl alcohol

CIA  
24 parts chloroform  
1 part isoamyl alcohol

TE  
10mM Tris.Cl pH 7.6  
1mM EDTA pH 8.0

Low TE  
10mM Tris. Cl pH 8.0  
0.1mM EDTA pH 8.0

20X SSC  
3M NaCl  
0.3M sodium citrate pH 7.0
Southern hybridisation solution

- 6X SSC
- 5X Denhardt's solution
- 0.5% w/v SDS
- 10% w/v dextran sulphate
- 20μg/ml salmon sperm DNA

100X Denhardt's solution

- 20g BSA
- 20g ficoll 400
- 20g polyvinylpyrrolidone-40
to 1 litre with distilled water

Salmon sperm DNA

500mg salmon sperm DNA was cut into small pieces with scissors. The DNA was dissolved in 50ml distilled water by stirring overnight. The solution was sheared by three passes through a 19g needle and boiled for 10 mins.

Denaturing solution (DS)

- 1.5M NaCl
- 0.5M NaOH

Neutralising solution (NS)

- 1.5M NaCl
- 0.5M Tris.Cl pH 7.2
- 1mM EDTA

2YT

- 10g bactotryptone
- 10g yeast extract
- 5g NaCl
to 1 litre with distilled water

10X TBE

- 108g Tris base
- 55g boric acid
- 40ml 0.5M EDTA
to 1 litre with distilled water
10X TAE

48.4g Tris base
11.42ml glacial acetic acid
20ml 0.5M EDTA

to 1 litre with distilled water
2.2. Screening recombinant DNA libraries

2.2.1. cDNA libraries

Four main cDNA libraries were utilised in this study: JY (B-lymphoblastoid cell line), DX3 (malignant melanoma), γ U937 (γ-IFN induced macrophage) and NW (B-lymphoblastoid cell line). These were a gift from Dr. David Simmonds (ICRF, Oxford). The libraries were constructed according to the method of Seed (1987) in the plasmid vector CDM8, and propagated in Escherichia coli MC1061/P3. The titer of the cDNA library was first determined by preparing a 10⁻² dilution of the frozen library stock and plating out dilutions on L-agar plates containing ampicillin and tetracycline at concentrations of 50 μg/ml and 7.5 μg/ml respectively. These plates were incubated overnight at 37°C and the 10⁻² dilution stored at 4°C.

The following day, the number of colonies obtained for each dilution were counted and used to calculate the volume of the 10⁻² dilution required to yield 250,000 colonies (optimal number for a 20 X 20cm filter). The calculated dilution volume was then spread onto four 20 X 20cm Hybond N+ membranes (Amersham) which were on the surface of L-agar (ampicillin and tetracycline selection) 245 X 245mm Nunc plates. These master plates were incubated at 37°C overnight and used the next day to prepare replica filters as follows. 12 fresh L-agar (ampicillin and tetracycline selection) 245 X 245mm Nunc plates were poured and overlaid with Hybond N+ membranes. Two replica filters were prepared from each master. The first master was removed from its agar plate and placed on 2 sheets of Whatman 3MM paper. The prewetted replica filter was carefully placed on top of the master. 2 sheets of Whatman 3MM paper were layered on top of the filters and the filters pressed together. The master filter was marked with a pattern of spots and the same pattern was also marked on the replica. The two filters were pulled apart and the replica filter returned to its plate. This process was repeated with the remaining 3 master filters resulting in 4 duplicated replica filters (8 in total). All 12 plates were incubated at 37°C for 4 hrs, the master plates were then stored at 4°C and the replica plates placed at room temperature overnight to regrow the colonies.

Once the colonies had regrown, the replica filters were processed as follows. Each filter was placed, colony side up, on a pad of Whatman 3MM paper soaked in DS for 9 mins. The filters were blotted dry and then dipped briefly in NS. DNA
was fixed onto the filters by placing them, colony side up, on pads of Whatman 3MM soaked in 0.4M sodium hydroxide (NaOH). The filters were blotted dry then placed on Whatman 3MM pads soaked in NS for 2 X 3 mins. Bacterial colonies stuck to the filters may lead to high background hybridisation levels so these were washed off using a tissue soaked in 5X SSC. The filters were rinsed in 5X SSC then blotted dry and stored at 4°C.

After hybridisation of the filters with the probe of interest, the presence of a positive hybridisation signal on the duplicate filters was used to identify the positive colony on the master plate. A 5mm area around the positive colony was scraped off the master plate using a sterile inoculating loop. This was transferred to 1ml LB containing ampicillin and tetracycline. After mixing, successive 10 fold dilutions were made of the stock tube and plated out onto 140 X 140mm L-agar plates (containing 50µg/ml ampicillin and 7.5µg/ml tetracycline). The stock tube was stored at 4°C and the plates were incubated at 37°C overnight. The next day, filter lifts were taken from those plates with well separated colonies. A Hybond N+ filter was overlaid onto the colonies for 1 min. During this time orientation marks were made on the filter and plate using a needle. The secondary filter was then placed, colony side up, on Whatman 3MM pads soaked in DS for 7 mins, blotted dry and placed on NS for 2 X 3 mins. The DNA was fixed to the filters by placing them on Whatman 3MM pads soaked in 0.4M NaOH. The filters were briefly rinsed in 2X SSC and hybridised again with the probe of interest. The plates were incubated at 37°C until the colonies had regrown and then they were stored at 4°C.

Single colonies positive with the probe were identified by hybridisation and autoradiography. These were picked into 5ml LB (ampicillin and tetracycline selection) and cDNA purified as described in section 2.6.

2.2.2. Cosmid libraries

Two cosmid libraries were utilised in the course of the work described in this report: 1) an enriched cosmid library made from the yeast artificial chromosome (YAC) 11.2 which is 450 kb in size and encompasses the region between the genes DMB and DOB in the class II region of the human MHC; 2) a flow-sorted chromosome 6 cosmid library. The construction and utilisation of these cosmid libraries is discussed, in detail, in chapter 3.
In parallel with the cosmid library constructed from YAC 11.2 and the flow-sorted chromosome 6 cosmid library, clones were also isolated from a P1 bacteriophage library.

2.2.3. P1 genomic library

The total P1 library (Francis et al., 1995) was prepared from DNA derived from the lymphoblastoid cell line GM1416B with a karyotype 48 XXXX. DNA was cloned into pAd10SacB11 vector and clones were recovered in the E. coli strain NS3145. The library was robotically arrayed onto high density grids, each containing 20736 clones (Nizetic et al., 1991) and represents a 1.2 fold genome coverage.

2.3. Preparation of DNA probes

Probe DNA fragments were excised directly from normal agarose gels, run in 1X TAE buffer and purified by centrifugation through a Whatman 3MM plug.

20-50ng DNA were diluted to a volume of 32μl with distilled water, boiled for 5 mins and cooled on ice. 10μl OLB, 2μl 10mg/ml BSA, 4μl [α-32P]dCTP (10μCi/μl) and 2μl Klenow fragment (5U/μl, NBL) were added and the reaction incubated at 37°C for 1 hour or, for probes < 700 bp in size, 4 hrs at room temperature. Unincorporated nucleotides were separated from probe DNA by centrifugation through a Sephadex G-50 column. This was prepared in a 1ml syringe plugged with glass wool and pre-equilibrated with TES. The labelled probe was collected and boiled for 5 mins before being added to the hybridisation solution.

Probes containing repetitive sequences (cosmid or P1 probes) were competed with human placental DNA (SIGMA) before use. 300μg human placental DNA was added to 50ng probe in a final concentration of 6X SSC. The mixture was boiled for 5 mins and incubated at 65°C for 100 mins before addition to the hybridisation solution.

-67-
OLB: 
- 100μl solution A
- 250μl solution B
- 150μl solution C

Solution A: 1.25M Tris.Cl pH 8.0, 0.125M MgCl₂, 10μl β-mercaptoethanol; 5μl each of 100mM dATP, dGTP, dCTP, dTTP (Pharmacia)

Solution B: 2M HEPES pH 6.6 (adjusted with NaOH)

Solution C: Hexanucleotide primers (Pharmacia) suspended in TE pH 8.0 at 90 OD₂₆₀ U/ml

TES: 10mM Tris.Cl pH7.6, 5mM EDTA, 0.2% SDS

2.3.1. Kinase reaction to label oligonucleotides

In some cases oligonucleotides of between 30-40 bp were used as probes for hybridisation reactions. These were labelled using T4 kinase and ³²P γ-ATP. For 100ml hybridisation solution the following mix was prepared: 500ng oligonucleotide, 100μCi ³²P γ-ATP, 10U T4 kinase and 1X kinase buffer in a final volume of 20μl. This mixture was incubated for 1 hr at 37°C and then directly added to the hybridisation solution. Hybridisations were carried out at 42°C. A typical wash would be 6X SSC, 0.1% SDS for 30 mins followed by 2X SSC, 0.1% SDS for 30 mins. The washing temperature was determined by the Tm of the oligonucleotide minus 12°C (Sambrook et al., 1989).

10X kinase buffer: 500mM Tris.Cl pH 7.5
- 100mM MgCl₂
- 150mM DTT
- 1mM EDTA
- 1mM spermidine
2.4. Filter hybridisation

Filters were prehybridised for 1-8 hrs at 65°C in hybridisation solution. Labelled probe, $5 \times 10^5$ cpm probe per ml (Sambrook et al., 1989), was added to fresh hybridisation solution and incubated with the filters overnight at 65°C. For high stringency hybridisations filters were washed down to a final concentration of 0.1X SSC, 0.1% SDS. When a cross species hybridisation was being performed lower stringencies were used, typically 6X SSC at 65°C. When hybridising the flow-sorted chromosome 6 cosmid library an initial wash was carried out down to 3X SSC, 65°C and the filters exposed to film. This stringency gave high background hybridisation enabling accurate positioning of cosmid clones on the cosmid grid. Further washing was then carried out down to 0.1X SSC to pinpoint the positive colony. Filters were exposed to Kodak XAR-5 autoradiography film at -70°C, backed with an intensifying screen.

Positive clones isolated from the cosmid library derived from YAC 11.2 were identified and picked as described in section 2.2.1. Positive clones from the chromosome 6 cosmid library were identified from their position on the cosmid grid. The cosmid clones were robotically arrayed on the filters in the pattern shown below.

![Diagram of cosmid grid](image)

Blocks of clones were gridded in a repeating pattern across the filter giving 144 X 144 colonies.

The x and y co-ordinates of the positive cosmid clones were identified and used to isolate a frozen glycerol stock of the appropriate clone. Clones identified in this manner were received as an agar stab. They were then streaked out onto agar plates containing 30μg/ml kanamycin and grown overnight at 37°C. Miniprep cultures of several colonies were carried out as described in section 2.7.
2.5. Transformation of bacterial cells with DNA by electroporation

2.5.1. Preparation of electro-competent *E. coli* cells

5ml of LB was inoculated with a single colony of the desired *E. coli* strain and grown overnight, shaking. 2.5ml of this culture was inoculated into 500ml LB in a 2 litre flask and grown at 37°C, shaking, to an OD$_{600}$ of 0.5-0.7. The cells were chilled in an ice-water bath for 15 mins, transferred to a prechilled 500ml centrifuge bottle and spun for 20 mins, 4000rpm at 2°C in a Beckman J-6B centrifuge.

The supernatant was decanted and the pellet resuspended in 5ml ice cold distilled water. A further 500ml ice cold water was added to the cells and mixed by inversion. The cells were centrifuged as before and the supernatant quickly decanted. The pellet was resuspended in ice cold water and centrifuged as before. The supernatant was decanted again and the cells resuspended by swirling in the residual liquid. 40ml ice cold 10% glycerol was added, and the cells pelleted at 6000rpm, 2°C, 10 mins. The cells were resuspended in an equal volume of 10% glycerol, aliquoted (50μl) into Eppendorf tubes, frozen on dry ice and stored at -70°C.

2.5.2. Electroporation

5-500pg of DNA in 1μl was added to 50μl thawed competent cells and mixed gently. The DNA and cells were transferred to a 0.2cm cuvette which had been chilled on ice for 5 mins. The cuvette was tapped slightly to settle the cells to the bottom and placed in the sample chamber of the electroporator which was set to 2.5kV, 25μF and 200Ω. The pulse was applied, the cuvette removed and 1ml SOC medium immediately added. The mixture was pipetted into a sterilin tube and incubated for 45 mins at 37°C with moderate shaking. Aliquots of the transformation culture were plated onto LB plates containing the appropriate antibiotics.
SOC medium:
0.5% Yeast extract
2% Tryptone
10mM NaCl
2.5mM KCl
10mM MgCl₂
10mM MgSO₄
20mM glucose

2.6. Preparation of DNA from transformed bacterial cells

2.6.1. Small scale plasmid preparations (mini-preps)

5ml aliquots of LB containing the appropriate antibiotic(s) were inoculated with single bacterial colonies containing the desired plasmid constructs. These were incubated at 37°C overnight, shaking. 1.5ml of the cultures were transferred into Eppendorf tubes and centrifuged for 1 min. The remainder of the cultures were stored at 4°C. The medium was removed by aspiration and the pellets resuspended in an ice cold solution of GTE. The tubes were left at room temperature for 5 mins. The cells were lysed by addition of 200μl of a freshly prepared solution of 0.2M NaOH/1% SDS. The tubes were mixed by rapid inversion and stored on ice for 5 mins. The bacterial chromosomal DNA was precipitated by adding 150μl ice cold potassium acetate (KAc) pH 4.8, vortexing and incubating on ice for 5 mins. The chromosomal DNA was pelleted by centrifuging at room temperature for 5 mins. The supernatants were transferred to fresh tubes containing an equal volume of PCIA. These were vortexed and centrifuged for 2 mins. The supernatants containing plasmid DNA were transferred to fresh tubes and precipitated by addition of 2 volumes of absolute ethanol. The tubes were left at room temperature for 5 mins and the DNA pelleted by centrifugation for 10 mins. The supernatants were decanted and the pellets washed in 70% ethanol. The pellets were then dried for 10 mins in a vacuum dessicator and resuspended in 30μl TE pH 8.0 containing 20μg/ml DNase-free RNase. 2-5μl were analysed by restriction endonuclease digestion.
2.6.2. Large scale plasmid preparations (maxi preps)

5ml of LB containing the appropriate antibiotics was inoculated with a single bacterial colony and grown during the day at 37°C, shaking. This culture was used to inoculate 400ml LB in a 1 litre flask and further grown at 37°C, shaking, overnight. The bacterial cells were harvested in a 500ml centrifuge bottle by spinning at 6000rpm for 10 mins at 4°C using a Beckman JA10 rotor. The pellet was resuspended in 10ml of GTE and left at room temperature for 5 mins. The suspension was transferred to a 100ml glass flask on ice, and 20ml 0.2M NaOH/0.1% SDS added. The suspension was mixed by swirling and left for 5 mins on ice to lyse the cells. 10ml of 5M KAc pH 4.8 was then added, mixed by sharp inversion and incubated on ice for 15 mins. The bacterial DNA and cell debris were precipitated by centrifugation for 15 mins, 3000rpm, 4°C in a Heraeus Minifuge T. The supernatant was strained through gauze into a 200ml glass bottle and the DNA precipitated by the addition of 0.6 volumes of isopropanol. After incubation at room temperature for 30 mins, the DNA was pelleted by centrifugation at 2000rpm for 15 mins in a Beckman J-6B. The supernatant was discarded and the pellet allowed to air dry. The crude plasmid DNA was then resuspended in 3ml TE pH 8.0. This was added to 4.4g caesium chloride (CsCl) in a sterilin tube along with TE pH 8.0 until a total of 4g of liquid had been obtained. 400μl of 10mg/ml ethidium bromide was added and the density of the solution adjusted to 1.5-1.6g/ml. The solution was centrifuged at 3000rpm for 10 mins at room temperature in a Heraeus minifuge T to precipitate bacterial proteins. The clear supernatant was transferred to a Beckman polyallomer Quick Seal centrifuge tube and the tube filled with 1g/ml CsCl in TE pH 8.0. The tubes were sealed and centrifuged at 58000rpm in an ultracentrifuge, 15°C for 18 hrs.

The plasmid DNA band in the resulting CsCl gradient was visualised under ultraviolet (uv) light. The top of the tube was pierced with a 25g needle and the plasmid DNA extracted with a 19g needle in 0.5-0.7 mls solution. Care was taken to avoid both the RNA band down the side of the tube and the upper nicked plasmid DNA band. The ethidium bromide was removed from the solution by repeated butan-2-ol extractions. The volume was increased three fold by addition of TE pH 8.0, two volumes of absolute ethanol were added and the DNA precipitated by centrifugation at 4000rpm, room temperature in a Heraeus minifuge T. The supernatant was decanted and the pellet resuspended in 400μl TE pH 8.0. This was transferred to an Eppendorf tube and precipitated by addition of
40μl NaAc pH 5.2 and two volumes of absolute ethanol. The DNA was pelleted by centrifugation for 10 mins in a microfuge, room temperature. Two further precipitations were performed before washing the pellet in 70% ethanol and final resuspension in 400μl TE pH 8.0 containing 20μg/ml DNase-free RNase. The yield was determined by measuring the absorbance of an aliquot at OD\textsubscript{260nm}.

2.6.3. Small scale P1 preparations

A single bacterial colony was inoculated into 10ml 2YT medium containing 25μg/μl kanamycin and grown overnight, shaking, at 37°C. The cells were spun down at 2000rpm for 10 mins in a Beckman centrifuge and the pellet resuspended in 300μl GTE. The solution was transferred to an Eppendorf tube and incubated on ice for 5 mins. 600μl of a freshly prepared solution of 0.2M NaOH/1% SDS was added, mixed by inversion, and the tube incubated on ice for 5 mins. The bacterial DNA was pelleted by centrifugation for 10 mins at room temperature in a microfuge. The supernatant was transferred to a fresh Eppendorf tube, the tube filled with isopropanol, and the P1 DNA left at room temperature for 30 mins to precipitate. The tubes were spun in a microfuge for 15 mins and the supernatant decanted. The pellet was air dried and resuspended in 100μl TE pH 8.0. The P1 DNA pellet was purified as follows. Ammonium acetate was added to a final concentration of 2.5M, the solution incubated on ice for 10 mins, and the debris pelleted by centrifugation for 10 mins in a microfuge. The supernatant was transferred to a fresh Eppendorf tube and extracted once with phenol, once with PCI and once with CIA. The aqueous phase was precipitated by the addition of NaAc pH 5.2 to a final concentration of 2.5M and two volumes of ethanol. The DNA was precipitated by centrifugation for 10 mins at room temperature in a microfuge. The supernatant was decanted, the pellet dried in a vacuum dessicator and resuspended in 20μl TE pH 8.0.

2.7. Restriction endonuclease digestion of DNA

For mapping and analysis of plasmid, cosmid and P1 DNA with restriction endonucleases, 200-600ng were digested in a total volume of 20μl, with 5-10 fold excess of enzyme for 1-2 hrs. When specific fragments were required for probe preparation or subcloning, the digests were scaled up accordingly. The volume of enzyme added not being allowed to exceed 10% of the total digest volume.
Digests were incubated at the optimum temperature in the appropriate buffer conditions as recommended by the manufacturer.

For preparation of genomic Southern blots 10µg of genomic DNA was digested overnight with a 10 fold excess of enzyme in a volume of 50µl. One tenth of the digest was run on a minigel to check the extent of digestion. Incomplete digests were diluted, supplemented with restriction buffer and enzyme before continuing digestion.

2.8. Electrophoresis of DNA

2.8.1. Agarose gel electrophoresis

All DNA samples were mixed with 0.2 volumes of loading dye before electrophoresis. Most restriction digests of plasmid, cosmid and P1 DNA were resolved in 0.8% agarose gels in 1X TAE buffer containing 1µg/ml ethidium bromide. The gels (10cm) were run in Hybaid apparatus at 50V for 3-4 hrs. Genomic digests and cosmid digests required for accurate mapping were run on longer (20cm) gels at 30V, overnight. The agarose concentration was varied between 0.6 and 2% depending on the size of fragments to be separated. The size markers used were either HindIII cut bacteriophage λ and 100 Base-Pair Ladder (Pharmacia).

5X loading buffer:  50% glycerol
                   0.2% bromophenol blue
                   0.2% xylene cyanol
                   0.1M EDTA pH 8.0

2.8.2. PFGE of P1 digests

PFGE was performed using a CHEF DRII apparatus (Biorad). A 150ml, 1% agarose gel was poured in 0.5X TBE, using a 30 well comb to create the wells. 125ng λ bacteriophage DNA concatamers (0.25 block) were used as size markers. These were loaded into the wells using sterile inoculating loops. 10µl of the P1 miniprep was directly loaded into the wells and the top of the wells sealed with 1% agarose in 0.5X TBE at 50°C. The gel was carefully placed in the tank containing 2 litres 0.5X TBE which was recirculated at 14°C. The gel was run at
180V for 20 hrs with a constant time pulse of 3-20 secs for undigested DNA and 4-20 secs for digested DNA. This separates molecules in the size range of 20-250 kb.

2.9. Southern blotting

After electrophoresis, a photograph of the gel was taken before soaking the gel in 0.25M HCl for 15 mins to partially depurinate the DNA. The depurinated sites are cleaved during the denaturing step, fragmenting long molecules resulting in more efficient transfer of high molecular weight DNA molecules to the filter. This step is unnecessary for fragments less than 10 kb in size. The gel is then rinsed in distilled water and soaked in DS for 35 mins. Following this, the gel is soaked for 2 X 15 mins in NS and capillary blotted onto nitrocellulose filters (Hybond N+) in 20X SSC for 18 hrs. The DNA was fixed to the filters either by baking the filters at 80°C for 1 hour in a vacuum oven or by soaking the filters in 0.4M NaOH for 20 mins. The filters were rinsed in 2X SSC before prehybridisation.

2.10. Sequencing of double stranded plasmid DNA

2-5µg of plasmid DNA in 8µl TE pH 8.0 were denatured by addition of 2µl 2M NaOH/20mM EDTA pH 8.0 and left at room temperature for 5 mins. The reaction was neutralised by addition of 3M NaAc pH 5.2. 7µl of distilled water was added and the DNA precipitated by addition of 75µl absolute ethanol. The mixture was incubated at 4°C for 10 mins, spun, washed with 70% ethanol, desiccated and the pellet resuspended in 7µl distilled water.

Sequencing was carried out based on the chain termination method of Sanger (Sanger et al., 1977) using the sequenase kit from USB. 2µl of sequencing buffer and 1pmol of primer were added to the DNA. The solution was heated at 65°C for 2 mins then cooled slowly to <35°C to allow the primer and template to anneal. 1µl 0.1M DTT, 2µl 1:5 fold diluted labelling mix, 0.5µl 35S dATP (10µCi/µl, Amersham), 2µl 1:8 fold diluted sequenase were added to the annealed DNA mixture and the reaction incubated at room temperature for 2-5 mins. 3.5µl aliquots of the labelling reaction were added to each of four tubes containing pre-warmed termination mix (ddG, A, T, C) and the reaction incubated at 37°C for 5 mins. The reactions were terminated by addition of 4µl stop solution. Immediately prior to use the samples were heated to 75°C for 2 mins then cooled on ice.
5X reaction buffer: 200mM Tris.Cl pH 7.5
100mM MgCl₂
250mM NaCl

5X labelling mix: 7.5μM each of dGTP, dCTP, dATP and dTTP

Enzyme dilution buffer: 10mM Tris.HCl pH 7.5
5mM DTT
0.5 mg/ml BSA

ddG termination mix: 80μM each of dGTP, dCTP, dATP and dTTP
8μM ddGTP
50mM NaCl

ddC termination mix: 80μM each of dGTP, dCTP, dATP and dTTP
8μM ddCTP
50mM NaCl

ddA termination mix: 80μM each of dGTP, dCTP, dATP and dTTP
8μM ddATP
50mM NaCl

ddT termination mix: 80μM each of dGTP, dCTP, dATP and dTTP
8μM ddTTP
50mM NaCl

Stop solution: 95% formamide
20mM EDTA
0.05% bromophenol blue
0.05% xylene cyanol

Sequencing gel electrophoresis was performed using Koch-Light apparatus. Before each run the gel plates were washed and cleaned with ethanol, acetone and distilled water. The back plate was siliconised with Repelcote. The plates were fastened together using Bulldog clips, separated by 0.4mm spacers. A 50ml urea denaturing acrylamide gel mix in 1X TBE was prepared and injected between the
plates with a syringe. This was allowed to set for 30 mins with the comb in
position. The gel was placed in the tank, 1 litre of 1X TBE poured in the
reservoirs and 2.5μl of each sample loaded. Gels were run at 40W for
approximately 2 hrs.

After the run, the gel was fixed in a tray containing 10% glacial acetic acid,
10% methanol for 15 mins. It was then transferred to Whatman 3MM paper and
dried at 80°C for 30 mins on a Biorad slab drier under vacuum. The gel was
exposed to Kodak XAR-5 film overnight at room temperature. Sequences were
analysed using Intelligenetics GCG and Blast computer software.

