Novel genes in the class II region of the human major histocompatibility complex

by

Isabel Mary Hanson

A thesis submitted for the degree of Doctor of Philosophy in the University of London

Human Immunogenetics Laboratory
Imperial Cancer Research Fund
44 Lincolns Inn Fields
London

and

Department of Genetics and Biometry
University College
Gower Street
London

February 1991
to Chris
ACKNOWLEDGEMENTS

I am very grateful to the friends and colleagues who have helped to make my three years at ICRF such an enjoyable time. First and foremost I would like to thank my supervisor, John Trowsdale, and my lab. mum, Pat Miller. Next, I gratefully acknowledge my fellow Human Immunogeneticists, past and present: Danny Altman, Ian 'Peachy Keen' Campbell, Vincent Cunliffe, Will Foulkes, Richard Glynne, Vikki Groves, Hitoshi Ikeda, Adrian 'Gene' Kelly, Lesley-Anne Kerr, Chris 'More tea, vicar?' Lock, Ruth 'Someone's stolen my bike' Lovering, Ian Mockridge, Steve Powis, Jiannis 'Disaster' Ragoussis, Philippe Sanseau, David 'Captain' Sansom and David Wilkinson. I am also indebted to Kathy Cheah, Paul Freemont, Pat Gorman, Ketan Patel, Sabine Myers, Sue 'Perfect Day' Rider, Melissa 'Ruby' Rubock, Lisa Stubbs, Susan Tonks and my external supervisor, Jonathan Wolfe. Beyond ICRF I would like to acknowledge my parents, Charles & Anne Hanson, and my parents-in-law-to-be, Ken & Penny Riley. I am particularly grateful to Patricia Cocks for providing me with a roof over my head in salubrious St. John's Wood. Last, but by no means least, I would like to thank Chris Riley for his support which never faltered even though I didn't name my new genes after him.
The aim of the work described in this thesis was to identify and characterise novel genes in the class II region of the human major histocompatibility complex (MHC).

A physical map of the region, spanning over 1Mbp, was constructed using pulsed field gel electrophoresis (PFGE) in conjunction with probes for the known class II genes. This map facilitated the direct localisation of the gene for the α2 chain of type XI fibrillar collagen, *COL11A2*, to a region just centromeric of the class II DP subregion. In addition the PFGE map revealed four clusters of sites for restriction endonucleases which cut preferentially in CpG-rich regions often found at the 5' ends of genes. Three of these clusters were cloned by cosmid walking and chromosome jumping. Genomic fragments from these regions were hybridised to cDNA libraries which resulted in the identification of five novel genes designated *RING1-5*. *RING1*, *RING2* and *RING5* were 95kb, 90kb and 85kb proximal respectively to *DPB2*. *RING3* was 35kb distal to *DNA*. *RING4* was 25kb proximal to *DOB*.

Nucleotide sequencing revealed that *RING1-5* were not related to each other or to the class II genes. The predicted protein product of *RING1* contained a novel cysteine-histidine motif which was conserved in a variety of other proteins and which was reminiscent of domains found in zinc-dependent nucleic acid binding proteins. *RING3* potentially encoded a protein with striking homology to the product of *fsb*, a *Drosophila* developmental gene. The putative product of *RING4* was a member of the 'ABC' superfamily of ATP-dependent transporter proteins. *RING5* was the human homologue of *KE4*, a gene in the mouse MHC region which is a candidate for the developmental lethal *iu5*.

These findings may be of importance in understanding MHC/disease associations and the role of the MHC in the immune response.
CONTENTS

1. The major histocompatibility complex
   1.1. Classical studies of the MHC .................................. 12
   1.2. Function of class I and class II gene products ............. 16
   1.3. Molecular genetics of the MHC ............................... 26
   1.4. Evolution of the MHC ......................................... 36
   1.5. Clinical relevance of the class II region .................. 42
   1.6. Evidence for other genes in the class II region .......... 45
   1.7. Approaches to finding novel genes ......................... 48

2. Materials and Methods
   2.1. Bacterial cell culture ........................................ 51
   2.2. Screening of recombinant DNA libraries .................... 52
   2.3. Subcloning of DNA fragments ................................ 55
   2.4. Transformation of bacterial cells with DNA by ............
        electroporation ............................................. 56
   2.5. Preparation of DNA from transformed bacterial cells .... 57
   2.6. Small scale preparation of bacteriophage λ DNA .......... 60
   2.7. Preparation of eukaryotic DNA from cells in culture ..... 61
   2.8. Preparation of very high molecular weight DNA for ........
        pulsed field gel electrophoresis ............................ 62
   2.9. Preparation of RNA ......................................... 63
   2.10. Restriction endonuclease digestion of DNA ............... 64
   2.11. Electrophoresis of DNA and RNA ........................... 66
   2.12. Preparation of DNA and RNA blots ........................ 67
   2.13. Preparation of DNA probes ................................ 69
   2.14. Hybridisation of blots .................................... 70
   2.15. DNA sequencing ............................................ 71

3. Identification of potential CpG islands in the class II region by PFGE mapping
   3.1. Strategy to identify potential sites of genes .............. 75
   3.2. Choice of materials for construction of the PFGE map ..... 77
   3.3. Construction of the PFGE map ................................ 79
4. Mapping of the **COL11A2** gene to the class II region

4.1. Introduction

4.2. Mapping of **COL11A2** using somatic cell hybrids

4.3. Mapping of **COL11A2** by PFGE

4.4. Cosmid walking between the DP subregion and **COL11A2**

4.5. Summary and discussion

5. Identification of novel genes associated with clusters of rare-cutter sites

5.1. Introduction

5.2. Cloning of cluster 1 by cosmid walking

5.3. Cloning of cluster 2 by chromosome jumping

5.4. Identification of cluster 3 in previously isolated cosmid clones

5.5. Analysis of cloned regions for coding sequences

5.6. Expression patterns of **RING1-5**

5.7. Refinement of the physical map of the class II region

5.8. Summary and discussion

6. Characterisation of novel genes by nucleotide sequencing

6.1. Introduction

6.2. Nucleotide sequence of **RING1**

6.3. Partial nucleotide sequence of **RING2**

6.4. Nucleotide sequence of **RING3**

6.5. Nucleotide sequence of **RING4**

6.6. Partial nucleotide sequence of **RING5**

6.7. Summary

7. Comparative mapping of novel genes in the MHC region of mouse and man

7.1. Introduction
Contents (continued)

7.2. Determination of the positions of sequences homologous to the human genes RING1-5 and COL11A2 in the mouse MHC 158
7.3. Determination of the positions of sequences homologous to three mouse genes, KE3-5, in the human MHC 161
7.4. Summary and discussion 168

8. RFLP analysis of novel genes
8.1. Introduction 172
8.2. Approach to identifying RFLPs 172
8.3. RFLP analysis of COL11A2 173
8.4. RFLP analysis of RING1 174
8.5. RFLP analysis of RING2 175
8.6. RFLP analysis of RING3 176
8.7. RFLP analysis of RING4 177
8.8. RFLP analysis of B51 178
8.9. Summary and discussion 179

9. Concluding discussion
9.1. Advances in MHC mapping 181
9.2. Potential role of novel genes in class II-associated phenotypes 183
9.3. Additional applications of probes generated in this study 188
9.4. More new genes in the class II region? 189
9.5. Function of the MHC gene cluster 190

10. References 192

Appendices
A. Cosmid clones 221
B. cDNA clones 224
C. Gene symbols and database accession numbers 225
FIGURES

1.1. Position of the MHC on chromosome 6 15
1.2. Schematic illustration of antigen presentation 17
1.3. Schematic structures of class I and class II molecules 23
1.4. Atomic structures of class I and class II molecules 24
1.5. Molecular genetic maps of the MHC 28
1.6. Evolutionary tree for the class II genes 38
1.7. Organisation of murine and human MHC regions 40

3.1. Diagram of PFGE apparatus 78
3.2. PFGE mapping of the class II region 80
3.3. PFGE mapping of the class II region 81
3.4. PFGE mapping of the class II region 82
3.5. Physical map of the class II region showing clusters of rare-cutter sites 86

4.1. Generalised structure of a fibrillar collagen molecule 89
4.2. Mapping of \textit{COL11A2} by \textit{in situ} hybridisation 91
4.3. Mapping of \textit{COL11A2} using somatic cell hybrids 93
4.4. PFGE mapping of \textit{COL11A2} 94
4.5. Cosmid clones between \textit{COL11A2} and the DP subregion 96

5.1. Cosmid clones in the region centromeric of \textit{COL11A2} 100
5.2. Detailed map of the new cosmid clones 101
5.3. Determination of the methylation status of the cloned rare-cutter sites in the genome 103
5.4. Determination of the methylation status of the cloned rare-cutter sites in the genome 104
5.5. Construction of a rare-cutter jumping library 106
5.6. Restriction maps of jumping clone \textit{λ}j2 and cosmid HPB.ALL 71 109
5.7. Determination of the methylation status of the cloned rare-cutter sites in the genome 110
5.8. Overview of region cloned by jumping 111
Figures (continued)

5.9. Map of overlapping cosmids in the DQ subregion 113
5.10. Determination of the methylation status of the cloned rare-cutter sites in the genome 114
5.11. Zoo blot analysis of genomic fragment 33X1 116
5.12. Northern blot analysis of genomic fragment 33X1 118
5.13. Novel genes in clusters 1, 2 and 3 120
5.14. Northern blot analyses of RING1-5 122
5.15. Mapping of the DNA gene 125
5.16. Physical map of the class II region showing the positions of RING1-5 and COL11A2 126