2.11. Shotgun sequencing

cDNA and whole cosmid inserts were completely sequenced by the shotgun
method of Bankier (Bankier et al., 1987). The DNA was isolated, self ligated,
sonicated, end repaired and ligated into Smal-cut M13mp18.

2.11.1. Vector preparation

5μg M13mp18 replicative form was digested with 20U Smal, using the
appropriate digestion buffer in a total of 50μl for 1 hour at room temperature. A
0.5μl aliquot of the reaction was run on a 0.8% agarose gel to assay for complete
digestion. After complete digestion 1μl 0.5M EDTA pH 8.0 was added to the
reaction. The solution was extracted with phenol and the DNA precipitated by
addition of 5μl 3M NaAc pH 5.2 and 2.5 volumes of ice cold ethanol for 10 mins
at -70°C. The tubes were centrifuged for 10 mins in a microfuge, the supernatant
discarded and the pellet washed in 70% ethanol. The centrifugation was repeated,
the pellet dried under vacuum and resuspended in 25μl TE pH 8.0. 20U CIAP was
added and the reaction incubated for 1 hour at 37°C.

The vector was purified by electrophoresis on a 0.8% agarose gel in 1X
TAE containing 0.5μg/ml ethidium bromide. The gel was run at 50V for 2 hrs and
visualised under uv light. The linear vector band was excised from the gel using a
scalpel and purified by electroelution. The agarose block was transferred to a
dialysis bag containing a minimal amount of 1X TAE. The bag was sealed at both
ends and placed in an electrophoresis tank containing 1X TAE. Electrophoresis
was carried out at 100V for 3-4 hrs by which time the DNA had passed out of the

-77-
gel into the surrounding solution. The solution was collected, phenol, PCIA, CIA extracted, ethanol precipitated and resuspended in TE pH 8.0 at 20ng/μl.

2.11.2. Insert preparation

~50μg of either the cDNA or cosmid clone of interest were digested with 10 fold excess of the appropriate restriction enzyme under optimal conditions to obtain complete digestion. The digested products were run out on a 0.8% agarose gel in 1X TAE until separation of the desired bands had been achieved. Either the cDNA insert or the cosmid fragment was excised from the gel using a scalpel. Purification was achieved by electroelution and phenol extractions as described in section 2.11.1. The resulting, pure fragments were resuspended in 70μl TE pH 8.0 and 0.5μl run on an analytical 0.8% agarose gel to estimate the DNA concentration.

10μg of the fragment was self-ligated in 100μl total volume by 1 X 10^7 U of T4 DNA ligase in 1X ligation buffer (NEB) at 14°C for 16 hrs. The following day 5μl of the ligation reaction was run out on a 0.8% agarose gel, 8V, overnight. To ensure efficient self-ligation λ HindIII was as a marker. If the self-ligation was effective the fragments were diluted to a total volume of 300μl with low TE pH 8.0 and sonicated to give fragment sizes of ~500 bp (3 X 7 second bursts, setting 6, 1 min cooling between bursts, on ice). 10μl of the sonicated DNA was run on a 1.5% agarose gel using 100 Base-Pair Ladder (Pharmacia) to check that fragment sizes of the desired length had been achieved. If not, sonication was repeated, checked, and the DNA ethanol precipitated overnight at -20°C. The sonicated fragments were then washed with 70% ethanol, vacuum dried and resuspended in 60μl low TE pH 8.0. The fragment ends were repaired with 16 and 20 units (5U/μl) of Klenow and T4 DNA polymerase respectively, and 1μl each of 100mM dATP, dCTP, dGTP and dTTP added. The DNA was size fractionated in a 0.8% agarose gel, excising fragments in the size range 350-700 bp. These were purified by phenol extraction. The DNA was ethanol precipitated and resuspended in 50μl low TE pH 8.0.

2.11.3. Ligation

Four ligation reactions were set up for each insert containing 0.1, 1, 2 and 4μl of insert DNA and 20ng vector DNA. Ligations were carried out in a total of 20μl
using 2 X 10^6 U T4 DNA ligase in the appropriate buffer conditions at 14°C overnight. Control ligations were also set up, one containing no insert DNA and the other containing no insert DNA, and no ligase. Ligations were transfected into JM101 competent bacteria.

2.11.4. Preparation of competent cells

A single colony of JM101 bacteria was inoculated into 5ml of 2YT medium and grown overnight, shaking, at 37°C. 300µl of this overnight culture was added to 30ml 2YT medium in a 500ml flask and grown to an OD_600 of 0.4-0.6. The culture was transferred to a Falcon 50ml disposable tube and centrifuged at 2000rpm for 10 mins at 4°C in a Heraeus minifuge T. The pellet was resuspended gently in 2.5ml cold TFB and incubated on ice for 15 mins. 100µl DMSO was added to the suspension and incubated on ice for 15 mins. 100µl of DTT/KAc was added and, after 10 mins, a further 100µl of DMSO. The cells were kept on ice until needed for transformation.

| TFB:          | 10mM methyl ethane sulphonate |
|              | 100mM rubidium chloride       |
|              | 45mM manganese chloride      |
|              | 10mM calcium chloride        |
|              | 3mM hexaminecobaltic chloride |

| DTT/KAc:     | 2.25M dithiothritol           |
|              | 40mM potassium acetate pH 6.0 |

2.11.5. Transformation

400µl competent cells and 20µl ligation reaction were added to a sterile glass culture tube and incubated on ice for 45 mins. The cells were heat shocked at 42°C for 3 mins and 6ml molten top agar (45°C), 100µl 2% Xgal and 50µl 2.5% IPTG added. The bacteria in top agar were quickly poured onto a pre-warmed (37°C) agar plate and swirled to effect an even distribution. The top agar was allowed to set and the plates incubated at 37°C overnight.
Top agar: 10g bactotryptone
8g sodium chloride
8g bacto agar
to 1 litre with distilled water

X-gal: 2% 5-bromo-4-chloro-3-indolyl-β-galactoside in DMF
(dimethylformamide)

IPTG: 2.5% isopropyl-β-D-thiogalactopyranoside in water

2.11.6. Template preparation

A single colony of JM101 was inoculated into 10ml 2YT and grown, with
shaking, at 37°C overnight. This was diluted 1:100 with 2YT. Individual plaques
were picked into 1.5ml aliquots of diluted JM101 and grown at 37°C for 4.5-5.5
hrs with vigorous shaking. The bacterial cultures were pelleted by centrifugation
in a microfuge for 5 mins and the phage-containing supernatant transferred to a
fresh tube. The phage were precipitated by the addition of 200μl 20% PEG
solution. The supernatant/PEG was centrifuged for 5 mins in a microfuge and the
PEG removed. The tube was respun for 30 seconds and all residual PEG removed.
The phage pellet was resuspended in 100μl TE pH 8.0 and extracted with an equal
volume of phenol to remove the phage coats. The aqueous layer was removed and
the DNA ethanol precipitated. The DNA was resuspended in 30μl TE pH 8.0 and
5μl used directly for sequencing as described in section 2.10.

20% PEG: 20g polyethylene glycol (8000mw)
14.6g sodium chloride
distilled water to volume of 100ml

2.12. RNA preparation and manipulation

2.12.1. Isolation of RNA

RNA was isolated using RNase-free reagents, glassware and plasticware.
10^8 cells from the lymphoblastoid cells lines LCL 721 and LCL 721.174 were pelleted by spinning at 2000 rpm for 5 mins in a Heraeus minifuge T. The pellet volume was estimated and 5 volumes of guanidium isothiocyanate were added. The pellet was dispersed by vortexing. 1g of CsCl was added to each 2.5ml of homogenate. The homogenate was then layered onto a 1.3ml cushion of 5.7M CsCl, 0.1M EDTA pH 7.5 in a polyallomer ultracentrifuge tube. Ultracentrifugation was carried out at 40000 rpm for 12 hrs at 20°C in a SW60 rotor. After centrifugation, the supernatant was discarded and the base of the tube containing the RNA pellet cut off with a scalpel. Any remaining supernatant in the base of the tube was carefully removed with a pipette and the pellet resuspended in 500μl 10mM Tris.Cl pH 7.4, 5mM EDTA, 1% SDS. This was extracted once with a 4:1 mixture of chloroform and 1-butanol. The aqueous phase was transferred to a fresh tube and the organic phase re-extracted with an equal volume of Tris solution, described above. The two aqueous phases were combined and the RNA precipitated by adding 0.1 volumes of 3M NaAc pH 5.2 and 2 volumes of absolute ethanol, storing at -20°C overnight. The RNA was pelleted by microfuging at 4°C for 10 mins and the pellet washed in 70% ethanol. The RNA was resuspended in 200μl distilled water and stored at -20°C.

Guanidium isothiocyanate:
- 4M guanidinium isothiocyanate
- 5mM sodium citrate pH 7.0
- 0.1M β-mercaptoethanol
- 0.5% sarkosyl

2.12.2. Isolation of poly (A)^+ RNA

Polyadenylated RNA was isolated from total RNA by oligo-dT cellulose chromatography using Fast-Track reagents (Invitrogen). Each total RNA sample was made 0.5M with NaCl and incubated with pre-equilibrated oligo-dT cellulose for 30 mins at room temperature. The mixture was transferred to a spin-column and washed three times with high-salt binding buffer to remove non-Polyadenylated RNA. Poly (A)^+ RNA was then eluted from the column in a low salt buffer and mixed with 0.15 volumes 2M NaAc pH 5.2 and 2 volumes absolute ethanol. The mixture was frozen on dry ice to precipitate the RNA and then microfuged for 15 mins. The supernatant was removed and the pellet resuspended in low-salt buffer before storage at -20°C.
2.12.3. Electrophoresis of RNA

RNA was resolved in formaldehyde agarose gels. 2.2g agarose was melted in 144.2mls distilled water and the gel allowed to cool to 60°C. 20ml 5X MOPS buffer and 35.8ml formaldehyde were added to the gel before pouring. RNA samples were prepared as follows. 5.5µl (10-15µg total RNA or 3µg poly (A)+ RNA) were mixed with 1µl 5X MOPS buffer, 3.5µl formaldehyde and 10µl formamide. The samples were incubated for 15 mins at 65°C and cooled on ice. 1µl loading dye (50% glycerol, 0.1% bromophenol blue) was added and the samples loaded on the gel. Electrophoresis was carried out in 1X MOPS buffer at 30-40mA overnight using 18S and 28S rRNAs as molecular weight markers. After electrophoresis, northern gels were stained for 30 mins in 10µg/ml ethidium bromide. Gels were rinsed in 20X SSC and blotted onto Hybond N+ filters overnight in 20X SSC. RNA was fixed to the filters by baking at 80°C for 2 hrs under vacuum.

\[5X \text{MOPS:} \quad 0.1M \text{3-(N-morpholino)propane sulfonic acid}\]
\[40mM \text{sodium acetate}\]
\[5mM \text{EDTA pH 8.0}\]

2.12.4. Hybridisation of RNA blots

Northern blots were prehybridised at 42°C for 1 hour in a buffer containing 50% deionised formamide, 10% dextran sulphate, 1M NaCl, 1% SDS. Hybridisation was carried out at 42°C in fresh buffer with the addition of 10⁶ cpm probe per ml and 100µg/ml sheared, sonicated salmon sperm DNA that had been boiled together for 5 mins.

After hybridisation of a human cDNA probe to a northern blot, the filters were washed down to a stringency of 0.1X SSC, 0.5% SDS at 65°C. After genomic hybridisations the final washing stringency was usually 2X SSC at 50°C. Filters were exposed to Kodak XAR-5 autoradiography film at -80°C between intensifying screens.
2.13. Exon amplification

Exon amplification was carried out by Dr. M. North (Genome Analysis Laboratory, ICRF). Cosmid fragments of ~4 kb (BamHI, BglII fragments or Sau3A partial digests) were introduced into a mammalian expression vector, pSPL1. The cloning site is contained within an intron which is flanked by 3' and 5' splice sites from the HIV *tat* gene. COS-7 cells were transfected with these constructs, and high levels of transcription are driven by the expression vector SV40 early promoter. During *in vivo* processing, splice sites of any exon contained within an inserted genomic fragment are paired with the *tat* splice sites so that intronic DNA is excised and the exon is retained in the mature RNA. Reverse transcription followed by PCR was used to amplify such 'trapped' exons. Dr. M. North carried out the exon amplification, producing exons which I isolated.

2.14. PCR amplification of exon trapped products

The entire reverse transcription reaction was subjected to PCR amplification in a Hybaid thermocycler using the oligonucleotide primers SD2 (\[G T G A A C T G C A C T G T G A C A A G C\]) and SA2 (\[ATCTCAGTGGATTTGTGAGC\]). For each reaction 20µmol SD2, 20µmol SA2, 1µl 2.5mM dNTPs, 0.01U Taq polymerase, 3µl MgCl2, 5µl 10X PCR buffer (Promega) were added in a final volume of 50µl. Cycles were as follows: 1 min at 94°C, 2 mins at 56°C and 3 mins at 72°C, 30 cycles. Products were run out on 2% agarose gels in 1X TAE and visualised by staining with 1µg/ml ethidium bromide.

A secondary (nested) PCR reaction was carried out with 1µl from the first reaction using the primer pairs SA1 (\[CCCGTCGACGTGGGTCCCTCAGGATGG\]) and SD1 (\[CCGATCCGCGAGACCTCCTCAAGGC\]) which flank the vector splice junctions and contain Sall and BamHI cloning sites respectively. For each reaction 20µmol SA1, 20µmol SD1, 1µl 2.5mM dNTPs, 400U Taq polymerase, 3µl MgCl2, 5µl 10X PCR buffer (Promega) were added in a final volume of 50µl. Cycles were as follows: 10 mins at 94°C, 1 cycle; 1 min at 94°C, 1 min at 60°C, 3 mins at 72°C, 30 cycles and 10 mins at 72°C, 1 cycle.
Amplified products were separated on 1.8% agarose gels and visualised by staining with ethidium bromide. The PCR products were purified from the gel either by excision and centrifugation through Whatman 3MM paper or by electroelution (described in section 2.11.1). The purified products were either labelled and used as probes directly onto cDNA libraries as described in section 2.3 or were subcloned and sequenced.

2.15. Subcloning of PCR products

Subcloning of PCR products was carried out using the CloneAmp system (Gibco, BRL), following the manufacturers instructions. Briefly, one of the PCR primers is tailed with the sequence 5'....... CAU CAU CAU CAU .......3' and the other one with the sequence 5'....... CUA CUA CUA CUA .......3'.

10-50ng (1μl) of the PCR product was mixed with 2μl (50ng) of pAMP vector, 1μl of uracil DNA glycosylase (UDG, 1U/μl) and 15μl 1X annealing buffer (20mM Tris.Cl pH 8.0, 50mM KCl, 1.5mM MgCl₂) in a total volume of 20μl. The reaction was incubated for 30 mins at 37°C. After amplification, the PCR products are produced with the dUMP-containing sequence at their 5' termini. Treatment with UDG renders dUMP residues abasic and unable to base-pair, resulting in 3' protruding termini. The denatured ends of the PCR products then anneal to the pAMP vector, which has been designed with 12-base 3' complementary termini. After incubation the reaction was placed on ice and 1μl used for electroporation into *E. coli* strain MC1061/P3. In all cases 95% of the recombinant clones contained plasmid.
CHAPTER 3: COMPLETION OF AN OVERLAPPING COSMID MAP OF THE HUMAN MHC CLASS II REGION

3.1. Introduction

Although the entire class II region of the human MHC has been cloned in YACs (Ragoussis et al., 1991) and most of the region has been cloned in cosmids (Blanck and Strominger, 1990; Blanck and Strominger, 1988), some small areas of the class II region have resisted cloning in cosmid vectors. Since the average insert size of cosmid clones is 40 kb compared to the average insert size of 650 kb in class II YACs (Ragoussis et al., 1991), cosmid clones provide a more useful source for investigating a region of the genome in detail as they are easier to manipulate and less likely to rearrange. At the start of this project two areas of the class II region had yet to be cloned in cosmids: 1) the region between the genes DMB and LMP2 and 2) part of the region between the genes DPA1 and DNA (figure 3.1).

![Figure 3.1](image_url)

**Figure 3.1.** Map of the class II region of the human MHC showing regions uncloned in cosmids. At the start of this work, a 15-20 kb gap existed in the cosmid map between the genes DPA1 and DNA (1). A 40 kb gap in the cosmid map existed between the genes DMB and LMP2 (2).
As part of the continuing effort to link, characterise and sequence the entire class II segment of the MHC it was decided to clone these two remaining regions in cosmid vectors. By screening these segments for potential coding regions, their cloning will provide a further step towards the ultimate aim of producing a complete picture of the genomic organisation of the class II region. This chapter will treat the two segments separately, starting with the region immediately telomeric of the DMB gene.

3.1.1. Cloning the DMB to LMP2 gap

This region was of interest as it was shown, by physical mapping in conjunction with PFGE, to contain a NotI site (Ragoussis et al., 1991) which is characteristic of a CpG-rich sequence. The mammalian genome can be divided into two fractions with respect to DNA methylation. In the major fraction (98% of the total) the dinucleotide CpG occurs on average every 50 to 100 bp and is heavily methylated. In the minor fraction (2%), CpG occurs approximately every 10 bp and is non-methylated. Although the minor fraction is only a small proportion of the total it is distributed throughout the genome in 45 000 short regions of about 1 kb, known as CpG islands (Craig and Bickmore, 1994). The function of these regions is not yet known, but they are found to colocalise with the 5' ends of expressed sequences (Bird, 1987). In humans, about 60% of genes are associated with CpG islands, including all housekeeping genes so far analysed and about 40% of tissue-restricted genes (Larsen et al., 1992). In most cases, the island contains the promoter and one or more exons of its associated gene. There are three NotI sites in the class II region of the MHC, two of which flank TAP1 and the third, located centromeric of the LMP2 gene, which was uncloned in cosmids at the start of this work. Since the two NotI sites flanking TAP1 are associated with coding sequences, it seemed likely that this third NotI site would provide a strong starting point in the search for novel genes.

The DMB to LMP2 gap was implicated at the start of this project as a candidate region for a gene(s) involved in the class II antigen processing pathway (figure 3.2). Several mutant cell lines were developed with deletions in the MHC and concomitant defects in antigen presentation through class II that produced evidence for the existence of such a gene(s). One class of mutants was isolated from a HLA-DR3-expressing B lymphoblastoid cell line (B-LCL), 8.1.6, which has a hemizygous MHC deletion, by mutagenising clones with ethyl methane sulphonate and immunoselecting with an anti-HLA-DR3 monoclonal antibody 16.23 (Mellins et al., 1990). The mutants are unable to stimulate specific T cell clones when exposed to
Figure 3.2. Class II region of the human MHC showing deletion mutants spanning the area. Deletions of DNA in the indicated human B cell mutants are shown by solid lines. Class II molecules encoded by indigenous genes are normal in 721.82 and abnormal in 721.174 and 5.2.4. This suggests that DNA present in 721.82 and absent in the other mutants is needed for normal class II-mediated antigen presentation. The DNA segment responsible is shown by the dashed line (from DeMars et. al. 1985). Later work showed that the deletion in 5.2.4 extended into the DMB gene and DMA and DMB were identified at the genes affecting the class II antigen processing pathway (Fling et al., 1994; Morris et al., 1994).
intact protein antigens. However, their class II molecules effectively present exogenously supplied peptides. Thus, the mutants appear unable to form intracellular MHC class II-peptide complexes from intact protein antigens. The class II structural genes and the surface abundance of class II molecules in these mutants are normal. However, the mutants demonstrate a subtle change in class II molecule conformation. Their HLA-DR3 molecules have lost expression of the epitope recognised by antibody 16.23, and class II dimers extracted from the mutants dissociate into monomers on SDS-PAGE, whereas dimers extracted from the progenitor cells under the same conditions do not. It was proposed that the conformational changes were related to underoccupancy of cell surface molecules resulting from defective peptide loading during class II molecule biosynthesis.

Further evidence for the presence of this gene was provided by a mutant B-LCL, 5.2.4, (Ceman et al., 1992) which was selected with an anti-DR monoclonal antibody. 5.2.4 has a homozygous deletion within the class II region of the MHC (figure 3.2). Although mutant 5.2.4 expresses DP molecules on the cell surface, it is defective in the presentation of DP-restricted protein antigens to antigen-specific T cells. However, it presents DP-restricted antigenic peptides efficiently. 5.2.4 manifests the class II presentation-defective phenotype and is non-complementary with the 16.23 selected mutants for the class II presentation defect. Both sets of mutants present peptides more efficiently than the progenitor cells suggesting that class II molecules on the cell surface of these mutants are underoccupied with self peptides. The fact that 5.2.4 has the presentation defect, characterising the 16.23 mutants, along with a homozygous MHC deletion further implicates the deleted region as the site of the involved gene(s). This DNA segment is highlighted by the dashed line in figure 3.2.

The third piece of evidence for the existence of this gene was provided by mutants from the LCL721 B cell line (DeMars et al., 1985). Mutant 721.174 was derived by γ irradiation and immunoselection against cell surface class II molecules from LCL721.45 in which a complete MHC class II haplotype had been deleted. In mutant 721.174 there is a reduction of HLA-A2 and complete loss of HLA-B5 at the cell surface. Exposure of 721.174 cells to peptide restores expression of HLA-A2 and -B5 showing that 5.2.4 and 721.174 cells share a similar phenotype, due to deletion of a gene(s) in the MHC. A further mutant, 721.82, isolated in the same way as 721.174, did not show the class II presentation-defective phenotype. Combining these studies led to the supposition that DNA present in mutants 721.82 and absent in mutants 721.174 and 5.2.4 is needed for the production of normally conformed DR molecules,
as defined by the 16.23 epitope. This led to the conclusion that the gene(s) affecting class II conformation were located between \textit{DMB} and \textit{DQB1} (figure 3.2) and was unknown. Since the \textit{DMB} to \textit{LMP2} gap was contained within this region it seemed likely that the gene(s) involved resided in the gap.

To clone the \textit{DMB-LMP2} region two cosmid libraries were utilised, one from a YAC clone spanning the region and the other from flow-sorted chromosome 6 DNA. A P1 bacteriophage library was screened as an additional approach.

3.2. Results and discussion

3.2.1. Estimating the size of the region uncloned in cosmids between the genes \textit{DMB} and \textit{LMP2}

To estimate the distance between the \textit{DMB} and \textit{LMP2} genes the YAC 11.2 (Ragoussis et al., 1991) was cut with a variety of restriction enzymes and the DNA separated by PFGE. A CHEF apparatus was used under the following conditions: 200V, 2 to 20 second pulse for 16 hrs at 12°C. This YAC is approximately 450 kb in size and covers the subregion between the \textit{DMB} and \textit{DOB} genes. After blotting, the DNA was hybridised with probes flanking both sides of the gap and the results compared with published maps (Blanck and Strominger, 1988; Blanck and Strominger, 1990). A typical hybridisation with the complete \textit{DMB} cDNA is shown in figure 3.3. A large NotI fragment (approximately 90 kb) was observed hybridising to the \textit{DMB} probe which corresponded to the distance between the centromeric end of the YAC and the NotI site in the gap. Since the distance between the centromeric end of the YAC 11.2 and the telomeric end of the cosmid HA14 was about 55 kb and the distance between the NotI site in the gap and the centromeric end of the cosmid U15 was about 5 kb (Ragoussis et al., 1991) then the length of the gap was [90+5 kb]-55 kb. Hence, the gap between the cosmids HA14 and U15 was estimated to be approximately 40 kb in size. \begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure3.3.png}
\caption{Hybridisation with complete \textit{DMB} cDNA.}
\end{figure}

3.2.2. Construction of a cosmid library from the YAC 11.2

Initially, an enriched cosmid library was constructed from the YAC 11.2 which is 450 kb in size and encompasses the region between the genes \textit{DMB} and \textit{DOB} in the class II region of the human MHC. YAC DNA was partially digested with the restriction enzyme MbolI. The DNA was kept within agarose blocks, each block containing 40\mu g
Map of the MHC class II region showing position of YAC 11.2 with respect to the cosimds HA14, U15 and U10.
Figure 3.3. Hybridisation of complete DMB cDNA onto a Southern blot of digested DNA from the YAC 11.2, which covers the region from the genes DMB to DOB in the class II region of the MHC (Ragoussis et al., 1991). YAC DNA was digested with KpnI (A), NotI (B), ClaI (C), KpnI/NotI (D), ClaI/NotI (E), KpnI/ClaI (F) and the DNA separated by PFGE. Final washing stringency was 65°C, 0.2X SSC. A 90kb NotI restriction fragment from YAC 11.2 hybridised to the DMB cDNA which corresponded to the distance between the centromeric end of the YAC and the NotI site in the gap. The distance between the centromeric end of YAC 11.2 and the telomeric end of cosmid HA14 is known to be 55 kb and the distance between the NotI site in the gap and the centromeric end of cosmid U15 is 5 kb (Blanck and Strominger, 1990; Blanck and Strominger 1988; Ragoussis et al., 1991). The length of the gap between the cosmids HA14 and U15 is [90 + 5 kb] - 55 kb; ie. ~ 40 kb.
DNA. One quarter of a YAC block was digested with a range of MboI concentrations varying from 0 to 4 units of enzyme in a total volume of 200μl. After 30 mins the reactions were terminated by the addition of 4μl 0.5M EDTA. The digested YAC blocks were loaded onto a 0.4% agarose gel and run overnight at 4°C, 27V using λ DNA digested with HindIII as a molecular weight marker. YAC DNA digested with 0.03 and 0.06 units of MboI gave fragments in the size range of 35-40 kb which is the optimal insert size for a cosmid library. 4 X 0.25 YAC blocks were digested with 0.03 U and 0.06 U MboI using the conditions described above. Each block was washed in 500μl TE pH 8.0 then placed in 100μl TE pH 8.0 containing 0.01 U calf intestinal alkaline phosphatase (CIAP). This dephosphorylates the DNA minimising recircularisation of the YAC DNA during ligation. The reactions were incubated for 30 mins on ice to allow enzyme diffusion into the block. They were then placed at 37°C for 45 mins to allow the dephosphorylation reaction to occur. The blocks digested with the same units of MboI were combined and incubated at 70°C for 20 mins to melt the agarose. An appropriate amount of the resulting digested-YAC-containing solution was aliquoted and used for ligation reactions, described below. The remaining solution was cooled to re-set the agarose and stored at 4°C with a layer of TE pH 8.0 on top of the agarose block.