6.1. Nucleotide and amino acid sequence of RING1 132
6.2. Alignment of proteins sharing the RING1 cysteine-histidine motif 133
6.3. Hypothetical structure of the RING1 cysteine-histidine motif 137
6.4. Nucleotide and amino acid sequence of RING2 138
6.5. Restriction map of RING3 cDNA clones 141
6.6. Nucleotide sequence of RING3 142
6.7. Alignment of the RING3 amino acid sequence with fsh 144
6.8. Nucleotide sequence of RING4 147
6.9. Hydropathicity plot of the RING4 protein product 148
6.10. Alignment of the RING4 amino acid sequence with members of the ABC superfamily 150
6.11. Schematic diagram of ABC structure 151
6.12. Alignment of the RING5 nucleotide and amino acid sequence with KE4 154

7.1. Molecular map of the proximal region of the mouse MHC 157
7.2. Determination of the positions of sequences homologous to the human genes RING1-5 and COL11A2 in the mouse MHC 159
Figures (continued)

7.3. Determination of the positions of sequences homologous to the mouse genes KE4 and KE5 in the human MHC 163
7.4. Restriction map of cosmid HPB.ALL 51 165
7.5. PFGE mapping of B51 166
7.6. Physical map of the proximal end of the human MHC region 167
7.7. Comparative map of the proximal regions of the mouse and human MHCs 169

A.1. Restriction map of the cosmid vector cos202 222
B.1. Restriction map of the cDNA vector CDM8 223
**TABLES**

1.1. Serologically defined HLA specificities 19  
1.2. Associations between the MHC and diseases 43  

3.1. Distribution of rare-cutter sites in human DNA 76  
3.2. PFGE fragments detected by class II probes 83  

4.1. PFGE fragments detected by COL11A2 and DPB1 95  

5.1. PFGE fragments detected by COL11A2 and 33X1 102  
5.2. Zoo blot results obtained with cosmid fragments 115  
5.3. cDNA clones isolated with cosmid fragments 121  
5.4. Expression patterns of RING1-5 123  

6.1. Homologues of RING1 134  

7.1. Genomic probes for the mouse genes KE2-5 162  
7.2. PFGE fragments detected by B51 and 33X1 168  

8.1. RFLP analysis of COL11A2 173  
8.2. RFLP analysis of RING1 (CEM15) 174  
8.3. RFLP analysis of RING1 (33X1) 175  
8.4. RFLP analysis of RING2 175  
8.5. RFLP analysis of RING3 176  
8.6. RFLP analysis of RING4 178  
8.7. RFLP analysis of B51 179  

A.1. Orientation of cosmid clones 221  
B.1. Orientation of cDNA clones 225
1. The 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 about 4Mbp of DNA and contains over 70 genes. Three features of the MHC in particular have stimulated the enormous amount of work on this gene cluster. First, it encodes molecules which play a central role in the regulation of the immune response. Second, it is associated with numerous diseases. Third, it represents 1/750th of the human genetic material, and as such is a paradigm for the detailed molecular genetic organisation of the genome.

1.1. Classical studies of the MHC

Many of the MHC loci were discovered by chance and were studied for several years in absence of any knowledge of their true function. This section summarises experiments in both human and murine systems which led to the discovery of the MHC.

1.1.1. Class I molecules

Some of the first experiments which identified a phenotype controlled by the MHC were performed seventy five years ago when it was shown that the ability to successfully transplant tumours between mice was genetically determined, being dependent on the strains selected as donor and host (Little and Tyzzer, 1916; Klein, 1986). Much later it was shown that that the outcome of tissue transplants segregated with a serologically defined blood antigen, H(histocompatibility)-2 (Gorer, 1937; Gorer et al., 1948). More refined genetic and serological analyses revealed that the situation was more complicated, with two segregant allelic series of antigens, H-2K and H-2D, playing a major role in determining the success of transplants (Amos et al., 1955). In humans, the presence of histocompatibility antigens was demonstrated when it was discovered that the sera of multiply transfused patients and
multiparous women contained antibodies which were capable of agglutinating the leukocytes of some donors but not others (Dausset, 1958; Payne and Rolfs, 1958; van Rood et al., 1958). The complex patterns of reactivity of these sera with large panels of donors were statistically analysed to define the first genetically controlled human leukocyte antigens (HLA; van Rood and van Leeuwen, 1963; Payne et al., 1964). Initially, two distinct loci, HLA-A and HLA-B, were found to control the expression of HLA molecules. It was soon demonstrated by family studies and population studies that these loci were closely linked (Ceppelini et al., 1967). Later, a third closely linked locus, HLA-C, was also described (Svejgaard et al., 1973). Studies of the effects of matching at the HLA-A, -B and -C loci on the survival of human tissue transplants suggested that the HLA antigens were analogous to the histocompatibility loci (H-2D and H-2K) which had been previously defined in mice. These molecules were classified together as the classical transplantation antigens, or class I antigens. The finding that humans and mice (and other species) had a gene cluster which played an important role in the survival of tissue transplants gave rise to the concept of the major histocompatibility complex.