Preparation of the cosmid vector DNA

The vector SuperCos 1 was used in the preparation of this cosmid library. This vector was chosen as it was designed for cosmid cloning using very small amounts of genomic DNA. Efficient DNA cloning is achieved as the vector contains multiple cos sites. SuperCos 1 contains genes for the amplification and expression of cosmid clones in eukaryotic cells. In addition, most genomic inserts can be excised in a single NotI restriction fragment. The vector was prepared as outlined below.

20μg of the SuperCos 1 vector (1 μg/μl) was digested with 18 U/μg of the restriction enzyme XbaI in a total volume of 200μl, in standard buffer conditions for 1 hr at 37°C. The DNA was extracted once with phenol and once with PCIA. The aqueous phase was adjusted to 0.3M sodium acetate (NaAc) pH 5.5 and precipitated with 2 volumes of absolute ethanol. The DNA pellet was washed with 70% ethanol, vacuum dried and resuspended in distilled water at a concentration of 1 μg/μl. 500ng
of cut vector was run on a 0.8% agarose gel; a single linear 7.6 kb SuperCos 1 cosmid band confirming complete digestion.

The digested cosmid DNA was incubated with 0.65 U CIAP/μg DNA at 37°C for 1 hr to dephosphorylate the DNA. EDTA pH 8.0 was added to a concentration of 15 mM and the mixture incubated at 68°C for 10 mins to terminate the reaction. The DNA was extracted once with phenol and once with PCIA. The aqueous phase was adjusted to 0.4M with ammonium acetate (NH₄OAc), precipitated with 2 volumes of absolute ethanol and washed with 70% ethanol. The DNA pellet was lyophilised and resuspended in TE pH 8.0 at a final concentration of 1 μg/μl.

To produce both vector arms necessary for ligation, the DNA was further digested with the restriction enzyme BamHI. 20μg of DNA was digested with 100 U of BamHI in standard buffer conditions for 1 hr at 37°C in a total volume of 200 μl. The DNA was extracted once with PCIA and once with CIA. The aqueous phase was adjusted to 0.3M NaAc pH 5.2 and ethanol precipitated. The DNA pellet was washed with 70% ethanol, vacuum dried and resuspended in TE pH 8.0 at a concentration of 1 μg/μl. 500ng of vector was run on a 0.8% agarose gel and visualised with ethidium bromide. Two bands of 1.1 kb and 6.5 kb showed complete digestion.

Ligation of vector and genomic DNA

Four ligation reactions were set up containing 1 μg digested, dephosphorylated SuperCos 1 vector, 2 μl 10X ligation buffer, and varying concentrations (1-3μg) MboI-digested YAC DNA in a final volume of 20 μl. 1 μl high concentration (2 X 10⁶ U/μl) T4 DNA ligase was then added and the reactions incubated at 14°C overnight.

Packaging the cosmid library

4 μl of each ligation reaction was added to freeze/thaw extract on ice (Gigapack II packaging kit; Stratagene). 15 μl sonic extract was added to this and mixed by stirring with a pipette. The reactions were incubated at room temperature for 90 mins. 500 μl phage dilution buffer and 20 μl chloroform were then added. The reactions were spun for 3-5 seconds in an Eppendorf centrifuge to sediment the debris. The supernatants were stored at 4°C ready for titration.
Titering the cosmid library

A 50ml culture of LB containing 0.2% maltose and 10mM magnesium sulphate (MgSO\textsubscript{4}) was inoculated with a single colony of \textit{E. coli} NM554 cells and grown, shaking, overnight at 37°C. The culture was spun for 10 mins, 2000 rpm at 4°C in a Heraeus minifuge T centrifuge. The supernatant was decanted and the pellet resuspended in 12.5ml MgSO\textsubscript{4}. Immediately prior to use, the cells were diluted to OD\textsubscript{600} in 10mM MgSO\textsubscript{4}. The packaged DNA was diluted 1:10 and 1:50. 25 µl of each dilution was mixed with 25 µl of diluted NM554 cells and allowed to stand at room temperature for 30 mins. 200µl LB was then added to each sample and the tubes incubated for 1 hr at 37°C with occasional shaking. The tubes were spun for 1 min and the pellet resuspended in 50µl fresh LB. The cells were plated out on L-agar plates containing 50µg/ml ampicillin and incubated overnight at 37°C. The following day the library titer was calculated and an appropriate quantity of the remaining ligations packaged and plated out. The cosmid library was plated onto Hybond N+ filters overlaid on L-agar Nunc plates (245 X 245mm) producing a total of 9000 colonies. These filters were replicated as described in section 2.2.1.

3.2.3. Flow-sorted chromosome 6 cosmid library

This library was constructed by Dr. Dean Nizetic as part of a collaboration with the Human Immunogenetics Laboratory. A flow-sorted chromosome 6 was used for the construction of the library. DNA from a lymphoblastoid cell line of normal karyotype, RPETO1, was digested, ligated into the Lorist vector and transfected into DH5α MCR (BRL). The library was robotically arrayed (Nizetic et al., 1991) onto high density grids and represents approximately 4 fold coverage of chromosome 6. It is part of the reference library database from the ICRF (Nizetic et al., 1991).

3.2.4. P1 genomic library

The total P1 library was prepared by Fiona Francis (Genome Analysis Laboratory, ICRF; Francis et al., 1994) as described in section 2.2.3. The library was robotically arrayed onto high density grids, each containing 20736 clones (Nizetic et al., 1991) and represents a 1.2 fold genome coverage.
3.2.5. Isolation of cosmid clones from the DMB to LMP2 region

The cosmid library isolated from YAC 11.2 was screened with various probes in order to isolate new cosmids spanning the area between the DMB and LMP2 genes. Two cosmids flanking the gap were available: HA14 and U15, both of which are part of cosmid libraries prepared by Blanck and Strominger (1988 and 1990). HA14 contains the DMA and -B genes and U15 contains the genes LMP2, TAP1, LMP7 and TAP2 (figure 3.4). Due to the high density of repetitive DNA on the ends of these cosmids it was difficult to isolate unique probes at their extreme ends (Beck et al., 1992b). The complete DMB cDNA was therefore used as a probe from the centromeric side of the gap and, since the genomic sequence of U15 was available, two oligonucleotides were used to produce a 2 kb, single-copy probe by PCR from the telomeric side of the gap. This latter probe was called GE1.

The DMB cDNA and GE1 PCR probe were used to screen the cosmid library prepared from YAC 11.2 resulting in the isolation of 10 new cosmids. Four cosmids isolated with the DMB cDNA were shorter than cosmid HA14. These cosmids contained no novel DNA so were not further analysed. Six cosmids were isolated with the probe GE1. These cosmids were digested with the restriction enzymes EcoRI, BamHI, NotI, ClaI, Stul, HindIII and XhoI, Southern blotted, and hybridised with cDNAs corresponding to the genes LMP2, TAP1, LMP7 and TAP2. Final washing stringency was 65°C, 0.2X SSC. Comparison of the restriction digest and hybridisation patterns of the new cosmids with that of cosmid U15 enabled rough mapping of the cosmids. Five of the six cosmids encoded the genes LMP2, TAP1, LMP7 and TAP2 and gave virtually identical restriction digest patterns to that of U15. These cosmids did not contain the NotI site residing in the gap. However, one cosmid, called N3, contained the NotI site predicted to reside in the gap by physical mapping (Ragoussis et al., 1991) and extended the cloned region by 4 kb (figure 3.4).

At the same time filters from the flow-sorted chromosome 6 cosmid library were screened with the DMB cDNA and GE1 probe. Three clones were isolated using the DMB cDNA. Hybridisation studies showed that two of these encoded the genes DMB and DMA (centromeric of DMB). The third clone, A12, was of interest since it showed a positive hybridisation signal with DMB yet a negative hybridisation signal with DMA indicating that it would extend furthest into the gap. From restriction enzyme mapping and comparison with cosmid HA14, cosmid A12 extended approximately 21 kb telomeric of HA14. On the other side of the gap eight cosmids...
Figure 3.4. Map of the MHC class II region between the genes *DMB* and *LMP2*. The map shows the position of cosmids and a P1 clone isolated, in this study, from the cosmid library constructed from YAC 11.2 (N3), the flow sorted chromosome 6 cosmid library (A12/A15) or from the P1 library (DP2). These are shown relative to the previously described cosmids HA14, U15 and U10 (Blanck and Strominger, 1988; Blanck and Strominger 1990). (+), a large number of StuI sites at the telomeric end of the map.
were isolated with the probe GE1 (figure 3.5). Seven of these cosmids were positive with the \textit{LMP2} cDNA probe and one was negative. This last cosmid, A15, contained the NotI site and extended 30 kb into the gap. Initial restriction enzyme mapping and hybridisation studies showed an overlap between the two cosmids A12 and A15 (figure 3.6). Comparison of the restriction enzyme digest patterns of cosmids A15 and A12 using the enzymes NotI, XhoI, SmaI and ClaI established a 6 kb overlap between the two cosmids. These results were confirmed by using fluorescence in situ hybridisation (FISH) on free chromatin fibers with both cosmids detected in different colours. The FISH was carried out by Dr. G. Senger in the Human Cytogenetics Laboratory at the ICRF.

The cosmid clones A12, A15 and N3 were used for FISH on free chromatin released from peripheral blood lymphocytes of a healthy donor (Senger et al., 1993). The fluorescence signals on released chromatin obtained with each cosmid were visible as either red or green dotted lines that were partially overlapping. In figure 3.7a, cosmid A12 is visualised in red and cosmid A15 in green, showing a 6 kb overlap. The overlaps between cosmids A12/HA14 and A15/U15 were demonstrated by the same technique showing a 20 kb overlap between A12 and HA14 and a 10 kb overlap between A15 and U15 (figures 3.7b and 3.7c).

### 3.2.6. Isolation of a P1 clone

In parallel with these approaches a P1 phage library was screened with the GE1 probe. Due to the large insert size of P1 clones (about 100 kb) compared to the average cosmid insert size of 40 kb they facilitate the cloning of areas which have, in the past, proved problematic due to large stretches of repetitive sequence, for example. Such is the case with the \textit{DMB} to \textit{LMP2} region. Since the repetitive sequence is spread over a greater distance in P1 clones it is less likely to cause problems with recombination, producing more stable clones. Thus the cloning of such areas should be aided by the use of P1s.

One clone (DP2) of approximately 100 kb was isolated from the P1 library using the GE1 probe. DP2 was positive with both the \textit{DMB} cDNA and GE1 probe thus linking both sides of the gap (figure 3.4). Moreover, DP2 was negative with cDNAs corresponding to the genes \textit{DMA} and \textit{DOB} but positive with the \textit{TAP2} cDNA. Therefore, the centromeric end of the P1 clone was located between the \textit{DMA} and...
Figure 3.5. Hybridisation of 2 kb 1CR probe, GE1, onto flow-sorted chromosome 6 cosmid library showing 8 positive clones. The filter is spotted with 20 730 clones in groups of 9 (Nizetic et al., 1991). The x and y coordinates of positive clones were identified and used to identify frozen glycerol stocks of the corresponding cosmids. These were plated out onto L-agar plates containing 30μg/μl kanamycin, grown overnight at 37°C and analysed by mini-preps and restriction enzyme digestion.
Figure 3.6. Hybridisation of Southern blot of EcoRI digested DNA from the cosmids A12 and A15 with an 8 kb Xho I fragment located at the centromeric end of cosmid A15. A common band of 4 kb was detected in both cosmids indicating that they overlap each other.
Figure 3.7. Two colour FISH on free chromatin released from peripheral blood lymphocytes of a healthy donor (Senger et al., 1993). Cells were harvested with 75mM KCl as hypotonic solution and methanol/acetic acid (3:1) as fixative. Cells, spread on slides, were rehydrated in PBS and chromatin was released from the nuclei by using 50mM NaOH/30% ethanol. DNA probes were labelled with biotin-dATP or with digoxygenin-11-dUTP by nick translation. In situ hybridisation and probe detection of several pairs of differentially labelled probes was then performed with the digoxygenin-labelled probes being detected with monoclonal mouse antidigoxygenin and FITC-conjugated sheep-anti-mouse IgG. Biotinylated probes were detected with avidin Texas red. The probe signals were analysed using a Zeiss Axioscope fluorescence microscope equipped with a cooled CCD camera. Digital grey-scale images were pseudocoloured, merged and photographed from the computer screen.

The FISH was carried out by Dr. G. Senger as part of a collaboration between the Human Immunogenetics and Human Cytogenetics Laboratories.

The hybridisation signals of all probe pairs are either shown in red or green. (A) Cosmids A12 (red) and A15 (green) showing a 6 kb overlap, (B/D) cosmids HA14 (green) and A12 (red) showing a 20 kb overlap, (C) cosmids A15 (green) and U15 (red) showing a 10 kb overlap. The overlapping regions are visible where both red and green signals are present. The scale bar indicates 10µm.
DMB genes and the telomeric end between LMP7 and DOB genes. This result agreed with the size of the clone, restriction mapping and FISH analysis (figure 3.8).

3.2.7. Cloning of the gap between DPAI and DNA

After the cloning of the MHC class II region from the genes DMB to LMP2, one other gap remained in the class II cosmid map: between the genes DPAI and DNA (figure 3.1). Mapping of the DPAI and DNA genes onto the existing cosmid map in conjunction with PFGE (Ragoussis et al., 1991; Blanck and Strominger, 1990) estimated the gap between the cloned sequences (centromeric end of the cosmid O19A on the distal side and telomeric end of cosmid MANN 3.6, containing DPAI, on the distal side) to be in the region of 10-15 kb. Cosmids O19A and MANN 3.6 have been described previously (Blanck and Strominger, 1990; Trowsdale et al., 1985). From analysis of the centromeric sequence of cosmid O19A (S. Beck, personal communication) it could be seen that the sequence was highly repetitive which could have caused problems in cloning. Initially, it was decided to try to clone the gap using the flow-sorted chromosome 6 cosmid library alone and reverting to the P1 library if this approach proved unsuccessful.

Since the sequence at the centromeric end of O19A was highly repetitive it was decided to use a 10 kb XhoI restriction fragment from the middle of the cosmid O19A as a probe onto the chromosome 6 cosmid library filters. The XhoI fragment was sonicated for one minute before radioactive labelling to high specific activity using the method of Feinberg and Vogelstein (1984). The fragment was competed with human placental DNA to remove repetitive sequences before hybridisation to the chromosome 6 cosmid library filters. Final washing stringency was 65°C, 0.2X SSC. Fifteen cosmids were isolated, and hybridised with the complete DPAI cDNA and the 10 kb XhoI fragment from cosmid O19A originally used to isolate the clones. Four cosmids (A1, A2, A3 and A5) showed positive hybridisation signals with DPAI (figure 3.9) and also hybridised with the 10 kb XhoI fragment from O19A. These cosmids (A1, A2, A3 and A5) hence bridged the gap between DPAI and the centromeric end of the cosmid O19A and completed the class II cosmid map. End sequencing of two of the four cosmids (A1 and A3) was carried out which allowed accurate positioning of the clones on the existing cosmid map. EcoRI, BamHI and Mspl restriction digests of the four cosmids produced similar band patterns indicating that the cosmids had not rearranged. Since the cosmid A1 was slightly longer than cosmids A2, A3 and A5 further mapping was carried out on this clone.
Figure 3.8. FISH mapping of the P1 clone, DP2, using the methods described in figure 3.7. DP2 is visualised in red, the cosmid HA14 is shown in yellow and cosmid U10 is visualised in green. Cosmids HA14 and U10 are part of the cosmid libraries previously described by Blanck and Strominger (1988 and 1990) and encode the genes DMA and DMB and the LMP/TAP gene cluster, respectively. Both HA14 and U10 completely overlap the P1 clone DP2 indicating that DP2 spans the MHC class II region from the genes DMA to TAP2. The FISH was carried out by Dr. G. Senger as part of a collaboration with the Human Cytogenetics Laboratory.
Figure 3.9. Southern blot of EcoRI-digested DNA (in duplicate) from six cosmids isolated from the flow-sorted chromosome 6 cosmid library using a 10 kb XhoI fragment from the cosmid O19A (previously described by Blanck and Strominger, 1990). The complete DPAl cDNA was hybridised to these cosmids at a final washing stringency of 65°C, 0.2X SSC. Positive signals were observed on four of the six cosmids (A1, A2, A3 and A5) showing that they contain the gene DPAl and thus span the region from the centromeric end of cosmid O19A to the gene DPAl (cloned in the cosmid MANN 3.6, Trowsdale et al., 1985) resulting in cloning of the remaining gap in the class II cosmid map.
Restriction digests of cosmid A1 in comparison with cosmid O19A and sequence present in the MHC database (MHCDB) allowed the construction of a restriction map for A1 (figure 3.10).

3.3. Conclusions

To maximise the chance of cloning the DMB to LMP2 region a saturation approach was initiated by utilising two different cosmid libraries and a P1 library. It is not known why certain regions of the human genome prove so difficult to clone in cosmids but one explanation for the difficulty in cloning this particular gap could be the presence of multiple repeat sequences such as Alu or LINES (Kariya et al., 1987; Jurka, 1989). Three new cosmids A12, A15 and N3 were isolated and mapped to the uncloned region. A12 and A15 were overlapping cosmids. The fidelity of these new clones with the YAC clone 11.2 and the existing genomic maps was established with physical mapping and the order of the cosmids was confirmed by FISH.

DP2 is the first P1 clone isolated from the class II region of the MHC and provides a clone intermediate in size between cosmids and YACs. The fidelity of DP2 was confirmed by comparison of the restriction pattern of the P1 clone with the new cosmids, existing genomic maps and FISH.

These data show that cosmid libraries rarely contain clones spanning this particular region. No such problems were found with the YAC or P1 libraries. The reason for the differences in the stability between cosmids and other systems is not known although one explanation could be the respective sizes of the clones. In the P1 or YAC clones the repeat sequences may be stabilised by the presence of long stretches of unique DNA not present in the smaller clones. Another explanation may be the low copy number of the P1/YAC clones in comparison to the cosmids.

In the mouse, the region between the DMB and DOB genes (Mb and Ob) has already been cloned in cosmids (Steinmetz et al., 1986) and appears to be approximately half the size of the human region (Cho et al., 1991). By contrast, the size of the gene cluster containing LMP2, TAP1, LMP7 and TAP2 is conserved between both species (Hanson and Trowsdale, 1991). Analysis of sequence data between DMB and LMP2 revealed the presence of two LINE repeat sequences in the human DNA at either ends of the recently cloned gap (S. Beck, personal communication). It is possible that the intervening sequence is not conserved between mouse and man as recent recombination events in the human MHC have resulted in the insertion of this stretch of DNA. There is no direct proof for this hypothesis and
Figure 3.10. Map of the class II region of the MHC between the genes *DPB1* and *DNA* showing position of the cosmid A1 in relation to the previously described cosmids MANN 3.6 (Trowsdale et al., 1985), O19A and O17A (Blanck and Strominger, 1990).
the origin of inserted DNA will only become apparent once further sequence of the human genome has been attained.

Cloning the gap between the gene DPA1 and the centromeric end of cosmid O19A proved to be an easier task than the cloning of the DMB to LMP2 gap, necessitating the use of only one cosmid library. The probable reasons for this are twofold. First, the DPA1-centromeric end of cosmid O19A gap was only 24 kb in size as opposed to 40 kb and was hence cloned in a single cosmid. Second, repetitive sequence was only present at one end of the DPA1-centromeric end of cosmid O19A region as compared to both sides of the DMB to LMP2 gap. This led to a higher representation of the cosmid clones covering the DPA1-centromeric end of cosmid O19A gap in the chromosome 6 cosmid library as compared to the number of clones covering the DMB to LMP2 gap.

The new cosmid clones (A1, A12 and A15) completed the cosmid map over the class II region of the MHC (figure 3.11). They were used in an ongoing project to complete the genomic sequence over the class II region and to search for new transcribed sequences either by direct screening of cDNA libraries or by using exon amplification. These experiments are discussed in the following chapters.
Figure 3.11. Complete map of the class II region of the MHC showing position of overlapping YACs and cosmids (taken from MHCDB, Newell et al., 1994). Genes encoded in the class II region are shown as black boxes. At the start of this work a 15-20kb gap in the cosmid map existed between the cosmids MANN 3.6 (which encodes the gene $DPA1$) and O19A (which encodes the gene $DNA$). This is now cloned in the cosmid A1. A 40kb gap in the cosmid map existed between the telomeric end of the cosmid HA14 (which encodes the genes $DMA$ and $DMB$) and the centromeric end of the cosmid U15 (which encodes the $LMP/TAP$ gene cluster). This is cloned in the overlapping cosmids A12 and A15.
CHAPTER 4: GENOMIC ORGANISATION AND TRANSCRIPT IDENTIFICATION IN THE CLASS II REGION KE3 TO DMB

This chapter discusses the genomic analysis of the class II region of the MHC, between the genes KE3 and DMB (figure 4.1). Screening the region for new transcripts is described along with mapping the gene for the retinoic acid receptor β chain and two novel exons between the DMA and DMB genes.

**Figure 4.1.** Map of the class II region of the MHC from KE3 to DMB (taken from Campbell and Trousdale, 1993).

4.1. Introduction

4.1.1. Screening cDNA libraries with cosmid inserts to identify novel genes

Once the class II region of the MHC had been completely cloned in cosmids a systematic approach was undertaken to produce a complete picture of the genomic organisation of the region, identifying all expressed transcripts. Three approaches were utilised to this end, the main method involving the screening of cDNA libraries with genomic (cosmid) fragments covering the region of interest. Four cDNA libraries were screened, which were either derived from B or T cells. All the genes expressed in the class II region, apart from one (RING3; Beck et al., 1992a), are involved in the immune system. These genes either encode class II structural products or proteins which are involved in the class I or class II antigen
processing pathways. Hence, it was assumed that any novel transcripts isolated would be highly expressed in cells of the immune system. By these methods we aimed to identify all immune system genes as well as all ubiquitously expressed genes. Those expressed only in highly specific tissues (kidney, for example) would be missed, discussed below. In this approach either cosmid fragments or whole cosmids were radioactively labelled with $\alpha^{32P}$-dCTP and competed with human placental DNA to remove repetitive sequences which would cause high background hybridisation levels. The cosmid probes were then hybridised to sets of cDNA library filters containing a total of 1 million clones. This approach had previously proved highly successful in our laboratory resulting in the identification of the TAP/LMP gene cluster along with the $\text{DMA}$ and $\text{B}$ genes (Powis et al., 1992; Trowsdale et al., 1990; Glynne et al., 1991; Kelly et al., 1991a/b).

4.1.2. Exon amplification of the MHC class II region

To complement the screening approach, exon amplification was also utilised in the search for expressed transcripts (Buckler et al., 1991). This method, based on the selection of functional splice sites in genomic DNA, has been used recently to identify the neurofibromatosis type 2 suppressor gene (Trofatter, 1993) and a copper transporter gene defective in individuals with Menkes disease (Vulpe et al., 1993), amongst others. Exon amplification is based on the in vivo selection for splice sites flanking exon sequences in genomic DNA. The advantage of this technique is that it is independent of tissue expression patterns and can lead to the isolation of rarely expressed transcripts, transcripts with a limited tissue expression pattern and transcripts expressed only in certain stages of development. Mammalian genomic DNA segments are inserted into an intron of the HIV-1 tat gene which is flanked by splice sites and exons of the viral gene, contained within the vector pSPL1. Upon transfection into COS cells, the inserted genomic fragment is transcribed using an SV40 early promoter. When a fragment containing an entire exon with flanking intron sequence is present, the exon is retained in the mature poly A+ cytoplasmic RNA. RNA-based PCR amplification using oligodeoxynucleotide primers specific for the plasmid exons is then used to detect the presence of the exon.
4.1.3. Sequencing the class II region

The third approach undertaken in the project is a sequencing programme in collaboration with Dr. Stephan Beck (DNA Sequencing Laboratory, ICRF). The ultimate aim of this collaboration is to sequence the class II region of the human MHC, providing a complete picture of the genomic organisation of the region. This will provide definitive information concerning gene structures and orientations, repeat sequences and predicted regulatory sequences.