1.1.2. Class II molecules
Evidence for a novel MHC-encoded function in man was obtained when it was found that strong proliferative responses were obtained when lymphocytes from different individuals were cultured together in vitro (Bach and Hirschorn, 1964). The locus responsible for this mixed lymphocyte reaction was closely linked to those encoding the classical transplantation antigens, but was distinct from them because the MLR proceeded even when the lymphocyte donors were matched for HLA-A, -B and -C (Yunis and Amos, 1971). The HLA-D region was thus defined. Later it became apparent from more detailed serological analysis that the HLA-D region contained two loci, HLA-DR and HLA-DQ (Tosi et al., 1978). Finally it was found that MLR occurred even between lymphocytes matched for DR and DQ and that there was a third D-region locus, HLA-DP (Shaw et al., 1980).

In mice, evidence for a new MHC linked function came from studies of the ability of different inbred strains to mount an immune response to synthetic antigens The 'immune response' (Ir) genes controlling this
phenotype were shown to map to a region of the mouse MHC distinct from those encoding the H-2K and D molecules (McDevitt et al., 1972). Two Ir molecules, I-E and I-A, were defined. Later it was shown that the mouse Ir and the human D-region molecules were related, and they were termed class II antigens.

1.1.3. Class III molecules
The first non-class I-class II gene was mapped to the MHC when it was discovered that mouse serum contained a serologically defined factor which was encoded in the region between H-2K and H-2D loci (Shreffler and Owen, 1963; Shreffler and David, 1972). Later it was demonstrated that this was the complement component C4 (Meo et al., 1975). Similarly in man it was found that two serologically defined serum proteins were encoded by the MHC, and later it was discovered that these proteins were the products of the two C4 loci (O'Neill et al., 1978). Complement components C2 and factor B (Bf) were also found to be encoded in the interval between HLA-D and HLA-B (Fu et al., 1974; Lamm et al., 1976). The genes encoding C4, C2 and Bf were inseparable from one another by recombination (Weitkamp and Lamm, 1982). The interval between the class II region and the class I region which encoded the complement components was designated the class III region.

1.1.4. Mapping of the classical MHC loci
The polymorphism of the MHC molecules facilitated the mapping of classical human MHC loci by recombination analysis, revealing the gene order DP-[DQ, DR]-[C4, C2, Bf]-B-C-A (Figure 1.1.; Weitkamp and Lamm, 1982).

The chromosomal location of the MHC was deduced from a variety of approaches. For example, HLA typing of the family of an individual with a cytologically detectable translocation breakpoint in region 21 on the short arm of chromosome 6 revealed that the breakpoint had occurred in the middle of the MHC region, thus positioning the MHC at 6p21 (Berger et al., 1979). More recently, HLA typing of γ-radiation-induced MHC-loss mutant cell lines coupled with high-resolution cytogenetic karyotyping facilitated the mapping of the MHC to the distal portion of band 6p21.3 (Spring et al., 1985). From recombination
Figure 1.1.
Position of the classical MHC loci relative to one another and to other polymorphic markers on human chromosome 6 (from Weitkamp and Lamm, 1982).
mapping of the MHC loci relative to other polymorphic markers on chromosome 6 the orientation of the complex on the short arm and its position relative to these markers was determined. These studies revealed that the DP locus was closest to the centromere (Figure 1.1.; Weitkamp and Lamm, 1982).

1.2. Function of class I and class II gene products

The class I molecules were initially defined as transplantation antigens (section 1.1.1.). The class II molecules were initially defined in man as being responsible for the mixed lymphocyte reaction and in mouse as the immune response genes (section 1.1.2.). However, it is now apparent that the class I and class II molecules are closely related, both structurally and functionally. By binding peptide fragments and presenting them on the cell surface, class I and class II molecules provide the context for recognition of peptide antigen by the T-cell receptor, and as such play a central role in the control of the immune response. 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 (Figure 1.2.). The interaction between the MHC/antigen complex and the T-cell receptor results in activation of the T-cell. Stimulated helper T-lymphocytes secrete lymphokines which promote antibody production by B-cells and assist in the activation of cytotoxic T-cells. Stimulated cytotoxic T-cells lyse the cell presenting the foreign antigen (Klein, 1986).

Class I molecules are expressed by virtually all nucleated cells. Class II molecules are expressed constitutively by B-lymphocytes, macrophages and activated T-lymphocytes and de novo expression can be induced by γ-interferon in a wide variety of cell types (Cresswell, 1987).

1.2.1. Polymorphism of class I and class II gene products

It was the polymorphism of the classical class I, class II and class III molecules which initially led to their discovery. In fact, the MHC is the most polymorphic gene system known.
Figure 1.2.
Highly schematic illustration of antigen presentation by class I and class II molecules to cytotoxic and helper T-cells respectively. APC, antigen presenting cell; TCR, T-cell receptor; ●, peptide. CD8 and CD4 are 'accessory' molecules on the surface of cytotoxic and helper T-cells respectively which are also involved in the interaction between the T-cell receptor and the MHC molecule.
The development of serological reagents to characterise the human and murine class I and class II antigens revealed that for each locus a remarkably large number of alleles were present in the population at a significant frequency (>1%). A list of the serologically defined human class I and class II specificities is shown in Table 1.1. (Bodmer et al., 1989b). More recently, the PCR-based cloning and sequencing of numerous class II genes has revealed even greater polymorphism at the DNA level (Bodmer et al., 1990). All the serologically detectable variation in class II DR molecules is attributable to polymorphism in the β-chain, because the α-chain is monomorphic. In contrast, both DQ chains are polymorphic. The DP α-chain is polymorphic, although markedly less so than the DP β-chain (Trowsdale et al., 1985).

Although some species, such as the Syrian hamster, seem to display limited MHC polymorphism, the highly polymorphic human and mouse systems are probably more typical (Klein, 1986). Because the polymorphism is especially marked in those loci which encode the functional antigen presenting molecules, and tends to be particularly concentrated within the exons encoding the domains which form the antigen binding groove, it is most likely that it has been established by natural selection rather than through random drift (Klein, 1986; Parham et al., 1988; Marsh and Bodmer, 1989; Klein and Takahara, 1990).

The major selection pressure on MHC class I and class II polymorphism has probably been exerted through pathogens (Klein, 1986). As described below, class I and class II molecules are important in determining which antigens an individual can respond to in two ways: through selective binding of specific peptide fragments, and through influencing the development of the T-cell repertoire. An individual whose MHC molecules fail to stimulate an immune response against a critical antigenic determinant on a pathogen, either because the antigen does not bind to the presenting molecules or because the individual's T-cells cannot recognise the MHC/antigen complex, has a selective disadvantage if infected by that pathogen. However, MHC polymorphism is of great advantage to the population as a whole because it ensures that other individuals will have combinations of class I and class II alleles which are effective in responding to the same
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th>DR</th>
<th>DQ</th>
<th>DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>B5</td>
<td>Cw1</td>
<td>Dw1</td>
<td>DR1</td>
<td>DQw1</td>
<td>DPw1</td>
</tr>
<tr>
<td>A2</td>
<td>B7</td>
<td>Cw2</td>
<td>Dw2</td>
<td>DR2</td>
<td>DQw2</td>
<td>DPw2</td>
</tr>
<tr>
<td>A3</td>
<td>B8</td>
<td>Cw3</td>
<td>Dw3</td>
<td>DR3</td>
<td>DQw3</td>
<td>DPw3</td>
</tr>
<tr>
<td>A9</td>
<td>B12</td>
<td>Cw4</td>
<td>Dw4</td>
<td>DR4</td>
<td>DQw4</td>
<td>DPw4</td>
</tr>
<tr>
<td>A10</td>
<td>B13</td>
<td>Cw5</td>
<td>Dw5</td>
<td>DR5</td>
<td>DQw5 (w1)</td>
<td>DPw5</td>
</tr>
<tr>
<td>A11</td>
<td>B14</td>
<td>Cw6</td>
<td>Dw6</td>
<td>DRw6</td>
<td>DQw6 (w1)</td>
<td>DPw6</td>
</tr>
<tr>
<td>A19</td>
<td>B15</td>
<td>Cw7</td>
<td>Dw7</td>
<td>DR7</td>
<td>DQw7 (w3)</td>
<td></td>
</tr>
<tr>
<td>A23 (9)</td>
<td>B16</td>
<td>Cw8</td>
<td>Dw8</td>
<td>DRw8</td>
<td>DQw8 (w3)</td>
<td></td>
</tr>
<tr>
<td>A24 (9)</td>
<td>B17</td>
<td>Cw9 (w3)</td>
<td>Dw9</td>
<td>DR9</td>
<td>DQw9 (w3)</td>
<td></td>
</tr>
<tr>
<td>A25 (10)</td>
<td>B18</td>
<td>Cw10 (w3)</td>
<td>Dw10</td>
<td>DRw10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A26 (10)</td>
<td>B21</td>
<td>Cw11</td>
<td>Dw11 (w7)</td>
<td>DRw11 (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A28</td>
<td>Bw22</td>
<td></td>
<td>Dw12</td>
<td>DRw12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A29 (w19)</td>
<td>B27</td>
<td></td>
<td>Dw13</td>
<td>DRw13 (w6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A30 (w19)</td>
<td>B35</td>
<td></td>
<td>Dw14</td>
<td>DRw14 (w6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A31 (w19)</td>
<td>B37</td>
<td></td>
<td>Dw15</td>
<td>DRw15 (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A32 (w19)</td>
<td>B38 (16)</td>
<td></td>
<td>Dw16</td>
<td>DRw16 (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A33 (w19)</td>
<td>B39 (16)</td>
<td></td>
<td>Dw17 (w7)</td>
<td>DRw17 (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A34 (10)</td>
<td>B40</td>
<td></td>
<td>Dw18 (w6)</td>
<td>DRw18 (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A36</td>
<td>Bw41</td>
<td></td>
<td>Dw19 (w6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A43</td>
<td>Bw42</td>
<td></td>
<td>Dw20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A66 (10)</td>
<td>B44 (12)</td>
<td></td>
<td>Dw21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A68 (28)</td>
<td>B45 (12)</td>
<td></td>
<td>Dw22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A69 (28)</td>
<td>Bw46</td>
<td></td>
<td>Dw23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw47</td>
<td></td>
<td>Dw24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw48</td>
<td></td>
<td>Dw25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw49 (21)</td>
<td></td>
<td>Dw26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw50 (21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw51 (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw52 (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw54 (w22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw55 (w22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw56 (w22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw57 (17)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw58 (17)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw60 (40)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw61 (40)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw62 (15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw63 (15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw64 (14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw65 (14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw71 (w70)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw72 (w70)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw75 (15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw76 (15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw77 (15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aw74 (w19)</td>
<td>Bw6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.1.**