Analysis of the intron/exon organisation of genes provides a measure of their structure, relatedness and in some cases their function and origin. For example, genomic analysis of the region encompassing the \textit{LMP/TAP} gene cluster clearly indicated the presence of two pairs of related genes (Beck et al., 1992b). \textit{LMP2} and \textit{LMP7} share the same number of six exons, one of which (exon 4) is identical in size. Typing of the intron/exon boundary classes revealed that four out of five boundary classes were identical. Boundary classes are determined according to the position of the introns. Class 0, splice site between codons; class 1, splice site after codon position 1; class 2, splice site after codon position 2.

In the case of \textit{TAPI} and \textit{TAP2} their relatedness is more obvious. Both genes share the same number of 11 exons, 8 out of 11 exons have the same size and all 11 intron/exon boundaries are identical in their classes. Based on this data, it was suggested that both the \textit{LMPs} and \textit{TAPs} arose by duplication of an initial transporter and proteasome gene pair followed by subsequent inversion of one of the genes (Beck et al., 1992b). However, recent work by Belich et al., (1994) has suggested that the proteasome genes (\textit{LMP2} and \textit{LMP7}) arose from two genes (MB1 and \textit{delta}) which duplicated and mutated in different locations. Copies of these genes, the \textit{LMPs}, were brought together to form a cluster of genes, with the \textit{TAPs} in the MHC.

The human genome is estimated to contain about 100 000 genes or consist of about 5\% of coding sequence (Bishop, 1974). A large percentage of the non-coding sequence consists of a growing number of repeat sequences with \textit{Alu} sequences predominating (Jurka and Smith, 1988). According to their evolutionary origin, \textit{Alu} repeats can be divided into two classes (J, old and S, new; Jurka and Smith, 1988). The quotient of S:J can be taken as a direct measure for the conservation or age of a sequence. Thus, analysis of non-coding as well as coding sequences can provide information concerning evolution of stretches of sequence.
Many genes within the human MHC are known to be regulated by one or a combination of the following two mechanisms. 1) Regulation by interferon via interaction with a short sequence motif, ISRE (Friedman and Stark, 1985). 2) Regulation by a class II specific promoter (C2) consisting of three short sequence motifs (Sugawara et al., 1991). By screening these regulatory sequences against the genomic sequence it is possible to ascertain which genes are controlled by which regulatory sequence. For example, two ISRE sequences were found to be located close to the LMP/TAP gene cluster, one of which is bidirectional and may control the entire gene cluster (Beck et al., 1992b).

As part of the sequencing collaboration new cosmids, in conjunction with existing cosmids over the class II region, were sonicated and shot-gun cloned into the vector M13mp18. These clones were then sequenced, compiled into contigs and the sequence data sent to the NIH for analysis by the Gene Recognition and Analysis Internet Link program (GRAIL; Uberbacher and Mural, 1991). This program is a computational approach for locating protein-coding portions of genes in anonymous DNA sequence. Sensor algorithms provide an indication of the coding potential of a region of sequence. A neural network is then used to integrate the sensor outputs and to predict the location of coding regions. The values of the sensor signals are calculated according to a number of criteria: 1) Codon usage. Particular amino acids preferentially use particular codons and this bias is used to identify potential coding regions and the preferred reading frame. 2) Periodicity. The 3-periodicity of each of the four bases is examined and compared to the periodic properties of coding DNA. The overall base composition of the test DNA is compared with the known composition for coding and non-coding DNA. 3) Dinucleotide occurrence. This is non-random, with dinucleotides such as AA and TC being common and CG being rare. By asking whether the next nucleotide is of similar or different commonality compared to the previous nucleotide, a sensor value is obtained which represents the difference between a reference value derived from intron DNA and the test DNA. 4) Word frequency. Sequences can be characterised by examining the frequency of occurrence of nucleotide "words" of a given length in the sequence. Introns and exons have different distributions of word occurrence and this difference is used to calculate a sensor value for the test sequence. 5) Repetitive DNA. The test sequence is compared with statistics for several classes of repetitive DNA. This is a negative coding indicator since it reflects the fact that highly repetitive DNA rarely encodes protein.
The program provides a probability score for the presence of exons in a particular stretch of sequence which is divided into three categories: excellent, good and poor. The GRAIL program can detect about 60% of exons, small exons usually accounting for the 40% which are missed.

4.1.4. Localisation of a gene provisionally mapped to the MHC: the human retinoid X receptor β gene (hRXRβ)

Retinoid X receptors (RARs) and retinoic acid receptors (RARs) form two distinct classes of the nuclear hormone receptor superfamily that mediate the effects of steroid hormones and retinoic acid (RA) on cell growth and differentiation (Mangelsdorf et al., 1990). Upon hormone binding, some receptors are translocated from the cytoplasm to the nucleus where they control the transcriptional expression of target genes. This involves binding of the receptors, often in homo- or heterodimeric form, to specific sequences in the target gene promoter. RXRs show high species conservation (Henrich et al., 1990) and respond to 9-cis RA as a ligand (Heyman et al., 1992). The RXR subfamily consists of at least three similar genes, RXRα, RXRβ, and RXRγ, all of which control transcription of target genes mediated by retinoids (Mangelsdorf et al., 1992).

Recently, the human RXRβ gene (hRXRβ) was mapped to the short arm of chromosome 6 (6pter-q13), which also contains the MHC (Fleischhauer et al., 1993). This was interesting as the murine RXRβ (mRXRβ) is known to activate transcription of MHC class I genes in response to RA and has been mapped to chromosome 17, which contains the murine MHC (Hamada et al., 1989). At this time it was not known whether hRXRβ was involved in retinoic acid mediated induction of MHC class I gene transcription in humans but localisation of the gene close to the MHC would provide circumstantial evidence for such a role.

Initial mapping of the hRXRβ gene in humans was achieved by Southern hybridisation of EcoRI digests of human-roden somatic cell hybrids with a full length mRXRβ cDNA probe (Fleischhauer et al., 1993). Analysis of several hybrids containing spontaneous breaks and well-characterised translocations permitted assignment of the gene to the short arm of chromosome 6. For example, one human-hamster hybrid contained a spontaneous break involving chromosome 6 with retention of only 6pter-q13 as well as chromosome 22 and X; this hybrid
retained the human RXRβ gene. I decided to localise the hRXRβ gene, aiming to ascertain if it mapped to the MHC class II region.

4.2. Results and Discussion

4.2.1. Exon amplification

Exon amplification was carried out on new and previously isolated cosmids covering the region of interest in the search for transcribed sequences. The areas covered were those between the genes DPA1 and DNA, cloned in the cosmids O19A and A1 (figure 3.11; Blanck and Strominger, 1990; Chapter 3), and the region surrounding the DMA and DMB genes, cloned in the cosmid HA14 (figure 3.11; Blanck and Strominger, 1990). These regions were selected to complete work in our laboratory screening all cosmids covering the KE3 to DMB region for transcribed sequences.

Exon amplification was carried out by Dr. M. North (Genome Analysis Laboratory, ICRF), producing exons which I isolated by PCR. The resulting exons were used as probes onto B and T cell cDNA libraries and further analysed by subcloning and sequencing. A schematic representation of exon amplification is shown in figure 4.2. Cosmid clones were digested into fragments of ~4 kb (maximal size that can be inserted into the exon amplification vector) using the restriction enzymes BamHI, BglII or Sau3A partial digests. These genomic fragments were introduced into the mammalian expression vector, pSPL1. The cloning site is contained within an intron which is flanked by 3' and 5' splice sites from the HIV tat gene. COS-7 cells were transfected with these constructs, and high levels of transcription were driven by the expression vector SV40 early promoter. During in vivo processing, splice sites of any exon contained within the inserted genomic fragment are paired with the HIV tat splice sites so intronic DNA is excised and the exon is retained in the mature RNA. Reverse transcription followed by PCR was used to amplify such 'trapped' exons.

In the course of this work, several improvements were made to the original exon amplification vector, decreasing the number of false positives and increasing the sensitivity of the procedure (Church et al., 1994). This was accomplished by altering the HIV tat exon sequences that flank the vector intron such that a BstXI recognition site was formed upon joining of the tat exons during RNA splicing. The BstXI site is only recognised in processed RNAs containing no trapped
Figure 4.2. Schematic representation of exon amplification (from Church et al., 1994). Genomic fragments are cloned into the multiple cloning site (MCS) in the pSPL3 vector intron. Transient transfection into COS7 cells allows for production of RNA driven by upstream SV40 early promoter sequences. Half restriction sites are joined by RNA processing. RNA is converted to double stranded cDNA followed by BstXI digestion. Digestion separates PCR primer hybridisation sites, thus eliminating this molecule as a substrate for PCR. The presence of a novel exon spliced to the vector exons prevents BstXI site formation and allows for subsequent amplification of this template.
exons. Digestion with this enzyme separates PCR primer hybridisation sites thus eliminating these molecules as substrates for PCR (figure 4.2).

Initially, cosmid U15 (containing the genes LMP2, TAP1, LMP7 and TAP2) was used as a model to test the efficiency of exon amplification (North et al., 1993). Amplified products from the cosmid were checked either by hybridisation to cDNAs mapping to the region or by direct sequencing. On separate rounds of exon amplification it was possible to identify at least one exon from each gene known to be present on the cosmid. Different amplification experiments generally gave a reproducible number of PCR products from cosmid U15. Sequencing of the PCR products revealed that the exons amplified were exons 2, 4 and 5 from LMP2; exon 4 from LMP7; exon 2 from TAP1 and exon 3 from TAP2. This confirmed that exon amplification was a reliable method of identifying coding sequences and was hence applied to the cosmids O19A and HA14. Cosmid A1 was isolated after this stage of the work and was not used for exon amplification. It was assessed for expressed transcripts by sequencing and hybridisation onto cDNA libraries (discussed later).

Since exons could be lost if an appropriate BamHI/BgIII cosmid fragment was too large to be efficiently subcloned into the pSPL1 or pSPL3 vector it was decided to partially digest the cosmids with the restriction enzyme Sau3A. After digestion, the cosmids were size fractionated to produce fragments between 1 and 6 kb which were subcloned into pSPL1/3. Two rounds of exon amplification were carried out per cosmid and the products analysed. Both cosmids contain known genes (DNA on O19A; DMA and DMB on HA14) which were used as internal controls.

4.2.2. Exon amplification of cosmid O19A

In both rounds of exon amplification of cosmid O19A two PCR products were produced. These were identified, by hybridisation, as being exons from the gene DNA. It was tentatively concluded, from these results, that no further genes were present on cosmid O19A. This conclusion was confirmed by hybridising the cosmid onto cDNA libraries from B and T cells, competing out repetitive elements with human placental DNA. Seven positive cDNA clones were isolated which were all found to originate from the DNA gene.
4.2.3. Exon amplification of cosmid HA14

In both rounds of exon amplification of HA14 four PCR products were produced. Two of these products were identified as being exons from the DMB gene and two products were found to correspond to novel exons. Exon HA14-4 is 135 bp in size and exon HA14-2 is 424 bp in size. Both exons were subcloned into the vector pAMP and sequenced (figure 4.3A/B).

Both exons were hybridised to Southern blots containing a range of cosmids covering the class II region (figure 4.4). Final washing stringency was 65°C, 0.2X SSC. Positive hybridisation signals were only observed on the cosmid HA14 showing that the exons were specific for this cosmid. On hybridisation of HA14-4 and HA14-2 exons to Southern blots of the DMA and DMB cDNAs no positive signals were seen indicating that the exons are not contained within the DMA and DMB genes (figure 4.4). This, however, does not rule out the possibility that the exons represent alternative splice products of the DMA and DMB genes.

Since the complete genomic sequence of the cosmid HA14 had been obtained (S. Beck, personal communication) it was possible to accurately position the exons on the genomic map in relation to DMA and DMB. This showed that both exons mapped between the DMA and DMB genes with HA14-4 residing 364 bp from the 5' end of DMB. Exon HA14-2 is located 2242 bp centromeric of the 5' end of DMB and 2191 bp telomeric of DMA. The two exons are separated from each other by 1845 bp. This close proximity raises the possibility that HA14-4 and HA14-2 may be alternative splice products of DMB. Both exons were used as probes onto cDNA libraries from B and T cells. In each case no positive hybridisation signals were observed. This could be explained by the fact that the gene encompassing these exons is not expressed in these tissues. In view of the close proximity of the exons they may be exons from the same gene.

Oligonucleotide primers were designed from both the exons and used to ascertain their tissue expression patterns. PCR reactions were carried out on genomic DNA samples obtained from various tissues using the cosmid HA14 as a positive control since this cosmid encodes both the exons. The different genomic DNA samples used in these experiments were derived from mammary gland, brain, testis, osteosarcoma cells, HeLa cells, foetal liver, placenta and fibroblasts. In both cases a PCR product was observed in the reactions using cosmid HA14 as template DNA but not with any of the tissue samples (data not shown). The reasons for this could be twofold. First, the tissue expression pattern of the exons
Figure 4.3A. Complete nucleotide sequence of exon HA14-2 showing predicted amino acid sequence in 6 frames. One open reading frame (ORF) is possible; numbered 6, allowing translation in one direction. Comparison of genomic and exon sequences indicated that HA14-4 is transcribed (5' to 3') in a telomeric to centromeric direction in the MHC class II region. Genomic/exon nucleotide sequence comparison also showed that exon HA14-4 splices according to the 'GT-AG' rule.
Figure 4.3B. Complete nucleotide sequence of exon HA14-4 showing predicted amino acid sequence in 6 frames. Three open reading frames (ORF) are possible, numbered 3, 4 and 5, allowing translation in either direction. As with HA14-2, comparison of genomic and HA14-4 sequences showed that exon HA14-4 splices according to the 'GT-AG' rule.
Figure 4.4. Hybridisation of HA14-2 (A) and HA14-4 (B) exons to Southern blots of EcoRI-digested cosmids covering the class II region. In both cases the only positive signal was on cosmid HA14. Figures 4.6C and 4.6D show hybridisation of the exons to Southern blots of the DMA and DMB cDNAs. No signals were seen, hence HA14-2 and HA14-4 are not encoded within the DMA or DMB cDNAs, (confirmed by sequencing).
could be limited to a particular tissue not present in this sample. Second, the exons are developmentally regulated and encode a gene which is only expressed for a specific time during development. The conclusive way to test this hypothesis would be to isolate genomic DNA from different mouse tissues at various stages of development and look for expression of the exons.

Recently, efforts to characterise the human genome and advances in mapping human genetic disorders have led to an interest in techniques for identifying transcribed sequences from specific chromosomes. One approach commonly used to screen for coding sequences in cloned genomic material involves the use of Zooblots (Monaco et al., 1986). As protein coding sequences are generally more conserved during evolution than non-coding sequences, genomic clones can be used as hybridisation probes on blots of genomic DNAs of other species. Conserved sequences, which are probably protein coding sequences, hybridise at a higher stringency with genomic DNAs of other species than do less conserved, putative non-coding sequences.

Since cross-hybridisations of the class II regions from human and mouse have shown remarkable gene colinearity (Hanson and Trowsdale, 1991) it was decided to study conservation of exons HA14-4 and HA14-2 between human and mouse. As the mouse class II region has been completely cloned in cosmids (Steinmetz et al., 1986) both HA14-4 and HA14-2 exons were hybridised to Southern blots of Xhol and EcoRI-digested DNA from mouse cosmids spanning this region (figure 4.5). Hybridisation conditions were identical to those used normally however, since a cross species hybridisation was being performed, the final washing stringency was lowered to 65°C, 6X SSC.

Exon HA14-4 hybridised to the overlapping mouse cosmids 10.13 and 5.22 indicating that HA14-4 was conserved in the mouse. Murine cosmids 10.13 and 5.22 encompass the genes Ma and Mb which are syntenic to the human DMA and DMB genes, thus positioning exon HA14-4 in an equivalent position in the human and mouse MHC class II regions. In control experiments exon HA14-4 did not hybridise to the mouse cosmids 4.24 and 55.2. Cosmid 4.24 encodes the mouse gene Aβ3 which corresponds to the human gene DPA1. Cosmid 55.2 encodes the mouse gene Eα which corresponds to the human gene DRA1. These results indicate that exon HA14-4 is specific to the overlapping mouse cosmids 10.13 and 5.22 and that no other copies of the exon are present in the mouse MHC class II region. Exon HA14-2 did not hybridise to mouse cosmids at low stringency (data not shown) indicating that HA14-2 could be a false positive generated by exon
Figure 4.5a. Comparative map of the human and mouse MHC class II regions showing the position of mouse cosmids (from Steinmetz et al., 1986).

Figure 4.5b. Hybridisation of exon HA14-4 to a Southern blot of mouse cosmids covering the class II region which were digested with the restriction enzymes EcoRI or XhoI. Final washing stringency was 6X SSC, 65°C. Strong signals were observed on the cosmids 10.13 and 5.22 which overlap each other and encode the mouse genes $Ma$ and $Mb$. $Ma$ and $Mb$ are homologous to the human genes $DMA$ and $DMB$ indicating that exon HA14-4 maps to the equivalent area in the mouse and human MHC class II regions. No signals were observed on the mouse cosmids 4.24 and 55.2. 

The two restriction

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Restriction Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>EcoRI</td>
<td>XhoI</td>
</tr>
</tbody>
</table>

The two restriction enzymes EcoRI and XhoI were digested with mouse cosmids 10.13 and 5.22 and the hybridisation was detected by hybridisation of exon HA14-4.
amplification or that this exon is not conserved between species. A small number of PCR products are known to arise from the amplification of noncoding sequences which do, however, contain regions of high homology to acceptor and donor splice junctions (North et al., 1993). For example, exon amplification of cosmid N3 which overlaps cosmid U15 resulted in the amplification of an Alu repeat. Analysis of the flanking genomic region showed that the Alu repeat contained regions that could act as 5' and 3' splice junctions so that the intervening sequence could be amplified. A similar situation could have given rise to exon HA14-2. Sequence analysis revealed that exon HA14-2 did not show homology with existing Alu families or the Alu repeat amplified from N3. When the specificity of exon amplification as a method of gene isolation was originally tested it was discovered that all single copy products showing species conservation were genuine (North et al., 1993; Buckler et al., 1991). It was thus suggested that non-single copy exon amplification products not showing species conservation were likely to be artefacts.

Since HA14-2 did not show cross-species conservation no further work was carried out on this product. To ascertain if HA14-4 was conserved in species other than the mouse, the exon was hybridised onto a Zooblot (figure 4.6). This was a Southern blot made from EcoRI-digested genomic DNA from a variety of species including chicken, hamster, mouse, rat, rabbit, deer and whale. Human DNA was also present on the Zooblot as a positive control. The initial washing stringency was 65°C, 6X SSC which produced strong hybridisation signals. It was possible to reduce the washing stringency to 65°C, 1X SSC and still see positive hybridisation signals. This is an indication of high species conservation (>80% at the DNA level). Exon HA14-4 hybridises to one EcoRI band in humans indicating that one copy of the exon is present in the human genome. Positive hybridisation signals were found with three bands in hamster and five bands in chicken indicating that three and five copies of exon HA14-4 are present in the hamster and chicken genomes respectively. Exon HA14-4 has thus been present in DNA before separation of the mammalian and avian lineages which occurred about 300 million years ago (Kroemer et al., 1990). These results suggest that HA14-4 is a differentially spliced product of DMB which would account for the high species conservation and the close proximity of the exon to the 5' end of DMB. Inability to detect the exon in cDNA libraries could be due to the fact that this particular spliced product of DMB is very rare. HA14-4 could still be part of a novel gene which is developmentally regulated or have a highly specific tissue distribution.
Figure 4.6. Hybridisation of exon HA14-4 onto EcoRI-digested genomic DNA from various species. Final washing stringency was 1X SSC, 0.1% SDS, 65°C. HA14-4 hybridised to five bands in the chicken, three bands in hamster, two bands in mouse, and three bands in rabbit showing high cross-species conservation. The exon hybridised to one band in humans indicating that the gene is single copy.
Isolation and sequencing of HA14-4 from the mouse would aid in the elucidation of the identity of HA14-4.

In conjunction with exon amplification and cosmid screening of cDNA libraries, genomic sequencing is being used as a third approach to identify novel transcripts and to provide a complete picture of the genomic organisation of the class II region of the human MHC. This project is being carried out in collaboration with Dr. Stephan Beck (DNA Sequencing Laboratory, ICRF). My role in this collaboration was to produce random M13 libraries from cosmids covering the class II region for sequencing. The complete sequence is sent to the NIH for analysis by the GRAIL program which can detect about 60% of genes (Beck et al., 1992b). On analysis of the sequence of the cosmids O19A, A1 and HA14 the only possible exon identified so far was HA14-4 which, although giving no further indication whether HA14-4 is a new gene or a further exon of DMB, does provide evidence that HA14-4 is not an artefact of exon amplification.

4.2.4. Localisation of the hRXRβ gene

The flow-sorted chromosome 6 cosmid library was screened with a 1.2 kb fragment from the hRXRβ cDNA (kind gift from K. Fleischhauer). Seven cosmids were isolated (H1-7) which gave similar banding patterns on EcoRI digestion. This showed that the cosmids overlapped each other. The presence of the hRXRβ gene in these cosmids was confirmed by hybridisation of the cDNA onto Southern blots of the cosmids at high stringency (65°C, 0.2X SSC). Initial mapping studies were carried out by FISH on metaphase spreads of normal human lymphocytes, described in Chapter 3. Cosmid H7 was used as a probe in conjunction with cosmid U15 (Blanck and Strominger, 1988) which spans the genes LMP2, TAPI, LMP7 and TAP2 in the MHC class II region. The hybridisation signals obtained were overlapping in most chromosomes but separated in some, distancing the probes by <1 Mb (figure 4.7). This places cosmid H7, containing hRXRβ, up to 500 kb centromeric of the MHC class II region which is consistent with the initial report (Fleischhauer et al., 1993).

To elucidate this further, the 1.2 kb hRXRβ probe was hybridised to YAC clones spanning the MHC class II region (Ragoussis et al., 1992). A positive hybridisation signal was observed on the YAC DPA and no signal was observed on the YACs 51.2 and 11.2 (figure 4.8a), confirming the in situ results and mapping hRXRβ between the genes KE3 and Y5. Further hybridisation studies
Figure 4.7. FISH of cosmids H7 (red) and U15 (green) on metaphase chromosomes. U15 spans the genes \textit{LMP2}, \textit{TAPI}, \textit{LMP7} and \textit{TAP2} in the MHC class II region (Blanck and Strominger, 1988); H7 contains the hRXR\(\beta\) gene and was isolated from the flow-sorted chromosome 6 cosmid library using the hRXR\(\beta\) cDNA as a hybridisation probe. As the cosmid hybridisation signals are either overlapping or close to each other, they are distanced by \(<1\) Mb, placing hRXR\(\beta\) up to 500 kb from the MHC class II region. The FISH experiments were carried out by Dr. G. Senger, using methods described in Chapter 3, as part of a collaboration between the Human Immunogenetics and Human Cytogenetics Laboratories.
Figure 4.8A. Map of the class II region of the human MHC showing YACs spanning the area. HRXRβ hybridised to the YAC DPA and not to the YACs 11.2 or 51.2, roughly mapping the cDNA between the genes RING2 and RING3.

Figure 4.8B. Detailed map of the MHC showing the precise location of hRXRβ with respect to the genes KE4, KE5 and COL11A2.

Figure 4.8C. Southern blot of cosmids digested with EcoRI and probed with hRXRβ cDNA. Lane 1: cosmid CosHcol.11, lane 2: cosmid HPB.ALL 31 and lane 3: cosmid HPB.ALL 33. The hRXRβ cDNA hybridises to an 11 kb EcoRI band on cosmids HPB.ALL 31, HPB.ALL 33 and an 8 kb band on cosmid CosHcol.11. The CosHcol.11 band is shorter since the cosmid does not contain the entire EcoRI band.
were carried out on cosmid clones covering the region (Hanson et al., 1989; figures 4.8b and 4.8c). Positive signals were obtained with cosmids HPB.ALL 31, HPB.ALL 33 and CosH.col 11 mapping hRXRβ to a 11 kb EcoRI fragment between the KE4 and KE5 genes. This has been confirmed in a recent report which mapped mouse Rxb between H2-Ke4 and H2-Ke5 on chromosome 17 and human RXRβ between KE4 and KE5 in the class II region of the human MHC (Nagata et al., 1995).