Complete listing of serologically defined HLA specificities from the Tenth International Histocompatibility Workshop (Bodmer et al., 1989b).
pathogen (Benacerraf, 1981; Klein, 1986). Natural selection is probably responsible not only for maintaining the multiple alleles seen at each functional class I and class II locus but also for maintaining the the multiple class I and class II loci. The potential of each haplotype to encode more than one class I or class II molecule increases the range of antigens to which an individual can respond (Doherty et al., 1976).

1.2.2. Functional significance of class I and class II polymorphism
A series of key experiments contributing to contemporary understanding of the function of class I and class II molecules was performed by Zinkernagel and Doherty (1975; Doherty et al., 1976). Different strains of mice were infected intracerebrally with lymphocytic choriomeningitis virus (LCMV). One week later the mice were killed and their T-lymphocytes tested for the ability to lyse LCMV-infected fibroblasts from a mouse with the H-2\(^k\) MHC haplotype. The target fibroblasts were killed only by T-cells derived from H-2\(^k\) strains. It was later shown that all the mouse strains produced cytotoxic T-lymphocytes in response to the viral infection but in each case these lymphocytes were only able to lyse target cells of the same (self) MHC haplotype. Specifically, the phenotype was determined by the class I \(\text{H}-2\text{K}\) gene. On the basis of this work it was proposed that the cytotoxic T-lymphocytes had dual specificity, simultaneously recognising the foreign viral antigen and the self class I molecule. Later it was demonstrated that guinea-pig T-helper lymphocytes were only activated by antigens presented by cells of the self class II haplotype (Sprent, 1978). These studies gave rise to the concept of MHC-restricted antigen presentation, 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, or what form the antigen was in. Over the subsequent years it became apparent that T-cells had a single receptor which recognised a single complex of MHC molecule with antigen in the form of a peptide of about 5-10 residues in length (Figure 1.2.; Owen and Crumpton, 1987; Davis and Bjorkman, 1988). Proteins from which MHC-presented antigens are derived are now known to processed by partial proteolysis before presentation (Townsend and Bodmer, 1989).
Classical *in vivo* studies of the control of the immune response had already indicated that class II molecules could present short synthetic peptides (Benacerraf, 1981) and it was later shown that peptides could bind directly to purified class II molecules *in vitro* (Babbitt et al., 1985; Buus et al., 1986). Similarly it was shown that the antigen presented by class I molecules could be mimicked by short synthetic peptides (Townsend et al., 1986) and that class I molecules purified from cells were bound to peptides of 8-9 residues in length (Elliott et al., 1990).

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). Later it was shown that the ability to mount an immune response correlated directly 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 amino acid sequences from a protein antigen, whether a T-cell response can be generated against that protein antigen (Benacerraf, 1981).