These results indicated that the human retinoic acid receptor β (hRXRβ) locus is at 6p21.3, approximately 3 kb centromeric of the KE5 gene. The steroid/thyroid hormones and metabolites of RA, whose effects are mediated by RXRβ, are critically involved in the regulation of cell growth and differentiation (Sporn and Roberts, 1983). One criterion of cellular differentiation is the expression of MHC class I genes on the cell surface. MHC class I gene transcription is controlled by two conserved upstream cis-acting sequences, the MHC class I regulatory element (CRE) and the interferon consensus sequence (ICS; Hamada et al., 1989). In the MHC CRE there are at least three sequences, regions I-III, that bind nuclear factors. The murine RXRβ has been shown to bind to one of these regions, region II, and activates murine class I gene transcription in response to RA in undifferentiated embryonal carcinoma cells (Hamada et al., 1989).

It has also been demonstrated that RA treatment of human embryonal carcinoma cells (EC) activates a heterodimer composed of RXRβ and the retinoic acid receptor (RARβ). This heterodimer binds to region II of the CRE resulting in induction of MHC class I gene expression (Segars et al., 1993). This was studied by chemical cross-linking and immunoprecipitation experiments. 35S-labelled RXRβ was chemically cross-linked with nuclear extracts from NT2 cells, precipitated with biotinylated region II oligonucleotide, reprecipitated by an anti-RARβ antibody and resolved in an SDS gel. Labelled RXRβ was precipitated by an anti-RARβ antibody and formed a cross-linked complex of 100 kDa. This band represented a RARβ-RXRβ heterodimer. RA induction of class I and β-2m surface expression was studied by treating EC cells with RA and testing for cell surface expression with two MAbs, one specific for class I molecules and one for β-2m. It was found that MHC class I and β-2m molecules are co-ordinately induced in EC cells following RA treatment.

It is of interest that a gene involved in activating the expression of MHC class I molecules lies just centromeric of the class II region. At present the
significance of this remains unknown but it may be of advantage to the individual that such genes with associated functions are co-ordinately regulated.

4.3. Conclusions

This chapter has described the complete analysis of the region from RING3 to DMB in the class II region of the MHC. This has resulted in the mapping of a retinoic acid receptor β gene which is implicated in transcriptional control of class I genes. Generally, genes involved in transcriptional control do not need to be located near their targets and can be situated on different chromosomes. Considering the clustering of genes with related functions in the MHC, it is highly interesting that a gene involved in transcriptional control of class I genes should be located so close to the class I region. It is attractive to suggest that this positioning is not fortuitous and presents some advantage to the individual, possibly allowing the co-ordinate regulation of genes involved in immune pathways.

Exon amplification of cosmids over this region has resulted in the identification of two new exons, located between the DMA and DMB genes. The function of these exons has yet to be determined. It is possible that both exons are derived from the same gene. HA14-4 could be an alternative splice product of DMB by virtue of its close proximity to the 5' end of the gene, although comparison of HA14-4 sequence with that of DMA, DMB, class I and class II molecules shows no significant homology at either the DNA or protein levels. At present experiments are underway to sequence HA14-4 from the mouse, which may provide clues as to the function of the exon.

Human and mouse DNA sequence comparisons have contributed a great deal to the identification of coding regions. Zooblots, which detect sequence conservation between new sequences and the genomic sequence of other species, are commonly used in the search for biologically functional domains. Furthermore, differences in genomic organisations can help identify insertion, deletion, rearrangement, gene conversion and other events that can lead to speciation. Recently, a study has been reported that presents the largest sequence comparison (~ 100 kb) between human and mouse gene regions: the T cell receptor Cα/Cδ region (Koop and Hood, 1994). This region was studied because its 100 kb size provided a good test for large-scale DNA sequencing strategies and the region had discrete boundaries. The sequence comparison showed that there
was strict conservation of the organisation of the coding elements and approximately 71% similarity between the human and mouse sequences. This sequence conservation was especially significant considering that 94% of the sequenced DNA had no known function. It was proposed that the organisation and sequence of the T cell receptor Cα/δ region is conserved by natural selection. Since the sequenced region is composed of only 6% coding elements, this selection was proposed to extend to the non-coding regions as well. The non-coding region may contain information in relation to storage, duplication and evolution of the region which would impose selective constraints on the DNA. It is possible that HA14-2 and HA14-4 correspond to such non-coding but functionally important regions of the MHC and sequencing the corresponding mouse regions will help to elucidate this further.

Screening and sequencing of cosmids covering the region confirmed that no other genes are present in the region. Thus it seems as if the class II region consists of clusters of genes with gaps of up to 100 kb between them. This is in distinct contrast to the class III region where the gene density is far higher and will be discussed in further detail later.
CHAPTER 5: SCREENING FOR NEW TRANSCRIPTS IN THE HUMAN CLASS II REGION, DMB TO DRBI

This chapter discusses the genomic analysis of that part of human MHC class II region defined by the genes DMB to DRBl (figure 5.1). Mapping of a phosphatase inhibitor pseudogene, four transcribed cDNAs and the presence of unspliced genomic sequences in cDNA libraries is described.

Figure 5.1. Map of the class II region of the human MHC from DMB to DRA (taken from Campbell and Trowsdale, 1993).

5.1. Introduction

The main method utilised in the search for expressed sequences was that of hybridising cosmids fragments onto cDNA libraries, as discussed in chapter 4. The first cosmids used for screening were N3, A12 and A15, spanning the DMB to LMP2 region. Restriction fragments from the cosmids (varying in length between 2-15 kb) were radioactively labelled with $\alpha^{32}$P dCTP and competed with human placental DNA. These fragments were then hybridised to cDNA libraries derived from B or T cells. Cosmid fragments were radioactively labelled by the random priming method of Feinberg (Feinberg and Vogelstein, 1984), with fragments
greater than 10 kb in size being sonicated for 1 min before labelling. This increases the likelihood that the whole fragment will be effectively labelled. To achieve efficient screening of the area of interest overlapping cosmid restriction fragments were used as hybridisation probes.

5.2. Results and discussion

5.2.1. Mapping a phosphatase inhibitor pseudogene to the class II region

One of the first cosmid fragments utilised in screening experiments was a 3.2 kb NotI fragment from cosmid N3. N3 was isolated from the cosmid library derived from the YAC 11.2 and maps to the class II region containing the LMP and TAP genes (chapter 3). The NotI fragment was used as a probe to screen a cDNA library made from the monocyte/macrophage cell line U937 (figure 5.2). Two types of clones of different sizes were isolated: ~1.3 kb and ~1.8 kb, called P19 and P22, respectively. When both cDNA inserts were used as probes on Southern blots containing EcoRI-digested DNA from the cosmid N3 or the YAC clone 11.2 covering the same genomic interval, an identical pattern of hybridisation was detected (figure 5.3). Therefore, the cDNAs are likely to be overlapping. The 1220 bp P19 clone was completely sequenced. The nucleotide and predicted amino acid sequences are shown in figure 5.4. The open reading frame (ORF) encodes a protein of 205 amino acids with a molecular mass of 23 kDa and a pI of 4.38. Comparison of the nucleotide sequence from clone P19 with sequences in the database (EMBL and GenBank) revealed 94% identity with rabbit phosphatase inhibitor-2 (IPP-2) nucleotide sequence over the ORF (Zhang et al., 1992). At the amino acid level, the predicted human IPP-2 peptide shares 92.7% identity with rabbit IPP-2.

5.2.2. Phosphatase inhibitors

Two small heat-stable proteins were first discovered in rabbit skeletal muscle by Huang and Glinsman (1976). They inhibited protein-phosphatase-1 (PP-1) which is one of the major types of serine/threonine phosphatases in eukaryotic cells (Cohen and Cohen, 1989). These two proteins were called phosphatase inhibitor-1 (IPP-1) and inhibitor-2 (IPP-2). IPP-1 is only active after phosphorylation on a threonine residue by cAMP-dependent protein kinase, while IPP-2 does not
Figure 5.2. Map of the class II region of the MHC between the *DMB* and *LMP2* genes, showing overlapping cosmids spanning the region. The large-scale map indicates, in bold, the 3.2 kb EcoRI fragment used to screen the U937 cDNA library (monocyte/macrophage) and the 3.5 kb EcoRI fragment sequenced from the cosmid N3.
Figure 5.3. Southern blot analysis of cDNA P19. 5μg of DNA from cosmid N3 was digested with EcoRI, blotted and hybridised with human IPP-2 cDNAs P19 (A) and P22 (B). Similar hybridisation patterns were detected with both probes indicating that the two cosmid cDNAs overlapped each other.
require phosphorylation (Cohen et al., 1977). Both proteins have been isolated and their amino acid sequences determined. IPP-1 is a protein of 165 residues with a molecular mass of 18.6 kDa, while IPP-2 is a protein of 203 residues with a molecular mass of 22.8 kDa.

IPP-2 interacts with the catalytic subunit of PP1 (PP1C) forming a heterodimer called PP1I. The heterodimer is inactive, but can be activated \textit{in vitro} by the reversible phosphorylation of IPP-2. This activation is catalysed by the glycogen-synthase-kinase, GSK3, by phosphorylation of a specific threonine residue (Thr 72) on IPP-2 (Ballou and Fischer, 1986). \textit{In vivo}, IPP-2 is phosphorylated on Ser 86, Ser 120, and Ser 121 (Holmes et al., 1987). These phosphorylations do not directly activate PP1I but enhance the rate of Thr 72 phosphorylation on IPP-2 by GSK3 and therefore activate the complex (DePaoli-Roach, 1984). The role of PP1I is unknown.

In a recent paper, it was suggested that IPP-2 is critical for the correct folding of PP1C and could be regarded as a new type of chaperone molecule (Alessi et al., 1993). IPP-2 is present in all vertebrate and invertebrate species so far studied and, through its interaction with PP1, is involved in many biological processes. These include glycogen metabolism (Dent et al., 1990), muscle contraction (MacDougall et al., 1991) and mitosis (Axton et al., 1990). Phosphatase inhibitors have also been found to stimulate the production of Tumour Necrosis Factor A in human monocytes (Sung-Sang et al., 1992) and to activate the Na\textsuperscript{+}-K\textsuperscript{+}-2Cl\textsuperscript{−} cotransporter in rat parotid cells (Paulais and Turner, 1992). Thus, phosphatase inhibitors have important functions in a wide range of processes.

5.2.3. IPP-2 is a pseudogene

Comparison of the predicted amino acid sequence of the human IPP-2 clone in the class II region, P19, with sequences in the protein databases (Swiss-Prot) showed a strong homology between amino acids 129 (leucine) and 137 (lysine) with a short stretch of sequence from the \textit{fos} gene family (figure 5.4). This homology was first noticed by Holmes and coworkers (Holmes et al., 1986). As previously described, the short sequence is also strongly conserved between members of the \textit{fos} family, either in the cellular proto-oncogenes (\textit{c-fos}) and their viral counterparts (\textit{v-fos}) or in the growth factor \textit{fosB} (Zerial et al., 1989). The homology starts exactly at the 5' end of the third exon of the \textit{fos} genes which is
Figure 5.4. Complete sequence of the human IPP-2 cDNA P19. The ORF encodes a protein of 205 amino acids (single amino acid code). The differences between the rabbit IPP-2 sequence and the human nucleotide sequence are shown above and the amino acid sequence differences are shown below. The grey box indicates homology between P19 and a fragment of the human c-fos protein. The bottom lane corresponds to the amino acid sequence of c-fos, dashes showing positions of identity with the human IPP-2 sequence. Two potential N-glycosylation sequences are underlined. The first one is conserved between both species but the second one, located just before the stop codon is only found in the rabbit.
part of a highly conserved domain in mammalian, avian and Xenopus fos proteins (Molders et al., 1987). This domain, indispensable for the transforming properties of fos proteins, is a highly charged helical region with structures involved in nuclear transport, DNA and protein binding (Molders et al., 1987). The IPP-2 motif, LSPEEREKKRQ shows strong homology with the fos domain.

End sequencing was performed on the longer cDNA, P22. The sequence showed a perfect overlap with P19 at the 5' end but a longer 3' untranslated end. This result was consistent with the hybridisation patterns of the cDNAs on cosmid clones (figure 5.3) and further work was carried out on the smaller cDNA, P19.

The cDNA P19 was used as a probe for Southern blot analysis of total human genomic DNA digested with EcoRI. Several bands of various intensities were detected by hybridisation (figure 5.5a). One of these bands, indicated by an arrow, was also revealed by hybridisation of an EcoRI digest of the cosmid N3 with the same probe (figure 5.3). The remaining signals detected on the genomic Southern blot indicate other highly related sequences in the human genome since hybridisation was carried out under stringent conditions (0.2X SSC, 0.1% SDS, 65°C).

Zooblot analysis with the IPP-2 probe showed strong evolutionary conservation, with signals present in all vertebrate species (final washing conditions 1X SSC, 0.1% SDS, 65°C; figure 5.5b). Multiple bands were observed in both mouse and hamster genomes. In other species like the rabbit, only one band was visible. This conservation is to be expected since, as mentioned previously, IPP-2 is present in all vertebrate and invertebrate species so far studied and shares homology with a stretch of sequence from the highly conserved fos family of proteins.

The tissue expression pattern of IPP-2 was analysed by probing a human multi-tissue northern blot with the complete IPP-2 cDNA (figure 5.6). Two transcripts of ~2 kb and ~4 kb were detected in all lanes. The small transcript probably corresponds to the 1.8 kb cDNA (P22). The large transcript could be due to another gene, differential splicing, or an extended 3' end. The latter explanation is favoured, since several cDNAs isolated from the DX3 cDNA library (malignant melanoma) shared the same ORF and 5' end but the 3' untranslated region varied between the clones.

To compare the genomic and cDNA sequences of IPP-2 an EcoRI fragment from the cosmid N3 was completely sequenced. This fragment contains part of the probe used to screen the U937 cDNA library (figure 5.2). The genomic, human
Figure 5.5. (A) Southern blot analysis of the cDNA, P19. 10μg of human genomic DNA was digested with EcoRI, blotted and hybridised with the human IPP-2 cDNA. Under stringent conditions several bands were detected. The 3.5 kb band, previously detected on cosmid N3, is indicated by an arrow. Other signals correspond to different IPP-2-related sequences in the genome.

(B) Detection of conserved sequences by cross-species hybridisation. 10μg of each genomic DNA was digested with the restriction enzyme EcoRI and analysed on a Southern blot with human IPP-2 cDNA as a probe. After a high stringency wash (0.2X SSC, 65°C) bands, indicated by arrows, were detected in various species. The gene is single copy in some species (chicken, dog, rabbit, rat, guinea pig and cat) but multi-copy in human, mouse and hamster genomes.
Figure 5.6. Northern blot analysis with the human IPP-2 probe. 2μg of poly A+ RNA were loaded in each lane of a multiple-tissue northern blot (Clontech). After a high stringency wash (0.1X SSC, 65°C) two bands were detected by autoradiography. The smaller transcript corresponds to the 1.8 kb cDNA (P22) isolated from B and T cell cDNA libraries. The larger transcript is most likely to correspond to clones with a longer 3’ untranslated region since several cDNAs of 4 kb were isolated from a malignant melanoma (DX3) cDNA library which shared the same ORF and 5’ end but differed from the cDNA P22 in the 3’ untranslated region.
cDNA and rabbit cDNA sequences are shown in figure 5.7. The homology between the genomic fragment of N3 and the human IPP-2 cDNA sequence was 90% at the nucleotide level (not shown) and 88% at the amino acid level. The cDNAs (P19, P22) are probably products of a gene located elsewhere in the genome. A frameshift was present in the genomic sequence at position 100 owing to a single base pair deletion in cosmid N3. The genomic sequence of two other clones spanning this region was determined (cosmid A15 and P1 clone, DP2). These clones did not contain the frameshift mutation and it was concluded that the mutation probably took place in cloning IPP-2 in cosmid N3. None of the sequenced IPP-2 cDNAs isolated from the two different cDNA libraries U937 and DX3 were identical to the gene present in the class II region. In addition, the gene on cosmid N3 did not contain introns. Taken together, these data suggest that the gene in the class II region of the MHC is a non-expressed pseudogene.

By screening various cDNA libraries from B and T cells with the 1.8 kb IPP-2 cDNA, P22, several clones of different sizes varying from 1.2 to 4 kb were isolated. The ends of these cDNAs were sequenced and compared to the genomic sequences from N3 and the overlapping cosmid U15 (figure 5.2). All clones were similar at the 5' ends, differences being due to the lengths of the 3' untranslated regions. The ORFs of 15 of the 4 kb cDNAs were partially sequenced with internal primers. No nucleotide differences were detected between the IPP-2 clone P19 and these long transcripts, hence it was concluded that all the isolated cDNAs were derived from the same gene.

To localise the IPP-2 gene in the human genome a 3.8 kb IPP-2 cDNA was used as a probe for FISH on metaphase chromosomes (figure 5.8). Signals were detected on three chromosomes: 3q29, 5q33 and 6p21.31 which contains the class II region of the MHC (Senger et al., 1993). This indicates that there are at least three IPP-2 sequences scattered in the human genome, one of which corresponds to the gene itself and the other two (including the signal on chromosome 6p21.31) are probably due to pseudogenes.

5.2.4. Isolation of four cDNAs from the DMB to LMP2 region of the class II region

When the cosmids A12 and A15 were isolated from the flow-sorted chromosome 6 cosmid library, they were mapped by a combination of restriction enzyme digestion and hybridisation to known genes (described in chapter 3). Once a
<table>
<thead>
<tr>
<th></th>
<th>hIPP-2</th>
<th>cIPP-2</th>
<th>rIPP-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-------H---I---NK---------Q---N---D---E---------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>MASTASQRPLKIGKDNTSTTSMVASAEHPFGSVHEQLSKSKQKWDEM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>-------H---------M---------A---------P-------I------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>NILATYRPADKDYGLMKIDEFSTFYSTMGDDEADCSDTETTEAMATDS*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>-------H---------MI---------D---------Y---------TP-----T</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 5.7.** Complete predicted amino acid sequences of the human IPP-2 cDNA (hIPP-2), genomic IPP-2 from the cosmid N14 (cIPP-2) and the rabbit IPP-2 cDNA (rIPP-2). Dashes show positions that are identical between at least two sequences. Position 99 (in bold) is the only one where the three sequences are different. ^^^^^^^ indicates homology with the human fos sequence. The frame shift mutation at position 100 and the stop codons are indicated by an asterisk.
Figure 5.8. Mapping of the human IPP-2 genes by FISH using a 3.8 kb IPP-2 cDNA as a probe on replication G-banded metaphase chromosomes. FISH methods are described in Chapter 3. (A) Specific hybridisation signals were detected on chromosome bands 3q29 (→), 5q33 (Δ) and 6p21.31 (▲) as shown in two G-banded partial metaphases. All three hybridisation sites were found consistently in more than 10 metaphases. Signals on other chromosome bands were random. The FISH experiments were carried out by Dr. G. Senger as part of a collaboration between the Human Immunogenetics and Human Cytogenetics Laboratories.

(B) Idiograms of chromosomes 3, 5 and 6 with bars showing the hybridisation sites of IPP-2.
restriction map had been established the cosmids were screened for the presence of new genes. This was achieved by hybridising fragments from the cosmids onto cDNA libraries. Combinations of restriction fragments were chosen (usually EcoRI, XhoI or NotI) so that the whole cosmid could be screened in five or less hybridisation experiments.

The first fragment to produce a positive result was a 9 kb XhoI fragment from the cosmid A15, screened onto the B cell cDNA library JY (figure 5.9). Four positive clones were isolated which were ~1 kb in size and could be cut in half by the restriction enzyme XhoI. End sequencing of these clones showed that they were identical and further work was carried out on one of the clones, S1. This clone was completely sequenced (figure 5.10) and compared with sequences in the EMBL database. S1 is 983 bp in length and showed no significant homology to sequences in the database at either the DNA or protein levels. The clone has no poly A+ tail and does not contain a long ORF (figure 5.10).

In parallel with the screening approach in the search for expressed transcripts, a sequencing approach was also undertaken in collaboration with Dr. Stephan Beck (DNA Sequencing Laboratory, ICRF). Cosmids A12 and A15 were sonicated into small fragments, subcloned into the vector M13mp18 and completely sequenced. As the genomic sequence of this region became available it was compared to the S1 cDNA sequence providing information on the genomic organisation of S1. S1 was identical to the genomic sequence and was composed of two exons of 874 and 112 bp (290 and 36 amino acids) in size separated by an intron of 5962 bp. As with the cDNA sequence, the genomic sequence of S1 did not contain an ORF. The derived amino acid sequence of S1 contained stop codons in all frames at short intervals throughout the length of the clone.

At this point I decided to screen other cDNA libraries with S1 to ascertain if similar clones were expressed in these libraries. This would provide an indication of whether S1 encoded a gene or was due to aberrant transcription Initial screening experiments gave very high background hybridisation levels. Further analysis of the genomic/cDNA sequence revealed the presence of an Alu repeat in the 872 bp exon. In general, Alu repeat sequences are about 300 bp long and consist of two directly repeating monomer units (Kariya et al., 1987). They are the most abundant family of repetitive DNA sequences, being present at about 500 000 copies per haploid human genome. The presence of Alu repeats in cDNA sequences has previously been described. For example, the tissue factor (TF) gene contains three full-length Alu elements and one partial repeat (Mackman et al.,
Figure 5.9. Cosmid map of the class II region from \textit{DMB} to \textit{LMP2}. The XhoI fragment from A15 used as a probe onto B cell cDNA library, JY, to isolate novel cDNAs is shown in bold.
Figure 5.10. Complete sequence of cDNA clone S1 showing six frame protein translation. The clone contains neither an ORF nor a poly A(+) tail. Comparison of the genomic and cDNA nucleotide sequences showed that one intron (5963 bp) is present in the genomic DNA. Splicing occurs between base pairs 872 and 873 of the S1 sequence.
Digestion of S1 with the restriction enzyme XhoI produces two fragments, the telomeric fragment being free of repeats. This telomeric fragment was thus used in further hybridisation experiments in conjunction with the 9 kb XhoI fragment originally used to isolate the clone S1.

Four cDNA libraries were screened with the telomeric probe producing different length cDNAs shown in table 1. Although the length of the cDNAs differed between libraries, within a library, the clones were the same length.

<table>
<thead>
<tr>
<th>CDNA LIBRARY SCREENED</th>
<th>LENGTH OF CLONE ISOLATED</th>
</tr>
</thead>
<tbody>
<tr>
<td>NW (B cell)</td>
<td>2.8 kb</td>
</tr>
<tr>
<td>γU937 (monocyte/macrophage)</td>
<td>1.5 kb</td>
</tr>
<tr>
<td>JY (B cell)</td>
<td>1.0 kb</td>
</tr>
<tr>
<td>DX3 (malignant melanoma)</td>
<td>1.0 kb</td>
</tr>
<tr>
<td>HPB.ALL (T cell)</td>
<td>450 bp</td>
</tr>
</tbody>
</table>

**Table 5.1.** Length of different clones isolated from a variety of cDNA libraries, screening with the probe S1.

End sequencing was carried out on the 2.8 kb (S3) and 1.5 kb (S2) clones. Comparison with the genomic sequence and the sequence of S1 allowed the construction of a transcription map for the clones (figure 5.11). Compared to S1, the 2.8 kb clone has a longer centromeric region by 1951 bp. The centromeric end of clone S2 starts at the same position as S1 but its telomeric end is longer by 403 bp. The telomeric ends of S1 and S3 terminate in the same place. All three clones contain the Alu repeat sequence in their centromeric regions. Further sequence analysis using primers derived from the S1 sequence determined that the three clones all splice at identical positions in the genomic DNA. Knowledge of clone length and complete genomic sequence of the region determined that all the clones were composed of two exons (S1: 874 and 112 bp; S2: 874 and 414 bp; S3: 2825 and 414 bp), with one intron of 5962 bp in length present in the genomic DNA.

The cDNA clone S4 was isolated using the 9 kb XhoI fragment from cosmid A15. End sequencing of this clone S4 allowed its complete sequence to be
Figure 5.11. Position and splicing patterns of clones isolated from the cDNA libraries NW (S3), U937 (S2), JY/DX3 (S1) and HPB.ALL (S4). Comparison of the genomic and cDNA sequences shows that S1-3 all contain two exons and one intron (5962 bp) which occurs at the same position in the genomic DNA. They differ from each other in length at both ends. S4 corresponds to 405 bp of genomic sequence with no intron/exon structure. This clone contains a sequence, predicted by the GRAIL program, as having a 'good' chance of containing an exon. Clones S2 and S3 contain a stretch of sequence with an 'excellent' chance of containing an exon. Clones S1, S2 and S3 all contain an Alu sequence at their centromeric ends.
determined (figure 5.12). This clone consists of 403 bp of contiguous genomic sequence. It was possible that this cDNA clone was incomplete and so the S4 clone was used as a hybridisation probe onto other cDNA libraries to try and isolate a full length clone. A total of 12 cDNAs were isolated from the cDNA libraries NW (B lymphoblastoid), DX3 (malignant melanoma) and JY (B lymphoblastoid), all of which were identical to S4 in length and sequence.

The complete clone S4 and the telomeric end of the clone S1 were used as probes onto genomic Southern blots and northern blots made from B cells to determine tissue expression patterns. No signals were observed from either probe. Thus, it would seem as if these cDNAs are expressed at such a low level as to be undetectable. Their presence in cDNA libraries could be caused by aberrant firing of transcription caused by sequences upstream of the telomeric ends of the clones. To further analyse the genomic sequence of this region, the complete A15 sequence was analysed by the GRAIL program, described in chapter 4 (Uberbacher and Mural, 1991). It was interesting that the GRAIL program predicted the presence of an exon which overlapped the telomeric end of the cDNAs S2 and S3 and was 186 bp upstream from the telomeric end of S1, giving the exon a probability score described as 'excellent'. GRAIL also predicted the possibility of an exon covering 160 bp of the sequence of S4, with a probability of 'good'. Oligonucleotide primers were designed from both of these predicted exons and used to make PCR probes. These PCR probes were screened onto a variety of cDNA libraries but the only clones isolated were identical to either S1, S2, S3 or S4. Thus, the conclusion from this work was that the clones isolated, probably did not correspond to genes or produce protein products. Their presence in cDNA libraries was due to aberrant transcription.

5.2.5. Class I-related sequence in the class II region

Analysis of the genomic sequence of A15 also revealed a 88 bp stretch of sequence that showed homology to class I sequences (figure 5.13a). This sequence homology is located in the intron of the cDNAs S1, S2 and S3. Matches of >72% identity were found with different class I sequences from humans, gorilla and orangutan. Matches of >60% identity were found with class I sequences from mice and rabbits. Comparison of the sequence in the class II region with class I α3 domain consensus sequences (Klein, 1986), revealed that the highest identity occurred between residues 249-274 in the α3 domain (figure 5.13b). It was of
Figure 5.12. Complete sequence of cDNA clone S4 which was isolated from the T cell cDNA library HPB.ALL using a 9 kb XhoI fragment from the cosmid A15. The nucleotide sequence is shown on top and the amino acid sequence (one letter code) is shown underneath. The clone has an ORF of 128 amino acids. Comparison of S4 sequence with genomic sequence (from cosmid A15) indicated that S4 represents a single contiguous stretch of genomic DNA with no intron/exon structure, polyadenylation signal or poly A+ tail (also see figure 5.11).
Figure 5.13 (A). 88 bp stretch of sequence from the cosmid A15 which shows homology with class I sequence from various species. The DNA sequence is shown above and the amino acid sequence (one letter code) is shown below. The short stretch of sequence contains an ORF.

(B). Amino acid sequences of MHC class I α3 domains from mouse (H-2K, H-2D, H-2L, H-2Q and H-2T), rat (RT1), human (HLA) and rabbit (RLA) showing residues 249-274. The bottom line shows the sequence of the stretch of genomic DNA from A15 which has >60% homology with all species at the amino acid level.
A

1  gttggctattggtgtccctctgagtgqagagctatgccat  50
   G G Y W L V P S G V R Q R Y L C H
51  atgcaacataaaaggttgtgcgtgacccctcagcttag  88
   M Q H K R L P E P L S L R

B

H-2Kb  VPLGKEQYTYTCHVYHQGLPEPLTLRW
H-2Kd  VPLGKEQNYTCHVHHKGLPEPLTLRW
H-2Kk  VPLGKEQYYTCHVYHQGLPEPLTLRW
H-2Kw29.7  VPLGKEQYYTCHVYHQGLPKPLTLRW
H-2Db  YPLGKEQYYTCHVYHQGLPEPLTLRW
H-2d  VPLGKEQYYTCHVYHQGLPEPLTLRW
H-2Ld  VPLGKEQNYTCREYHGLPEPLTLRW
H-2Q7d  VPLGKEQNYTCHVHHEGLPEPLTLRW
H-2Q10b  VPLGKEQNYTCHVYHEGLPEPLTLRW
H-2Q10q  VPLGKEQNYTCHVYHEGLPEPLTLRW
H-2T1ab  VPLGKEQKYTCHVYHEGLPEPLTLRW
H-2T1ac  VPSGEEQKYTCHVYHEGLPEPLTLRW
RT1.1  VPLGKEQNYTCLVEHGLPEPLTLRW
RT1.2  VPLGKEQNYTLCVEHGLPEPLTQRW
HLA.A2  VPSGQEQRYTCHVQHEGLPKPPLTLRW
HLA.A3  VPSGEEQRYTCHVQHEGLPKPLTLRW
HLA.Aw24  VPSGEEQRYTCHVQHEGLPKPLTLRW
HLA.A28  VPSGEEQRYTCHVQHEGLPK
HLA.B7  VPSGEEQRYTCHVQHEGLPKPLTLGW
HLA.Cw3  VPSGEEQRYTCHVQHEGLPEPLTLRW
HLA.12.4  VPSGEEQRYTCHVQHEGLPEPLTLRW
HLA.328  VPSGQEQRYTCHMQHEGLQPEPLTLSW
RLA.1  VPSGEEQRYTCRVQHEGLPEPLTLTW
RLA.2  VPSGEEQRYTCRVQHKGLPEPLTLRW

COSMID  VPSGVRQRYLCHMQHKRLPEPLSLR
A15
interest to discover if this stretch of sequence corresponded to a gene which had not been detected by the screening approach.

To this end, oligonucleotides were designed from the class I-like sequence and used to produce a PCR probe. This probe was used to screen five cDNA libraries (from B and T cells) in the search for an expressed transcript. No such transcript was discovered. The sequence could be a remnant of a class I gene which was present in the class II region at one time. The ancestral class I and class II genes are thought to have evolved in their present form before the radiation of vertebrates which took place 400 million years ago (Kaufman et al., 1990). Both class I and class II ancestral genes have been duplicated many times to give the multiple loci seen in the class I and class II regions of the mammalian MHC. There is significant homology between the class II α2 domain, the class II β2 domain and the class I α3 domain which are all related at the amino acid sequence level to the conserved antibody fold domain found in members of the immunoglobulin supergene family (Korman et al., 1982). Alternatively, the class I-related sequence could have been inserted between the DMB and LMP2 genes on a fragment from the class I region. From analysis of sequence data between DMB and LMP2 (S. Beck, personal communication) it can be seen that two LINE repeat sequences are present in the DNA close to the DMB and LMP2 genes. It is possible that recent recombination events in the human MHC have resulted in the insertion of this particular stretch of DNA. There is no direct proof for this hypothesis as yet but when the human class I region is sequenced the origin of this tract of DNA may become apparent.

5.2.6. Completion of cosmid screening of the class II region

Figure 5.14 shows a map of probes derived from cosmids, covering the class II region which have been screened onto cDNA libraries in the search for expressed transcripts. In all cases, cosmid and Southern blots were present in the hybridisation mixture to ascertain that the probes were specific. In some cases other genes on the cosmid fragments acted as positive controls. All the cosmid fragments shown in figure 5.14 have also been used in exon amplification experiments. By combining these two approaches it seems likely that there are no unidentified genes in the MHC class II region between DPB2 and DRB1.

In the course of these screening experiments, several cDNAs have been isolated which apparently are expressed transcripts of contiguous stretches of
Figure 5.14. Map of the MHC class II region from KE3 to DRB1 showing complete cosmid inserts which have been screened onto B and T cell cDNA libraries.
genomic DNA without introns (figure 5.15). One such transcript, SUC1, was isolated using a 15kb XhoI fragment from cosmid A15. This cDNA is 2.5 kb in size and overlaps a similar transcript, previously discovered in our laboratory, called RING8. RING8 is a 2.3 kb cDNA which was isolated with a NotI fragment from the cosmid U15. Like SUC1 it contains no introns and has no polyadenylation sequence (ATAAA). It represents a single contiguous stretch of genomic DNA with no intron/exon structure. Hybridisation of SUC1 and RING8 onto genomic Southern blots and multi-tissue northern blots did not detect transcripts of either cDNA. This suggested that they were unlikely to be expressed single copy genes (data not shown).

A second genomic transcript, GEN1 (figure 5.15), was isolated using a 10 kb BamHI fragment from the cosmid U10 (previously described; Blanck and Strominger, 1988). This cDNA maps precisely to a 3 kb genomic sequence centromeric of TAP2 which is flanked Alu repeat sequences in opposite orientations. Initiation of transcription seems to have originated in the repeat sequences. There is little published on the significance of transcription of 'non-gene' regions of the genome which, from these studies, seems to occur on average once every 150-200 kb in the class II region of the human MHC.

5.2.7. Sequence analysis

So far 300 kb of sequence of the class II region of the MHC has been completed (figure 5.16). On test runs, the GRAIL program has been reliable in locating known exons, except LMP2 where the exons are smaller than 80 bp. Sequence analysis has proved extremely useful in localising all the transcripts described in this thesis. For example, it was possible to immediately identify SUC1 as corresponding to an overlapping transcript of RING8. Differences between IPP-2 cDNA sequence and the genomic sequence suggested that IPP-2 was a pseudogene. Knowledge of the repeat sequences centromeric of TAP2 provided an explanation as to the isolation of the genomic transcript GEN1.

5.3. Conclusions

This chapter has described the genomic analysis of a 280 kb stretch of the class II region of the human MHC between the genes DMB and DRB1. Such analysis led to the localisation of a phosphatase inhibitor pseudogene, IPP-2, centromeric of
Figure 5.15. Map of the class II region of the MHC from DMA to DOB showing position of expressed transcripts of genomic DNA isolated by cosmid screening of B and T cell cDNA libraries. RING8 and SUC1 are overlapping transcripts of length 2.3 and 2.5 kb, respectively isolated from the cosmids U15 and A15. They have no intron/exon structure, initiation or polyadenylation signals. GEN1 is a 3 kb contiguous stretch of genomic DNA, isolated from cosmid U10. It is flanked by two repeats from the human interspersed repeat (HIR) family. S1-4 are cDNA clones isolated from the cDNA libraries NW (S3), U937 (S2), JY/DX3 (S1) and HPB.ALL (S1). Comparison of the genomic and cDNA sequences shows that S1-3 contain two exons and one intron which occurs at the same position in the genomic DNA. All three cDNA clones differ from each other in length, at both ends. S4 corresponds to 405 bp of contiguous genomic sequence with no intron/exon structure.
Figure 5.16. Map of the human MHC indicating areas of the class II region that have been sequenced. The sequence data was obtained from cosmid fragments which are shown as overlapping solid lines. Dashed lines represent cosmid fragments from which sequence is in progress. The cosmid fragments were sonicated and randomly subcloned into the vector M13. *E. coli* cells were transformed with the vector/cosmid constructs, DNA isolated and used for sequencing. Sequence data was compiled into contigs and sent to the NIH for analysis by the GRAIL programme (Uberbacher and Mural, 1991).
the gene \textit{LMP2}. The IPP-2 gene itself has been mapped to either chromosomes 3 or 5 by \textit{in situ} hybridisation.

Three transcripts that are split into two exons and one intron at the genomic level but contain no other obvious feature that would suggest that they code for an expressed gene have been localised to the region between the genes \textit{DMB} and \textit{LMP2}. They contain no poly A\(^+\) tail and are not expressed on northern blots or could be detected on genomic Southern blots. The most likely explanation for these transcripts is that they result from aberrant transcription. Upstream of the telomeric end of these transcripts is a stretch of sequence, identified by GRAIL, as coding for an exon. Hybridisation of probes made from this sequence onto cDNA libraries did not find the relevant gene, nor was this sequence detected by exon amplification.

At the start of this project the class II region between the genes \textit{DMB} and \textit{LMP2} was implicated as a candidate region for a gene(s) involved in the class II antigen processing pathway. Evidence for this hypothesis was discussed in chapter 3. Extensive analysis of this region failed to find any such gene(s). During the course of this study the class II processing genes were identified as being \textit{DMA} and \textit{DMB} (Morris et al., 1994; Fling et al., 1994). Mutant cell lines with deletions in the MHC class II region and concomitant defects in antigen processing through class II were initially reported to map the class II processing gene(s) telomeric of the \textit{DMB} gene (DeMars et al., 1985). Further analysis of these mutants along with the development of new mutant cell lines enabled fine mapping of the region containing the class II processing genes. This led to the identification of the class II processing genes as being \textit{DMA} and \textit{DMB}. The mutant cell lines which were used to map the class II processing genes were unavailable to the Human Immunogenetics Laboratory.

Three unspliced contiguous genomic fragments have also been identified in the course of this study. These fragments contain no polyadenylation signals, no obvious initiation codons and have no intron/exon structure. They are, therefore, further examples of the presence of genomic transcripts resulting from aberrant transcription in cDNA libraries. Sequence analysis of the regions adjacent to these fragments has been unable to determine why transcription should be occurring. It is possible that if enough fragments of this nature are reported a consensus sequence causing such transcription can be identified.

Studies on the expression of genes in the testis have shown that transcripts producing sharp bands on northern hybridisation do not always encode functional
protein products (reviewed in Ivell, 1992). The single-copy pro-opiomelanocortin (POMC) gene is expressed as a 1100 bp mRNA in the pituitary gland and encodes the polyprotein precursor for adrenal cortex trophic hormone (ACTH), α- and γ-MSH (melanin stimulating hormone) and β-endorphin. In the testis, ovary, epididymis and placenta of several species, POMC-specific probes detect an 800 bp transcript which represents the major POMC gene present in these tissues.

However, subsequent studies showed that the 800 bp transcripts were not responsible for the neuropeptides. cDNA sequencing and nucleotide protection assays showed that the 800 bp transcripts in rat and human testis represented transcripts which were missing the 5' upstream non-coding region responsible for translation initiation, as well as the region encoding the signal peptide which targets the hormone precursor into the lumen of the ER. Therefore, the 800 bp transcripts are unable to encode the physiologically relevant peptides.

Similar aberrant transcripts have been reported for the pro-enkephalin A and vasopressin genes (Ivell, 1992). Pro-enkephalin A is the protein precursor for met-enkephalin and is encoded by 1.4 kb transcripts in brain tissues and the adrenal gland. In rat and mouse testis, the major pro-enkephalin A transcript is 1.7 kb. Sequence analysis showed that the 1.7 kb RNA resulted from differential splicing. This 1.7 kb transcript is only found in rats and mice, in humans only the conventional 1.4 kb transcripts are expressed. In rats the 1.7 kb transcripts are non-functional whereas in mice the transcripts associate with polysomes and produce peptides.

Two testicular transcripts are observed on northern hybridisation using a vasopressin-specific probe. These are intermediate in size between the hypothalamic and adrenal transcripts, where physiological peptide production has been observed. cDNA sequence and PCR analysis showed that the testis transcripts were the products of differential splicing. Two new exons, from 10 kb upstream in the genome, were spliced directly onto exon 2, replacing exon 1. Sequence and polysome analysis showed that the transcripts are non-functional.

Similar variant transcripts have been reported for the CREB (cAMP response element binding protein) gene, the oncogene pim-1 and the oncogene c-mos. In general, many variant transcripts are non-functional and they often represent the major form of the transcript in the tissues where they are found. The functional transcripts are less common and appear to be present in different cell types to the non-functional transcripts.
One explanation is that these aberrant transcripts may represent examples of leaky transcription. Within the nucleus, all regions of DNA in an activated state are accessible to many enzymes, transcription factors and DNA-binding proteins that are involved in the expression of a particular phenotype. It is thought that these factors interact with many regions of the genome until an appropriate fit to a regulatory element is found. Hence, it is not unlikely that combinations of consensus elements will occur which can lead to spurious transcriptional initiation, and the presence of such transcripts in cDNA libraries.

The S1, S2 and S3 transcripts in the human MHC class II region could represent non-functional transcripts present in B and T cells. The functional transcripts could be present in a different cell type. However, since the transcripts could not be detected on multi-tissue northern blots, this is unlikely to be the case. It is also possible that transcription irrespective of the encoded function may have some, as yet, unknown function.

Complete screening of the class II region from DP to DR has been achieved by a combination of exon amplification, sequencing and direct screening of cosmid fragments onto cDNA libraries. In the class II region gene density is approximately 1 gene for every 40 kb of DNA. In fact, there are segments of up to 70 kb which are gene-free. This is in contrast to the class III region where there is a gene at least every 15 kb (Campbell and Trowsdale, 1993). As yet, the class I region has not been so intensively investigated but initial studies using cDNA selection on YACs as a method of gene identification estimates the presence of a gene every 20 kb (Geraghty, 1993; Wei et al., 1993).

Thus, the gene density in the class II region is less than both the class I and III regions of the human MHC. Why this should be the case is unknown. The differences in gene density between the class I, class II and class III regions may be a consequence of the function of the genes encoded in these areas. The class III region contains a mixture of genes with different roles, many of which are widely expressed. This may also turn out to be the case for the class I region, as more genes are characterised. In contrast, the class II region is composed of genes with functions in the immune system which have limited tissue expression patterns. The only exception to this is RING3 which is homologous to a Drosophila maternal effect gene (Hanson et al., 1991). There may be some evolutionary advantage conferred on an individual if the class II loci can be regulated as a whole. It is possible that co-ordinate gene regulation enables the individual to respond more effectively to a range of antigenic challenges producing a broader
immune response. However, these questions will only be answered when the class I region has been fully analysed.

The human genome has long range G+C% mosaic structures that are related to chromosome bands (Bernardi et al., 1985). The Giemsa-dark G bands are composed mainly of AT-rich sequences, and the Giemsa-pale R bands mainly of GC-rich sequences (Craig and Bickmore, 1994). Gene density, codon usage, chromosome condensation and repeat sequence density are related to chromosome bands and long-range G+C% mosaic structures (Bernardi et al., 1985; Craig and Bickmore, 1994). Gene-dense R bands with loose chromatin structures replicate early in S phase and are rich in Alu repeats, while G bands with condensed chromatin structures replicate late and are rich in LINE-1 repeats (discussed in section 1.6.1). Class II and class III sequences differ in their G+C content, class II sequences being AT-rich whilst class III sequences are GC-rich (Fukagawa et al., 1995). The transition between the AT-rich and GC-rich domains, i.e. the boundary of long-range G+C% mosaic domains, is approximately 180 kb from DRA in the class II/III boundary region. AT-rich sequences encode fewer genes than GC-rich sequences (Bernardi et al., 1985) which may partly explain the difference in gene density between the class II and class III regions.
CHAPTER 6: CLONING AND CHARACTERISATION OF GENES LINKED TO THE LMP-RELATED PROTEASOME GENE, MB1

6.1. Introduction

6.1.1. The proteasome

Proteins in eukaryotic cells are continually broken down and replaced for several reasons: cells selectively degrade proteins with abnormal conformations, the accumulation of which could be harmful; rapid degradation of regulatory peptides is essential for the control of metabolic pathways and the cell cycle; and the breakdown of dispensable proteins in starvation provides amino acids for energy metabolism (Goldberg and Rock, 1992). The degradation of cytoplasmic proteins is carried out by a multi-subunit protease, called the proteasome.

Proteasomes were first characterised 25 years ago when protein particles were discovered by electron microscopy in extracts of human erythrocytes (Harris, 1968). These particles were later found in cytosolic and nuclear fractions from a broad range of eukaryotic cells, including yeasts and plants (Orlowski et al., 1991). In each case these particles were shown to be 700kD protein complexes with a sedimentation coefficient of 20S, composed of 12-15 different polypeptide subunits. The particles contain three different peptidase activities which cleave specifically after basic, hydrophobic and acidic amino acid residues (Wilk and Orlowski, 1983).

cDNAs encoding all the proteasome subunits have been cloned. These genes are related to each other and are highly conserved from yeast to man. Clarification of the molecular organisation of the 20S proteasome was aided by the discovery of a similar complex in the archaebacteria of the genus Thermoplasma. In Thermoplasma, the proteasome subunits are stacked into a barrel shaped complex of four layers which is ring shaped with a central hole. The two outer rings are composed of seven identical α-subunits, while each inner ring contains seven identical β-subunits (Puhler et al., 1992). Thus the molecular organisation of the particle is \( \alpha_7\beta_7\gamma_7\). A similar type of organisation has been proposed for eukaryotic proteasomes since all their different subunits known to date are closely related in sequence to either the α- or the β- chain of Thermoplasma. Assuming seven subunits in each protein ring, the maximal number of different polypeptides
in the eukaryotic particle is 14. There is sequence similarity, not only within the α- and β-subunit families, but also between the α- and β-sequences themselves, indicating that all subunits were derived from a single ancestral gene early in evolution (Zwickl et al., 1992).

The 20S proteasome is now known to form the proteolytic core of a larger protein complex, which is responsible for protein degradation in the ubiquitin pathway. This proteolytic complex has been termed the 26S proteasome (Peters et al., 1993). The ubiquitin system is a non-lysosomal ATP-dependent proteolytic pathway in which the covalent attachment of ubiquitin, a highly conserved 76-residue polypeptide, is thought to mark proteins destined for degradation (Hershko and Ciechanouer, 1992). Successive enzyme action leads to the formation of isopeptide bonds between the ε-amino groups of Lys residues in the substrate protein and the carboxyl terminus of ubiquitin. Ubiquitin is then itself ubiquitinated. The reaction is repeated, resulting in the formation of large conjugates that are specifically recognised by the 26S proteasome and degraded.

6.1.2. Involvement of the proteasome in the class I antigen processing pathway

The involvement of the proteasome in antigen processing was first suggested in 1989 (Townsend and Bodmer, 1989). A pathway was proposed suggesting that antigenic proteins were degraded, in the cytoplasm, into peptides which were transported through the ER. These peptides then bound to MHC class I molecules and travelled to the cell surface for presentation to T cells. It was proposed that the proteasome was involved in such peptide production and recently evidence has been produced to support this hypothesis.

A group of 16 proteins of molecular weight 15-30kD were precipitated from mouse cell lines using anti-MHC alloantisera and were called Low Molecular Weight Proteins, or LMPs (Monaco and McDevitt, 1982). The LMPs were expressed on monocyte and macrophage cell lines, as well as normal macrophages and spleen cells. Precipitation of LMPs from mice containing cross-overs within the H-2 region with an anti-H-2d serum led to the discovery that two subunits responsible for such precipitation were genetically linked to the murine class II region. These two subunits, named LMP2 and LMP7, were localised to a region of the mouse MHC between the genes Oa and Ob, the equivalents of the DNA and DOB genes in the human.
It was pointed out that the 20S proteasome resembled the LMPs in size and subunit composition (Parham, 1990) and a relationship between the two complexes was soon discovered. Anti-LMP and anti-proteasome immunoprecipitates were compared from a murine macrophage cell extract (Brown et al., 1991). An anti-H-2d antiserum was used to immunoprecipitate LMP antigens and a rabbit antiserum was made against purified rat liver proteasomes. Mixing of LMP and proteasome precipitates before electrophoresis demonstrated that 16 LMP subunits migrated with identical mobility to 16 of the polypeptides in the anti-proteasome precipitate. Two proteasome subunits migrated identically to the MHC-linked LMP2 and LMP7 polypeptides indicating that normal proteasomes contained these subunits. However, four proteasome subunits were not found in LMP precipitates and proteasome could still be precipitated from lysate that had been precleared with anti H-2d serum. This suggests that although the LMP and proteasome complexes are similar they are not identical. The precise relationship between the LMP complex and the proteasome remains to be elucidated.

6.1.3. LMP2 and LMP7: two proteasome subunits localised within the MHC class II region

Screening of cDNA libraries with cosmid clones spanning the MHC class II region \(DNA\) to \(DOB\) in the human and \(Pb\) to \(Ob\) in the mouse allowed the precise localisation of the two genes \(LMP2\) and \(LMP7\) (Glynne et al., 1991; Martinez and Monaco, 1991; Kelly et al., 1991). In the human, these genes flanked the \(TAP1\) transporter gene with the order being \(LMP2, TAP1, LMP7\) and \(TAP2\). Sequence comparisons confirmed that these two genes were members of the proteasome complex. The localisation of \(LMP2\) and \(LMP7\) within the MHC and their close proximity to the transporter genes strongly implicates the proteasome as having a role in antigen processing. The immune system, during the course of evolution, may have recruited a previously existing structure required for normal cellular turnover by attaching two extra subunits that serve to direct its output to the \(TAP\) transporter. Alternatively, the two MHC-linked subunits may serve to modify the proteolytic action of the proteasome to produce peptides better suited for MHC class I binding. Recently, several reports have provided data supporting the latter hypothesis (Gaczynska et al., 1993; Gaczynska et al., 1994; Rock et al., 1994).
As mentioned above, the proteasome exhibits multiple peptidase activities including three activities that preferentially hydrolyse small peptides on the carboxyl side of hydrophobic, basic or acidic residues. These activities are regulated by the cytokine interferon γ (IFN-γ) which is a stimulator of MHC class I presentation. Treatment of cells with IFN-γ increases the capacity of the proteasomes to cleave oligopeptides after hydrophobic and basic residues and reduces cleavage after acidic residues (Gaczynska et al., 1993). IFN-γ treatment also induces the addition of the LMP2 and LMP7 subunits to the proteasome and the loss of two other units (Yang et al., 1992). These changes in peptidase activity are not seen on IFN-γ treatment of the deletion mutant 721.174 which lacks the LMP2 and LMP7 genes. These findings led to the suggestion that the incorporation of LMP2 and -7 into the proteasome was responsible for the alterations in peptidase activities.