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 (Schwartz, 1989; Marrack and Kappler, 1988). This process takes place in the thymus and involves both the positive and negative selection of maturing T-lymphocytes. The positive selection process is poorly understood but is thought to involve the recognition by the T-cell receptor of thymically-expressed self class I or class II molecules, and results in only those T-cells bearing receptors which recognise self MHC molecules reaching the pool of mature T-cells. Positive selection is the means by which all mature T-cells recognise self MHC molecules as restriction determinants for foreign antigens. The process of negative selection eliminates those T-cells bearing receptors which have high affinity for self MHC + self peptide, and which would thus be autoreactive in the periphery. There is good evidence for the role of class II molecules in the negative selection procedure. Using monoclonal antibodies which recognise T-cell receptors containing the Vβ17a epitope it was shown that T-cells carrying such receptors were selectively deleted in the thymuses of
mice expressing the I-E antigen (Schwartz, 1989). Negative selection is a major mechanism through which tolerance to self proteins is established, and is thought to involve the recognition by maturing T-cells of thymically expressed MHC molecules presenting peptides from self proteins. However, the clonal deletion process does not remove all autoreactive T-cells, especially in the case of tissue-specific antigens not expressed in the thymus. Potentially autoreactive T-cell clones in the periphery may be controlled by additional mechanisms, such as suppressor T-cells, about which little is known (Schwartz, 1989).

1.2.3. Structure of class I and class II MHC molecules
Class I and class II molecules were initially characterised using standard biochemical techniques (Strominger, 1987). The functional molecules are cell surface glycoproteins. Both are heterodimers of α- and β-chains, as illustrated schematically in Figure 1.3. The mature class I or class II molecule has four extracellular domains, each of approximately ninety amino acids. For class I molecules, three of these domains are contained within the ~43kD α-chain, which is encoded within the MHC. The fourth domain is provided by β2-microglobulin (~12kD), which is encoded on human chromosome 15. In the case of class II molecules, two domains are contained within the ~34kD α-chain and two within the ~29kD β-chain, and both chains are encoded within the MHC.

More recently, the crystal structure of two class I molecules, HLA-A2 and HLA-Aw68, have been determined (Figure 1.4.; Bjorkman et al., 1987a; Garrett et al., 1989). The membrane-proximal domains α3 and β2-microglobulin interact to form the base of the molecule upon which the polymorphic α1 and α2 domains are supported. The α1 and α2 domains interact to form a cleft, the floor of which is composed of eight anti-parallel β-strands and the walls of which are formed by two α-helices. The crystal structure of a class II molecule has not yet been obtained, but a hypothetical structure has been modeled based on the amino acids conserved between class II and class I (Brown et al., 1988). In this model, the α1 and β1 domains interact to form a cleft similar to that observed in the class I structure (Figure 1.4.).
MHC class I

\[ \alpha^1 \alpha^2 \alpha^3 \]

Cytoplasm

Existed by virtually all nucleated cells

MHC class II

\[ \beta_1 \beta_2 \alpha_1 \alpha_2 \]

Cytoplasm

Restricted expression

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

Figure 1.3.
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.
Figure 1.4.

(a) Three-dimensional crystal structure of the class I molecule HLA-A2 (i) from the side, showing the antigen binding groove created by the α1 and α2 domains, and (ii) looking down on the groove from above (from Bjorkman et al., 1987a).

(b) Predicted structure of the antigen binding site of a class II molecule (from Brown et al., 1988).
The groove identified in the membrane-distal surface of the class I and class II molecules is thought to be the site of peptide binding. The dimensions of the cleft are appropriate for the accommodation of a peptide of between 8 and 25 amino acids in length, depending on the conformation of the peptide. In the two X-ray diffraction studies it was observed that unidentified material had co-crystallised in the cleft of the class I molecules, and theoretically this could be peptide (Bjorkman et al., 1987a; Garrett et al., 1989). In addition, many of the polymorphic residues and many of the residues predicted from functional studies to be important for peptide binding were found to line the groove (Bjorkman et al., 1987b; Parham et al., 1988; Marsh and Bodmer, 1989). The shape of the groove created by the polymorphic residues in HLA-A2 was quite distinct from that in HLA-Aw68, providing a structural basis for the observed allelic specificity in peptide binding (Garrett et al., 1989).

1.2.4. Peptide binding to class I and class II molecules

Class I molecules, which bind and present antigens derived from intracellular proteins, are thought to become complexed with peptide in the endoplasmic reticulum (ER) or a closely associated compartment (Townsend et al., 1989; Yewdell and Bennink, 1990). On the basis of current evidence, it seems likely that peptides are generated by proteolytic activity in the cytoplasm and then transported into the lumen of the ER where they bind to newly synthesised class I heavy chains (Townsend and Bodmer, 1989; Townsend et al., 1989). Interestingly, a gene encoding a function thought to be important for normal association of class I heavy chains with peptide is probably encoded within the MHC (section 1.6.2.). Formation of the heavy chain/peptide complex induces association with β2-microglobulin and subsequent export from the ER (Townsend et al., 1989). The class I/peptide complex moves rapidly from the ER to the Golgi apparatus, where modification of carbohydrate occurs, and then to the cell surface, where presentation of the bound antigen to T-lymphocytes occurs (Neefjes et al., 1990).