This hypothesis was investigated in the following manner. LMP2 and LMP7 cDNAs were either transfected into HeLa cells or 721.174 B-LCLs (Gaczynska et al., 1994). Proteasomes were then isolated from transfectant and control cells, the levels of the LMP2 and LMP7 subunits assessed, and their peptidase activities compared. LMP7 transfection into the strain 721.174 enhanced the hydrophobic and basic site activities, whilst LMP2 transfection lowered the acidic site activity. Similarly, transfection of HeLa cells with LMP2 and LMP7 mimicked the effects of IFN-γ treatment, which enhances hydrophobic and basic activities and reduces cleavage after acidic residues. This effect favours the generation of the kinds of oligopeptides that preferentially bind to MHC class I molecules providing circumstantial evidence for involvement of the proteasome in antigen processing.

The final piece of evidence confirming the role of the proteasome in the class I processing pathway appeared recently (Rock et al., 1994). Ovalbumin was introduced into the cytosol of murine B lymphoblastoid cells and the effect of peptide aldehydes, which are known to inhibit peptidase activities, on its presentation was tested. When antigen presenting cells were exposed to the peptide aldehyde, presentation of ovalbumin was inhibited in a dose dependent manner up to 100%. Ovalbumin presentation required the TAP transporter and was inhibited by brefeldin A. Aldehydes had no effect on the presentation of ovalbumin-derived peptide in cells infected with a vaccinia viral construct containing a minigene encoding this peptide. Therefore, aldehydes selectively blocked the production of the immunogenic peptide and not any subsequent step in the class I pathway. All these results show that the proteasome is responsible
for the generation of peptides presented to MHC class I molecules and thus has a major role in the class I antigen processing pathway.

6.1.4. Two proteasome components related to \textit{LMP2} and \textit{LMP7}

It was mentioned previously that IFN-\(\gamma\) treatment induces the addition of the \textit{LMP2} and \textit{LMP7} subunits to the proteasome and the loss of two other units. Recently two genes that encode strong candidates for these replaceable subunits have been identified (Belich et al., 1994). It is known that disruption of proteasome subunits in yeast results in a lethal, or severely deleterious phenotype (Heinmeyer et al., 1991; Emori et al., 1991), but \textit{LMP}-negative LCL721.174 mutant cells grow normally. In addition, transfection of the \textit{TAP} cDNAs into 721.174 cells restores their ability to process and present antigen through the class I pathway (Momberg et al., 1992). These results raised questions as to the function of the proteasome and, more specifically, the \textit{LMPs} in antigen processing. One hypothesis that would account for the observations suggested that expression of alternative proteasome subunits in the \textit{TAP}-transfected 721.174 mutant cells would allow the proteasome to function normally. In order to test this hypothesis cDNA libraries were screened with \textit{LMP2} and \textit{LMP7} cDNAs under non-stringent hybridisation conditions to isolate \textit{LMP2} and \textit{LMP7}-related cDNAs (Belich et al., 1994).

Two genes were isolated by this process which encode new proteasome subunits. \textit{MB1} is encoded on chromosome 14 and is closely related to \textit{LMP7} (67\% identity at the amino acid level). \textit{Delta} is encoded on chromosome 17 and is closely related to \textit{LMP2} (59\% identity at the amino acid level). Expression of the \textit{MB1} and \textit{delta} genes is reciprocal to that of the \textit{LMP} genes: \textit{MB1} and \textit{delta} are upregulated in mutant cell lines lacking \textit{LMPs} and down-regulated in the presence of IFN-\(\gamma\). Thus it seems as if the proteolytic activities of the proteasome are altered by differential expression of two subunits. Incorporation of the \textit{LMP} genes into the proteasome alters its cleavage abilities such that peptides are produced that preferentially bind to MHC class I molecules. However, in cell lines that lack the \textit{LMPs}, expression of the homologous \textit{MB1} and \textit{delta} subunits allows sufficient peptides to be produced, so that class I expression and antigen presentation are not significantly impaired.

Amino acid sequence comparisons of proteasome subunit genes reveals that \textit{LMP2} and \textit{LMP7} differ significantly from each other (only 32\% identity). \textit{MB1}
and \textit{delta} show similar, weak, identity (29\%) suggesting that the two gene pairs are not recent duplicates of each other. Considering the much stronger homology of LMP2 to \textit{delta} and LMP7 to \textit{MBI} it seems likely that LMP2 and \textit{delta} originated from a common ancestral gene, and LMP7 and \textit{MBI} from another. Further sequence comparison of proteasome subunits from different species provided candidates for these primordial loci in other species: the yeast gene \textit{PRE-3} is closely related to \textit{delta}, and \textit{PRE-2} is homologous to \textit{MBI} (Heinemeyer et al., 1993; Hilt et al., 1993). Two models have been proposed to explain the genetic origins of these genes (Belich et al., 1994). 1) The primordial genes were LMP2 and LMP7. These were initially linked and, after duplication and mutations, remained linked in the MHC. The duplicated genes (\textit{MBI} and \textit{delta}) split onto different chromosomes. 2) The primordial genes were \textit{MBI} and \textit{delta}. These duplicated and mutated on different chromosomes. The duplicated copies (LMPs) were brought together within the MHC to form a gene cluster with the TAPs. This second hypothesis seems the most attractive since the LMPs show less amino acid identity with their respective yeast homologues than do \textit{MBI} and \textit{delta}. Also, \textit{MBI} and \textit{delta} encode constitutively expressed proteasome subunits found in most tissues whereas LMP2 and LMP7 are up-regulated in cells of the immune system. Therefore, it seems likely that the LMP loci arose later in the evolution of vertebrates than the \textit{delta} and \textit{MBI} genes.

\textbf{6.1.5. Evolution of the LMP/TAP gene cluster: are the LMP-related proteasome genes, \textit{MBI} and \textit{delta}, linked to TAP-related genes?}

On considering the second hypothesis for the evolution of the LMP genes, outlined above (Belich et al., 1994), it is interesting to ascertain whether the theory could be extended to the TAP genes as well. Could the LMP/TAP gene cluster have been "captured" by the MHC due to an evolutionary advantage in having the antigen processing and presentation genes linked? A finding consistent with this idea is provided by analysis of the \textit{cim} phenotype in the rat where different RT1.A alleles seem to be better provided with peptides from linked TAP2 alleles (Joly and Deversen, 1994). However, the TAP loci are far more polymorphic in the rat, as compared to the human, and the advantages conferred by the different TAP alleles in the \textit{cim} system may not apply to humans. It may be unnecessary to have alternative TAP alleles in view of the large number of class I genes in humans.
If the \textit{LMP/TAP} gene cluster was captured by the MHC then the primordial \textit{TAP} genes may be located close to the primordial \textit{LMP} genes (\textit{MB1} and \textit{delta}). The initial mapping work revealed the \textit{TAP} and \textit{LMP} genes to be tightly clustered, only 593 bp separating the ATG translation initiation codons of \textit{LMP2} and \textit{TAPI}, which share a bi-directional promoter (Wright et al., 1995). As part of my analysis of the genomic organisation of the class II region, I decided to try and discover more about the evolution of the \textit{LMP/TAP} gene cluster with the aim of determining how the cluster had come to reside in the class II region. Studying the evolution of this gene cluster may make it possible to answer questions concerning the evolutionary advantage of keeping the \textit{TAP} and \textit{LMP} genes together. I therefore chose to determine if the \textit{LMP/TAP} linkage was also a feature of non-MHC proteasome genes.

\textbf{6.2. Results and Discussion}

\textbf{6.2.1. Isolation of cosmids encompassing the \textit{MB1} and \textit{delta} genes}

To ascertain whether any \textit{TAP}-associated genes were situated close to either of the \textit{MB1} or \textit{delta} genes it was decided to isolate cosmids encompassing the two genes and screen for the presence of novel cDNAs on these cosmids. Full length \textit{MB1} and \textit{delta} cDNAs were hybridised onto a complete human genome cosmid library and a flow-sorted chromosome 17-specific cosmid library, respectively (kind gift from Genome Analysis Laboratory). Five cosmids were isolated using the \textit{MB1} cDNA as a hybridisation probe (MB2-6) and three cosmids were isolated using the \textit{delta} cDNA (D1-3). The two sets of cosmids were first analysed by restriction enzyme digestion. The cosmids showed similar banding patterns on EcoRI digestion indicating that both sets overlapped each other. Initial work was carried out on the \textit{MB1}-associated cosmids. Since all the \textit{MB1}-associated cosmids overlapped each other further work was carried out on the longer cosmid, MB4. MB4 had an insert of \textasciitilde39 kb\textsuperscript{2} in size compared to insert sizes of between 34-37 kb for the remaining cosmids.

\textbf{6.2.2. Isolation of cDNA clones with the MB4 cosmid insert}

The complete MB4 cosmid insert was labelled with \textsuperscript{32}P-dCTP, competed with human placental DNA to remove repetitive sequences and hybridised onto a B

-164-
cell cDNA library (JY) under stringent conditions. The final washing conditions were 0.2X SSC, 65°C. The filters were exposed to film and then developed, showing many positive colonies. The filters were immediately re-hybridised with the full length MB1 cDNA under identical conditions. Any positive colonies that hybridised to the MB4 cosmid but not the MB1 cDNA would be novel cDNAs that were located within ~40 kb of the MB1 gene on chromosome 14.

Thirteen positive clones were identified in this way which were further characterised. Comparison of insert sizes enabled the clones to be divided into three groups. Seven clones had an insert size of ~1 kb, 3 clones had an insert size of ~1.8 kb and 3 clones had an insert size of ~2.3 kb.

6.2.3 Partial nucleotide sequence of cDNA clones encoded within the cosmid MB4

The two ends of the cDNA clones isolated from the JY cDNA library, described above, were sequenced using primers complementary to sequences either side of the cloning site of the cDNA vector, CDM8. Sequencing confirmed that the clones could be divided into three groups, each group containing clones of identical size and end sequence. Hence, further work was carried out on one clone from each group (M11: 1.8 kb, M14: 1 kb and M15: 2.3 kb). One end of the clones M14 and M15 (primed with the oligonucleotide CDM8-F) contained a long tract of A residues, characteristic of a poly (A) tail. Sequence from the clone M15 could not be read further, presumably due to secondary structure created by the homopolymeric tail. The nucleotide sequence of the other end of all the clones, presumably near the 5' end of the genes, was obtained by priming the sequencing reaction with the oligonucleotide CDM8-B. This nucleotide sequence from the three clones was used to search the EMBL sequence database. No significant matches were found with the cDNAs M14 and M15 but cDNA clone M11 was found to contain Alu repetitive sequences (discussed in section 6.2.5). The initial sequence was also directly compared with the sequence of TAPI and TAP2 but no significant homology was found at either the DNA or protein level (data not shown). It cannot be ruled out that these are TAP-related genes, especially if the only regions to have been sequenced were untranslated regions. Also, the sequences of the new clones may have significantly diverged from primordial sequences necessitating the acquisition of further sequence data.
6.2.4. Initial mapping of the cDNA clones M11, M14 and M15

Localisation of the cDNA clones was achieved in two ways: 1) hybridising all the clones to each other and 2) mapping the clones to the cosmid MB4 to ascertain if they mapped to similar restriction enzyme fragments. In the first approach, Southern blots were made from the three representative cDNA clones and each clone hybridised to the other two. In every case the relevant clone hybridised to itself but not to the two other clones, indicating that the three clones represented three different cDNAs. In the second approach, the cosmid MB4 was digested with a range of restriction enzymes (double and single digests), the digested DNA blotted and hybridised against the three cDNAs. Representative examples of the hybridisation patterns thus produced are shown in figures 6.1 and 6.2. The cDNAs all mapped to a 12 kb BgIII band from the cosmid MB4. Hybridisation of the complete MBl cDNA to a similar cosmid blot showed that MBl also mapped to this 12 kb BgIII band (figure 6.3). Thus, the four cDNAs (including MBl) were most likely located in close proximity to each other in a gene cluster.

6.2.5. Genomic sequencing of cosmid MB4

At this point the most efficient method of obtaining more data on the function of the cDNAs and their spatial relationship would be to obtain the complete genomic sequence of the 12 kb BgIII band from cosmid MB4. This would also provide information on the genomic structure of all the genes, including MBl. The 12 kb BgIII restriction band was therefore self ligated, sonicated and subcloned into the vector M13mp18. The resulting constructs were transformed into E. coli and sample sequenced by the DNA Sequencing Laboratory. Sample sequencing provides an extensive but incomplete sequence (~90%) of the region analysed.

Analysis of the sequence data showed that one of the clones (M11) was entirely composed of Alu repeat sequences. Hence, no further analysis was carried out on this cDNA clone.

Comparison of the partial cDNA sequences from M14 and M15 (229 and 193 bp, respectively) to the genomic sequence demonstrated complete homology (figures 6.4 and 6.5). Sample sequencing produced seven stretches of contiguous DNA sequence, contigs 1-7, which were approximately 600, 1500, 2000, 2500, 200, 2600 and 1300 bp in size. This resulted in 87% sequence coverage of the 12 kb fragment from cosmid MB4. The Alu sequence (clone M11) spanned
Figure 6.1. Hybridisation of cDNA clone M14 to a Southern blot containing cosmid MB4 which had been digested with the restriction enzymes BamHI, HindIII, BamHI/HindIII, PstI, BgIII and NotI (double and single digests). Final washing conditions were 0.2X SSC, 65°C. The filter was exposed to Kodak X-AR autoradiography film for 30 mins at room temperature. Longer exposure times showed positive hybridisation signals in all lanes. M14 hybridises to a number of different restriction bands, including a 12 kb BgIII band.
**Figure 6.2.** Hybridisation of cDNA clone M15 to a Southern blot of cosmid MB4 which had been digested with the restriction enzymes BamHI, BamHI/HindIII, HindIII, PstI, EcoRI, ClaI/EcoRI, NotI, NotI/SalI and SalI (single and double digests). Final washing conditions were 0.2X SSC, 65°C. The filter was exposed to Kodak X-AR autoradiography film for 30 mins at room temperature. Longer exposure times showed positive hybridisation signals in all lanes. M15 hybridises to a number of different restriction bands, including a 12 kb BgIII band.
Figure 6.3. Hybridisation of complete MB1 cDNA to a Southern blot of cosmid MB4 which had been digested with the restriction enzymes BgIII, BamHI, BamHI/HindIII, HindIII, Clal, Clal/EcoRI, EcoRI, NotI, NotI/Sall and Sall (single and double digests). Final washing conditions were 0.2X SSC, 65°C. The filter was exposed to Kodak X-AR autoradiography film for 30 mins at room temperature. Longer exposure times showed strong hybridisation signals in all lanes. MB1 hybridises to a number of different restriction bands, including a 12 kb BgIII band.
Figure 6.4. Partial nucleotide sequence of the 3' end of cDNA clone M14 which matches the genomic sequence, obtained from cosmid MB4. The predicted amino acid sequence is shown which allows one open reading frame (ORF), numbered 1. The poly (A) addition signal AACAAA is underlined. No significant matches of M14 with database sequences at either the amino acid or nucleotide levels were found.
Figure 6.5. Partial nucleotide sequence of the presumed 5' end of cDNA clone M15 which matches the genomic sequence, obtained from cosmid MB4. The predicted amino acid sequence is shown allowing one open reading frame (ORF), numbered 1. Two methionine codons within the ORF are shown in bold. However, neither of these methionine residues are in a good context for initiation of transcription. No significant matches of M15 with database sequences at either the amino acid or nucleotide levels were found.
approximately 2 kb of genomic sequence in contig 3. The 3' end of cDNA M14 was contained within contig 1. The partial nucleotide sequence and predicted 6 frame amino acid sequences of M14 are shown in figure 6.4. One ORF is possible, numbered 1. The sequence AACAAA (underlined), which resembles the consensus poly (A) addition signal AATAAA, started 17 nucleotides upstream of the first 9 A residues. In a search of nucleotide sequence databases, the AACAAA variant was detected in 0.8% of vertebrate cDNAs, and therefore may be a naturally occurring poly (A) addition signal.

The presumed 5' end of cDNA M15 was contained within contig 2. The M15 partial nucleotide sequence and predicted 6 frame amino acid sequences are shown in figure 6.5. One ORF is possible, numbered 1. Two methionine codons within the open reading frame are also shown. However, neither of these methionine residues are within a good context for initiation of translation. The first methionine residue is in the context AGCTCATGC and the second methionine is in the context AATCCATGA, both of which differ from the consensus translation initiation signal CCACCATGG (Kozak, 1986). It is possible that these methionine residues are in a long 5' untranslated region and acquisition of more cDNA sequence will determine the position of the translation initiation signal of M15. It is not possible to tell at this stage if the cDNA clone is that encoding the M15 protein product. This question will be resolved with the solution of the complete nucleotide sequence of M15.

The predicted nucleotide and amino acid sequence from contigs 1 and 2 (containing M14 and M15 cDNA sequences) were used to search the EMBL and Swiss-Protein sequence databases but no significant matches were found. Therefore, it seems likely that M14 and M15 cDNAs are unrelated to the TAP transporters. The cDNAs code for proteins with unknown functions.

6.2.7. Genomic organisation of MB1

The partial intron/exon structure of MB1 was determined from comparison of the complete MB1 cDNA sequence with the genomic sequence contained in contigs 5, 6 and 7 (figure 6.6). This sequence analysis showed the organisation of three exons (exons 1, 2 and 3) which are 33 bp, 306 bp and 283 bp in size, respectively, and are separated by two introns of >510 bp and >2.2 kb. As the genomic sequence of MB1 is contained within three different contigs it is not possible to determine an accurate size of the introns of MB1. However, it is likely that the
**Figure 6.6.** Comparison of the genomic organisation of human \textit{LMP2, LMP7} and mouse \textit{delta} genes with the human \textit{MB1} gene. All transcripts are shown in a 5’ to 3’ orientation.
gaps in the genomic sequence between contigs 5 and 6, and between contigs 6 and 7 are less than 200 bp (S. Beck, personal communication). Hence, it is probable that intron 1 is 510-710 bp in size and intron 2 is 2220-2420 bp. Therefore, MB1 spans approximately 4 kb of genomic sequence. Both introns 1 and 2 are full of Alu repetitive sequences which could explain why it has been difficult to obtain complete sequence coverage of MB1 at the genomic level (S. Beck, personal communication). Analysis of the MB1 cDNA sequence revealed that the cDNA contains a 5' untranslated region of 121 bp.

Comparison of the genomic organisation of MB1 with that of LMP2, LMP7 and mouse delta showed that MB1 differs significantly from the other proteasome genes at the genomic level (figure 6.6). The genomic organisation of exons and introns in human LMP2, human LMP7 and mouse delta are consistent with each other. All three genes are made up of six exons and five introns, although the sizes and locations of the introns are variable. Human LMP2 is closely related to human delta, showing 59% identity at the amino acid level. Comparison of the intron/exon organisation of the LMP2 gene with mouse delta supports this close relationship. The human delta gene is most likely to have a similar exon/intron organisation to that of LMP2, since the equivalent mouse gene has this pattern. Human MB1 and LMP7 genes are also closely related, showing 67% identity at the amino acid level. It would be expected that this relationship would also be echoed in the genomic organisation of the two genes. However, this is not the case. An MB1-related sequence is thought to have duplicated to produce LMP7 (Belich et al., 1994) which was subsequently brought together with LMP2 and the TAPs in the MHC. Since the intron/exon organisation of MB1 and LMP7 differs so widely it is likely that MB1 and LMP7 have been separated for a long evolutionary time or the altered gene structures have functional or regulatory consequences.

6.3. Conclusions

Analysis of the molecular anatomy of genes and gene families has been used to provide answers concerning the evolution of such genes. MHC class I and class II molecules, for example, show clear homology to immunoglobulin domains (Orr et al., 1979) and therefore belong to the immunoglobulin superfamily (Williams and Barclay, 1989). Exactly when MHC molecules first evolved remains unclear, however similar genes have been reported in amphibians, fish, reptiles and birds.
This implies that they were present prior to vertebrate evolution which occurred about 400 million years ago (Kaufman et al., 1990). The primordial gene for MHC antigens is thought to have encoded a single molecule similar to a class II gene which associated to form heterodimers. Duplication and divergence led to a multigene family encoding heterodimers like present-day class II antigens. Eventually a gene coding for a chain with three extracellular domains was formed and associated with β2-m giving rise to a class I-like molecule. Duplication and divergence of this gene led to the multiple class I loci in the MHC (Kaufman et al., 1984; Hughes and Nei, 1993).

Nucleotide and amino acid analysis of class II genes shows that all the α-chain genes are equally diverged from one another suggesting that they arose by duplication at roughly the same time (Auffray et al., 1984). Similar studies performed on β-chain sequences suggest that DPB1, DQB1 and DRB are equally related and that DOB may have diverged prior to events that gave rise to the other β-chain loci (Tonnelle et al., 1985). More recent duplication events are responsible for the highly related genes DQA1 and DQA2, DQB1 and DQB2, DPA1 and DPA2 and DPB1 and DPB2.

The TAP and LMP genes in the MHC class II region represent a gene cluster where the protein products are involved in the class I antigen processing pathway. Belich et al. (1994) have proposed a model to explain the genetic origins of the LMP genes, discussed in section 6.1.4. This chapter has described a preliminary study aiming to determine how the TAP/LMP gene cluster came to reside in the class II region.

Cosmids encompassing the LMP7-related gene, MB1, were isolated and used to identify novel cDNAs in close proximity to the MB1 gene. Three cDNAs were isolated and mapped by hybridisation to cosmid blots. The four cDNAs (including MB1) were all localised to a 12 kb BglIII restriction fragment, their close proximity hinting at the presence of a gene cluster. Genomic sequencing of the 12 kb BglIII restriction fragment was carried out and showed that one of the clones was composed of Alu repeat sequences. The genomic sequence and partial cDNA sequences of the remaining two cDNA clones were used to search the EMBL and Swiss-Protein databases but no significant matches were found. The cDNA clones mapping close to MB1 are, therefore, unrelated to the TAP transporters. Genomic sequencing also revealed that the 5' end of MB1 is located 81 bp from a BglIII restriction enzyme site. If the 12 kb BglIII fragment, encoding the MB1 gene and the cDNA clones M11, M14 and M15, is located at one end of
cosmid MB4 it is still possible that TAP-related sequences reside close to MBl. Further restriction enzyme mapping and sequencing of cosmid MB4 will determine if this the case.

A tight cluster of genes that are apparently unrelated by function has been discovered on chromosome 16q22.1 (Larsen et al., 1993). The genes for a protein serine kinase (PSKH1), the previously cloned lecithin: cholesterol acyl transferase (LCAT), a protein of unknown function and the proteasome subunit, MECL-1 are located in a 12 kb genomic region which includes a CpG island. The expression of MECL-1 is inducible with \(\gamma\)-IFN and has reciprocal expression to that of the proteasome subunit Z (Tanaka, 1995). It has been suggested that \(\gamma\)-IFN may induce the subunit replacement of Z by MECL-1 (along with the replacement of delta and MB1 by LMP2 and LMP7, respectively), producing proteasomes that are more appropriate for antigen processing through the class I pathway (Tanaka, 1995). The cluster of genes on chromosome 16 includes genes with restricted tissue expression patterns and widely expressed genes. The tight clustering of these genes suggests that they may be subject to reciprocal transcriptional regulation (Larsen et al., 1993). Further genomic sequence analysis of the MB4 cosmid, on chromosome 14, will determine if MB1 exists in such a gene cluster.

Comparison of the MB1 cDNA sequence with the genomic sequence enabled determination of the MB1 intron/exon organisation. MB1 is composed of three exons which are 33, 306 and 283 bp in size. Introns 1 and 2 are \(>510\) bp and 2.2 kb in size, respectively. This organisation is widely different to that of LMP7 and other proteasome genes. Gene duplication, a major mechanism in evolution, is responsible for the existence of all present-day genes (Ohno, 1970). For example, the chicken ovomucoid gene is thought to have evolved by triplication of an ancestral protein-coding DNA segment split by one intron. It is now composed of seven introns and eight exons (Breathnach and Chambon, 1981). There is also evidence that the conalbumin gene, composed of sixteen introns and seventeen exons, evolved by duplication from an ancestral gene with seven or eight exons (reviewed in Breathnach and Chambon, 1981). By analogy, MB1 probably duplicated to produce LMP7 which is composed of six exons and five introns. From this data it is impossible to comment on when the duplication event occurred.

Figure 6.6 shows the intron/exon structure of mouse delta which is similar to that of human LMP2. Since human and mouse class II genes are related both in function and genomic organisation it is to be expected that human delta will have
a similar molecular anatomy to that of mouse *delta* and hence human *LMP2*. This is in contrast to that of human *MB1* and *LMP7*. The reasons for this are unclear. However, determination of the genomic organisation of human delta may help to understand the evolution of the proteasome genes.
CHAPTER 7: CONCLUDING DISCUSSION

7.1. Genomic analysis of the human MHC class II region

Recent studies using the powerful techniques of reverse genetics have led to the discovery of numerous novel genes in the human MHC. In the class II region these fall into three broad categories: 1) genes with no apparent association with the immune system. Examples of these include \textit{RING1} which encodes a novel zinc finger motif (Freemont et al., 1991) and \textit{RING3} which encodes a protein homologous to the \textit{Drosophila} maternal effect gene, \textit{Female sterile homeotic} or \textit{Fsh} gene (Beck et al., 1992a). The significance of the location of these genes in the MHC, if any, remains unclear. 2) Novel class II genes, \textit{DMA} and \textit{DMB}, which are distantly related to the other major class II loci and associate to form a heterodimer. The DM molecule is thought to aid peptide binding to class II molecules in the class II antigen processing pathway (Sanderson et al., 1994). 3) The \textit{TAPs} and \textit{LMPs} which are involved in the class I antigen processing pathway. Antigenic proteins are degraded in the cytoplasm by the \textit{LMP}-encoding proteasome. Peptide products then traverse the ER membrane via the \textit{TAP} complex and associate with MHC class I molecules ready for presentation to T cells.

Genomic analysis has proved to be a highly successful approach in gaining insight into both the class I and class II antigen processing and presentation pathways. The preceding chapters have described the application of molecular mapping and cloning techniques in the class II region resulting in the completion of the class II cosmid map. Six expressed transcripts and two exons were identified and characterised. The general significance of this work in the context of advances in MHC mapping is discussed below along with suggestions for future experiments.

Recently, efforts to characterise the human genome and advances in mapping human genetic disorders have led to an interest in techniques for identifying transcribed sequences from specific chromosomes and sub-chromosomal regions. The human genome consists of 3 X 10^9 base pairs, but only 1-2% of this is expressed as mRNA in any given tissue (Hastie and Bishop, 1976). Although there are no reliable data on the total number of structural genes, estimates range from 5 X 10^4 - 1 X 10^5 (Bishop, 1974). Whilst defining the regulatory and structural features of the genome will be necessary for a complete
understanding of the genome, identification of transcribed sequences and their products is of immediate interest.

Identification of such transcribed sequences will allow efforts to sequence the genome to concentrate on expressed regions and will aid in analysis of these sequences by providing information about their expression patterns. In addition, transcribed sequences can be positioned on the continuously refined genetic and physical maps of human chromosomes, leading to an integrated map of the human genome which not only includes the genetic and physical location of all genes but also their expression patterns. Such integrated maps would provide information on the organisation of the human genome in terms of gene number, expression patterns and the clustering of genes. Finally, transcript identification from a particular chromosome or region will greatly facilitate the positional candidate approach for isolating genes responsible for inherited human diseases.

The "positional candidate" approach used in disease gene isolation assumes no functional information and relies on a combination of mapping the disease gene of interest to the correct chromosomal subregion followed by a survey of the interval to see if attractive candidates reside there (reviewed in Collins, 1995). An increasingly dense transcript map will greatly aid this approach, compared to pure "positional cloning" which locates the disease gene purely on the basis of its map position. Positional candidate cloning has located the genes for Marfan syndrome (Dietz et al., 1991), Hirschsprung's disease (Edery et al., 1994), and hereditary non-polyposis colon cancer (Papadopoulos et al., 1994; Leach et al., 1993) amongst others. It has been predicted that more than half of the human transcripts will be placed on the human genome map in the next 18 months, mainly by positional candidate cloning (Collins, 1995).

The MHC spans 4 Mb of DNA and represents 1/750th of human genetic material. As such, the MHC is a paradigm for the detailed molecular organisation of the human genome. The intensive efforts to map, clone and sequence the MHC genes have made the MHC one of the best studied regions of the human genome. The detailed maps which are available for the human MHC and the progress made towards cloning the entire region make the MHC an obvious starting point for large scale genomic sequencing, envisioned by the Human Genome Project. At present 300 kb of the MHC class II region has been sequenced and projects are underway to sequence the class I and III regions.

Initially, the MHC was mapped by PFGE. This technique allowed accurate sizing of the MHC, 3.8 Mb, and established the orientation of the complement and
21-hydroxylase gene loci relative to the class I and class II loci (Dunham et al., 1987). Detailed analysis of the MHC class II region has been achieved by cloning the region in YACs and cosmid vectors. Cosmids were isolated from the class II -DP, -DQ and -DR subregions but it was not possible, at first, to link all the DRB genes with each other or to DRA (Spies et al., 1985; Rollini et al., 1985). The DQ region was cloned (Blanck and Strominger, 1988), but was not linked to DR or the DNA gene (Blanck and Strominger, 1990). Similarly, the DNA gene had not been linked to the DP subregion. To clone the entire class II region, a 4X YAC library was constructed (Ragoussis et al., 1991) resulting in a contig over 1.5 Mb long which extended between the DRA gene in the class II region and the C4 gene in the class III region. The gaps between DRA and DRB, DRB and DQA and DOB and DPA were thus bridged with YAC clones.

However, a major problem associated with analysing YAC clones is the inability to isolate purified cloned DNA in large quantities, due to the absence of amplification methods specific for the recombinant chromosome. For applications such as cDNA library screening, cosmid clones are easier to manipulate and utilise.

At the start of this project, two areas of the class II region were uncloned in cosmid vectors. 1) The 40 kb region between the DMB and LMP2 genes and 2) 24 kb of the region between the genes DPAI and DNA. The region telomeric of DMB was of special interest as it was shown, by PFGE, to contain a NotI site which is characteristic of a CpG island. CpG islands are found to colocalise with the 5' ends of expressed sequences (Bird, 1987). In addition, the DMB to LMP2 gap was implicated at the start of this project as a candidate region for (a) gene(s) involved in the class II antigen processing pathway. However, during the course of this study, the class II antigen processing genes were identified as being DMA and DMB (Morris et al., 1994; Fling et al., 1994).

The DMB-LMP2 region was cloned in three overlapping cosmids: N3, A12 and A15. N3 was isolated from an enriched cosmid library constructed from the YAC 11.2 which spans the MHC class II region from DMB to DOB. A12 and A15 were isolated from the flow-sorted chromosome 6 cosmid library. In parallel with these approaches, screening of a P1 phage library resulted in the isolation of a P1 clone, DP2, which mapped between the DMA and DOB genes.

The order of the three new cosmids was confirmed by fluorescence in situ hybridisation (FISH). FISH is the technique of depositing fluorescent molecules in chromatin at the sites of specific DNA sequences (Trask, 1991). Different
combinations of cosmids were fluorescently labelled and hybridised to chromatin released from fixed cells. Regions of overlap between two probes, one detected in red and the other in green, appeared as signal with a yellow colour when viewed with a dual band pass filter on a microscope (Senger et al., 1993).

Multicolour FISH offers a fast, simple approach for probe ordering. Three probes can be mapped simultaneously: one detected in red, one in green and a third in yellow. The technique can thus provide a rapid map of three different cosmid clones in one experiment. Such FISH mapping depends on the assumption that chromatin is randomly folded after fixation and hybridisation, which may not always be the case (Senger et al., 1993). Hence, to obtain accurate sizing of cosmid overlaps, FISH should be used in conjunction with other mapping techniques such as restriction enzyme digestion and hybridisation studies.

The remaining gap in the class II cosmid map, part of the region between the genes $DPA_1$ and $DNA$, was cloned in one cosmid (A1) using the flow-sorted chromosome 6 cosmid library. This gap was easier to clone in comparison with the $DMB$ to $LMP2$ region which was probably due to its reduced size (24 kb compared to 55 kb) and the presence of fewer repetitive sequences. Cloning the $DMB$ to $LMP2$ region in cosmid vectors showed that cosmid libraries rarely contain clones spanning this region. Only one cosmid (N3) was isolated from the enriched cosmid library which contained novel DNA. Two clones (A12 and A15) were isolated from the flow-sorted chromosome 6 cosmid library which spanned the uncloned region. The other eight cosmids isolated from this library were very similar to either cosmid HA14 (Blanck and Strominger, 1990) or cosmid U15 (Blanck and Strominger, 1988). The difficulty in cloning the $DMB$ to $LMP2$ region and under-representation of the area in cosmid libraries could be due to the presence of multiple $Alu$ repeat sequences (Beck et al., 1992b).

Once the class II region had been completely cloned in cosmids, molecular mapping techniques were utilised to identify and characterise novel transcribed sequences within this region. Many methods are available to screen for coding sequences in cloned genomic material including identification of CpG islands which associate with the 5' ends of expressed sequences (Bird, 1987; Hanson et al., 1991); cDNA selection (Korn et al., 1992); island rescue PCR (Valdes et al., 1994); Zooblot analysis (Monaco et al., 1986); cDNA library screening with genomic fragments (Elvin et al., 1990); exon amplification (Buckler et al., 1991) and direct genomic sequencing (Beck et al., 1992b). The last three approaches were utilised in the course of this study, the main method involving screening
cDNA libraries with cosmid fragments. This approach had proved highly successful in both the MHC class II and class III regions (reviewed in Milner and Campbell, 1992).

These techniques resulted in the identification of a phosphatase inhibitor pseudogene, mapping of a retinoic acid receptor β gene, isolation of six expressed transcripts, and identification of two exons in the MHC class II region.

**Phosphatase inhibitor pseudogene, IPP-2**

A 1.3 kb cDNA was isolated from a monocyte/macrophage cDNA library by cosmid screening. This cDNA showed 94% nucleotide sequence identity with rabbit phosphatase inhibitor-2 (IPP-2). IPP-2 is thought to be involved in glycogen metabolism (Dent et al., 1990), muscle contraction (MacDougall et al., 1991) and mitosis (Axton et al., 1990). The human IPP-2 sequence in the class II region, found during this work, was shown to be a pseudogene by comparison of the cDNA and genomic sequences. These differed by 10% and 12% at the nucleotide and amino acid levels, respectively. FISH indicated that there are three IPP-2 sequences in the human genome on chromosomes 3q29, 5q33 and 6p21.31. The gene itself is thus located on either chromosome 3 or 5, the other two sequences due to pseudogenes.

**Mapping human retinoid X receptor β gene (hRXRβ)**

Initial studies indicated that hRXRβ mapped onto the short arm of chromosome 6 (Fleischhauer et al., 1993). Since the murine RXRβ had been mapped to chromosome 17, which contains the murine MHC (Hamada et al., 1989), and is known to activate transcription of MHC class I genes in response to retinoic acid, it was of interest to accurately map hRXRβ.

This was achieved by hybridising the human cDNA to YACs and cosmids spanning the MHC region and mapped hRXRβ to a locus 3 kb centromeric of KE5 in the class II region. Retinoic acid treatment of human embryonal carcinoma cells activates a RXRβ/retinoic acid receptor heterodimer resulting in induction of MHC class I gene expression (Segars et al., 1993). It is of great interest that a gene involved in activating the expression of MHC class I molecules is located centromeric of the MHC class II region. The significance of this localisation is...
unclear but co-ordinate regulation of genes with associated functions like these, hRXRβ and class I, may present a selective advantage.

Expressed transcripts in the class II region

Three overlapping cDNAs which are located 10 kb centromeric of LMP2 were isolated from different cDNA libraries in the course of this study. The cDNAs are composed of two exons of varying lengths and one intron (5962 bp), contain no poly A+ tail, and are not expressed on northern blots. It is likely that these transcripts result from aberrant mRNA transcription in cDNA libraries.

Three unspliced contiguous genomic fragments were also isolated from cDNA libraries. SUC1 is 2.5 kb in size and is located centromeric of LMP2. GEN1 is 3 kb in size and maps precisely to 3 kb of genomic sequence centromeric of TAP2, which is flanked by Alu repeat sequences. S4 is 450 bp in size and is located 12 kb centromeric of LMP2. These fragments have no intron/exon structure, no obvious initiation codons and no polyadenylation signals.

Similar non-functional transcripts have been reported for the pro-opiomelanocortin (POMC) gene, the vasopressin gene, the oncogene pim-1 and the oncogene c-mos (reviewed in Ivell, 1992). It is thought that these aberrant transcripts may be examples of leaky transcription resulting from spurious transcriptional initiation. It is likely that S1, S2, S3, S4, GEN1, RING8 and SUC1 represent examples of such transcription. From these studies it seems that transcription of "non-gene" regions of the genome occurs once every 150-200 kb in the MHC class II region. The significance of this is unknown but acquisition of further sequence data may enable the identification of any shared characteristics in or near these non-coding transcripts causing their transcription.

Exons HA14-2 and HA14-4

Exon amplification was used in conjunction with cosmid screening to identify novel transcripts in the MHC class II region. As exon amplification is independent of tissue expression patterns it was hoped that this technique would identify any transcripts missed by the screening approach.

Two exons, HA14-2 and HA14-4, were isolated from the cosmid HA14 and mapped centromeric of the DMB gene. Zooblot analysis showed that exon HA14-4 was highly conserved with five copies present in the chicken genome. Thus,
HA14-4 has been present in DNA before separation of the mammalian and avian lineages which occurred about 300 million years ago (Kroemer et al., 1990). Exon HA14-2 was not conserved and is possibly a false positive generated by exon amplification. Exon amplification of well-characterised cosmids containing genes mapping to the MHC class II region revealed that false positives are generated at a frequency of approximately one false positive for 80 kb of DNA analysed. However, this rate of false positive generation may be specific for the MHC class II region. Exon HA14-4 maps 364 bp centromeric of the 5' end of DMB and, due to this proximity, could be a differentially spliced product of DMB. PCR reactions carried out on genomic DNA using oligonucleotide primers designed from the sequence of exon HA14-4, did not detect HA14-4 expression in brain, testis, osteosarcoma cells, HeLa cells, mammary gland, foetal liver, placenta or fibroblasts.

Southern blot analysis revealed that exon HA14-4 maps to an equivalent position in the human and mouse class II regions. Precise location of HA14-4 with respect to the mouse gene Mb (homologous to human DMB) may help to determine if the exon is encoded by a differentially spliced product of DMB. If exon HA14-4 is located close to the mouse Mb gene, then this will provide circumstantial evidence that HA14-4 is encoded by a differentially spliced product of Mb (human DMB). Finally, isolation of this exon from different mouse tissues at various stages of development would help to discover if HA14-4 is developmentally regulated.

Sequencing

The third approach in this project to study the genomic organisation of the MHC class II region was that of genomic sequencing. This was carried out in collaboration with Dr. Stephan Beck (DNA Sequencing Laboratory, ICRF). At present 300 kb of class II genomic sequence has been completed. All the sequence data has been sent to the NIH for analysis by the GRAIL program which is a computational approach for locating protein-coding gene portions in anonymous DNA sequences (Uberbacher and Mural, 1991). This program predicted the presence of three novel exons in the class II region. One exon corresponded to HA14-4 which is possibly part of a differentially spliced DMB gene. The second exon was predicted to overlap the transcripts S2 and S3, centromeric of LMP2. Comparison of the cDNA and genomic sequence revealed that these transcripts
are composed of two exons and one intron, but they are not expressed on northern blots. The third exon was predicted to cover 160 bp of the cDNA S4. This transcript consists of 403 bp of contiguous genomic sequence with no intron/exon structure. PCR probes derived from the latter two predicted exons and screened onto B and T cell cDNA libraries identified only clones identical to S2, S3 and S4 which are thought to derive from aberrant transcription. Overall, in the MHC class II region, the GRAIL program was highly successful in predicting the presence of exons. Of the known genes, only \textit{LMP2} was missed, since its exons are smaller than 80 bp.

Genomic sequence analysis also allowed the identification of an 88 bp stretch of sequence centromeric of \textit{LMP2} that showed > 60% nucleotide sequence identity with class I sequences. A PCR probe was derived from the class I-like sequence and used to screen B and T cell cDNA libraries in the search for an expressed transcript. No such transcript was isolated. This class I-related sequence could have been inserted between the \textit{DMB} and \textit{LMP2} genes on a fragment from the class I region. Genomic sequencing revealed the presence of two LINE repeat sequences close to the \textit{DMB} and \textit{LMP2} genes which could have directed such a recombination event (Beck et al., 1992b).

In the mouse, the \textit{DMB to DOB} region (\textit{Mb to Ob}) is approximately half the size of the equivalent human region whereas the size of the \textit{LMP/TAP} gene cluster is conserved between the two species. This provides further indication that an insertion event has occurred in the human class II region. At present it is not possible to comment on the origin of this tract of DNA but I hypothesise that it is from the class I region.

\textbf{Gene density in the class II region}

Cosmid screening, exon amplification and genomic sequencing have revealed that, in the MHC class II region, one gene occurs approximately every 40 kb of DNA. This is in contrast to the class III region where the gene density is much greater, at least one gene every 15 kb (Campbell and Trowsdale, 1993). Initial studies on YACs spanning the class I region estimates the presence of a gene every 20 kb (Geraghty, 1993). This difference in gene density could be related to chromosome bands. Gene-dense Giemsa-dark G bands are composed of AT-rich sequences, and relatively gene-sparse Giemsa pale R bands mainly of GC-rich sequences (Craig and Bickmore, 1993). Class II sequences are AT-rich whilst
class III sequences are GC-rich (Fukagawa et al., 1995). The transition between the AT-rich and GC-rich domains occurs in the class II/III boundary region, approximately 180 kb from DRA. This AT-rich/GC-rich transition may explain the difference in gene density between the class II and III regions.

The difference in gene density between the MHC class II and III regions may also be a consequence of co-ordinate gene regulation. The class II region is composed of genes with functions in the immune system whereas the class III region contains many genes with different functions. Regulation of the class II loci as a unit may enable an individual to respond to a broader range of antigenic challenges. Co-ordinate regulation may limit the gene density in the region under such control (Kelly and Trowsdale, 1994). However, these questions will only be answered when a complete picture of the genomic organisation of the MHC, including gene density, gene function and genomic sequence has been attained.

Several projects are underway, at present, to achieve this goal. Genomic sequencing may aid in the identification of disease genes linked to the MHC. For example, the sleep disorder narcolepsy is associated with the class II -DR subregion (Aldrich, 1990). DR and DQ cDNAs from narcoleptic and normal individuals have been sequenced and found to be identical, leading to the assumption that the gene involved is unlikely to be DR or DQ but may located close to DR. The DR region has not been as intensively studied as the other class II subregions. Use of the flow-sorted chromosome 6-specific cosmid library will allow the cloning of specific DR haplotypes which differ in their gene organisation. Experiments are underway, in the Human Immunogenetics Laboratory in collaboration with Dr. Duncan Campbell (Dept. of Biochemistry, University of Oxford, UK), to clone the DR region in cosmids isolated from a cosmid library with a DR2 haplotype. This haplotype was chosen as it most commonly associated with narcolepsy (Aldrich, 1990). These cosmids will be mapped and sequenced allowing identification of novel genes in the DR region. Information will also be generated concerning gene structure, repeat sequences and regulatory sequences. Knowledge gained from the complete characterisation of the MHC will be invaluable to the Human Genome Project as well as providing information concerning antigen processing and presentation and MHC disease associations.

The genetic and physical data relating to the MHC have been compiled into a public database using a genome analysis program (MHCDB: Newell et al., 1994). The current contents of the database include the location of over 250 YAC
and cosmid clones, and over 100 genes, 150 kb of genomic DNA sequence and
cDNA sequences of currently known class I and class II alleles. This provides a
useful tool for the study of the human MHC and can be extended to analyse the
entire chromosome 6.

7.2. Isolation of genes linked to the LMP-related proteasome gene,
MBI

The LMP/TAP gene cluster

Colocalisation of genes with related functions is a feature of the MHCs of
numerous species and many of the MHC loci appear to be the result of past gene
duplication events. Theories have been proposed to explain this maintenance of
linkage (Kelly and Trowsdale, 1994). 1) The MHC genes may be co-ordinately
regulated. 2) The MHC genes may be maintained together to allow co-evolution
of function. The close association of polymorphic proteins may be advantageous
in that, once formed, beneficial associations of alleles are more frequently kept
together by linkage disequilibrium. 3) Clustering of genes within the MHC may
simply suggest a common, fortuitous ancestry.

Considering this last point, the mouse surfeit locus contains six
housekeeping genes and has been conserved over 600 million years of divergent
evolution (Colombo et al., 1992). An additional gene cluster has been discovered
on chromosome 16 and contains five genes with unrelated functions spanning 40
kb (Larsen et al., 1993). One of these genes encodes the proteasome subunit,
MECL-1. A possible explanation for such tight gene associations is that once a
cluster has been established it is effectively locked into the genome, unable to
separate because of small distances between the genes.

TAP1, TAP2, LMP2 and LMP7 form a tightly linked cluster of IFNγ-
inducible genes that function in the class I antigen processing and presentation
pathway. Their residence in the MHC may therefore reflect a common function.
LMP2 and LMP7 are thought to have evolved from primordial delta and MBI-
related genes, respectively (Belich et al., 1994). It will be interesting to discover if
the TAP genes also evolved from two separate primordial genes or a gene pair.
The LMP and TAP genes may have been brought together in the MHC to form a
gene cluster containing genes with related functions. If this was the case, the
primordial TAP genes may be located close to the primordial LMP genes.
A preliminary study was initiated aiming to discover how the LMP/TAP gene cluster had come to reside in the class II region. Three cDNAs were isolated which clustered within 12 kb of MB1 on chromosome 14. One of the cDNAs was entirely composed of Alu repeat sequences. Initial sequence analysis of the two remaining clones showed no homology with genes in the Swiss-Protein or EMBL databases. It is possible that these sequences were too limited to show significant homology with other sequences, specifically those corresponding to TAPI and TAP2, present in the two sequence databases. Therefore, genomic sequencing of the cosmid fragment containing MB1 and the two associated genes was performed. Again, no significant matches to sequences in the Swiss-Protein or EMBL databases were found. The data at present are consistent with the TAP transporters having evolved independently to the LMPs in different locations of the human genome. The LMPs and TAPs may have been brought separately into the MHC to form a gene cluster although how this occurred remains obscure. However, the 5' end of the MB1 gene is located close to one end of the 12kb BglII restriction enzyme fragment from cosmid MB4. If this fragment is located at one end of the MB4 cosmid insert, it is still possible that TAP-related transcripts map close the MB1 gene. More extensive sequence analysis of the genomic DNA encompassing the MB1 gene could reveal TAP-related transcripts.

Genomic sequencing enabled determination of the intron/exon organisation of the MB1 gene. MB1 is composed of three exons which are 33, 306 and 283 bp in size. These are divided by two introns of >190 bp and >2.2 kb. Comparison of MB1 and LMP7 genomic organisations revealed that MB1 is unlikely to have duplicated recently to produce LMP7 which is composed of six exons and five introns. At present it is not possible to state when, in evolutionary time, MB1 and LMP7 shared a common ancestor. The genomic organisation of mouse delta has been determined and is similar to human LMP2 with six exons and five introns. This suggests that human delta will have an analogous organisation.

Thus, the class II region of the MHC can be viewed as a cluster of genes encoding class II α and β chains (DP, DQ, DR, DO and DM), some with divergent functions; a cluster of antigen processing genes (LMPs/TAPs) and very few other functional sequences (RING3), except for several transcribed fragments of unknown significance. Our knowledge of the class II region is nearing completion but further insights may come from two areas: 1) Comparison between different human haplotypes. 2) Comparison with other species, for example the chicken. The chicken MHC is composed of three loci, B-F- and B-L-
encoding molecules which are homologous to class I and class II, respectively, and the B-G gene cluster. There is no mammalian equivalent of the B-G gene cluster, although two genes in the human class I region, the myelin/oligodendrocyte glycoprotein (MOG) gene and the butyrophilin (BT) gene share similarities with the B-G antigens of the chicken MHC (Vernet et al., 1993a).

In contrast to the human MHC, the chicken class I and class II B-F and B-L genes are tightly linked and are associated with other assorted genes. For example, a GTP-binding protein gene (12.3) which is located upstream of the most telomeric B-G gene (Guillemot et al., 1989). The chicken MHC does not contain analogues of class III genes, such as C4, C2 of Bf (reviewed in Trowsdale, 1995). It is also thought that the chicken MHC contains TAPs but not LMP genes (J. Kaufman, personal communication). Further analysis the chicken class II region will provide insight into the genetic events which have contributed to the differences between the human and avian MHCs.
CHAPTER 8: REFERENCES


-192-


Chicz RM, Urban RG, Lane WS, Gorga JC, Stern JC, Dario AA, Vignali AA and Strominger JL. (1992). Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. Nature 358: 764-768.


Shukla H, Gillespie GA, Srivastava R, Collins F and Chorney MJ. (1991). A class I jumping clone places the HLA-G gene approximately 100 kb from HLA-H within the HLA-A subregion of the MHC. *Genomics* **10:** 905-914.


-218-


-221-


White PC, Grossberger D, Onufer BJ, Chaplin DD, New MI, Dupont B and Strominger JL. (1985). Two genes encoding steroid-21-hydroxylase are located...