Class II molecules, which bind and present peptides derived from extracellular proteins, are thought to become complexed with peptide in an endosomal compartment (Yewdell and Bennink, 1990). Newly
synthesised class II molecules become associated in the ER with a protein known as invariant chain, which plays a key role in differentiating the exogenous antigen presentation pathway taken by class II molecules from the endogenous antigen presentation pathway taken by class I molecules. *In vitro* studies with purified invariant chain and class II molecules have shown that binding of invariant chain to class II molecules inhibits the binding of peptides (Teyton et al., 1990). This work led to the suggestion that association of invariant chain with class II could prevent binding of cytoplasmically-derived peptides to class II in the ER. From pulse-chase and immuno-electron microscopy experiments evidence has been obtained that the class II/invariant chain complex proceeds rapidly from the ER to the Golgi apparatus but instead of continuing directly to the plasma membrane the complex is diverted to a novel post-Golgi cytoplasmic compartment where it is retained for several hours before the appearance of mature class II molecules on the cell surface (Neefjes et al., 1990). Targetting of the invariant chain/class II complex to the post-Golgi vesicles is dependent on an amino acid sequence, identified by deletion mapping, in the N-terminus of the invariant chain molecule (Bakke and Dobberstein, 1990). It has been proposed that the compartment to which the complex is targeted may be an endosomal vesicle in which invariant chain becomes dissociated, facilitating binding of class II to peptides which have been generated by proteolysis from endosomally imported exogenous proteins (Neefjes et al., 1990; Teyton et al., 1990). This is consistent with the observation that antigen presentation by class II molecules, but not by class I molecules, is sensitive to inhibitors which disrupt the endosomal trafficking and processing pathway by which extracellular proteins are taken into the cell and degraded (Yewdell and Bennink, 1990).

1.3. Molecular genetics of the MHC

1.3.1. Application of molecular cloning and mapping techniques to the MHC region

With the advent of molecular cloning technology, it soon became apparent that the complexity of the MHC was much greater than had been previously indicated by the classical studies. The power of this
technology is illustrated by comparing the map in Figure 1.1., which summarises the classical studies, to that in Figure 1.5.b., which summarises ten years of molecular studies on the MHC region. The MHC provides an excellent example of the way in which a variety of cloning and mapping methods can be used to characterise in detail the molecular genetic organisation of a region of the mammalian genome.

The first important advance made possible by cloning technology was the isolation of probes for the classically defined MHC genes. These facilitated the sequencing, and analysis of gene organisation, of individual loci. In addition, through hybridisation of 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 are now known to correspond to seventeen class I genes, whereas only three class I loci had been detected by serological methods. Similarly class II α- and β-chain probes revealed class II genes in addition to those encoding the serologically defined DP, DQ and DR antigens. These class I- and class II-related genes were isolated by exhaustive screening of genomic and cDNA libraries with probes for the classical MHC genes. All these cross-hybridising human MHC genes have now been isolated, but additional class I- and class II-related sequences may exist which have diverged too greatly to be detected by this approach.

Initially, somatic cell hybrids containing fragments of chromosome 6 and γ-irradiation-induced mutants with deletions in chromosome 6 were used to show that these related genes also mapped to the MHC region. In some cases the relative positions of genes which are in very close physical proximity, such as the four genes in the DP subregion, could be determined by cosmid cloning. However, the most important advance in the understanding of the organisation of the region was facilitated by the development of the powerful long range physical mapping technique of pulsed field gel electrophoresis (PFGE; Schwartz and Cantor, 1984; Carle and Olson, 1984). In conjunction with the numerous MHC probes obtained from gene cloning studies, PFGE allowed the accurate mapping of the genes and subregions which had not been previously linked on overlapping genomic clones.
Figure 1.5.

Molecular genetic organisation of the human major histocompatibility complex in (a) 1987 (b) 1991. These maps are based on data described in section 1.3 and results presented in this thesis. The centromere is towards the left. Open boxes represent genes; hatched boxes represent genes known to be pseudogenes. Arrows show the direction of transcription, where known (5' → 3'). Horizontal bars beneath each map indicate regions which have been cloned. The arrangement of genes shown in the class II DR subregion is as for a DR2 haplotype. The relative order of genes within brackets is not known. The small gaps in the map of the telomeric end of the class I region indicate that the physical distances between gene clusters is uncertain. The genes numbered 1-18 in the class III region are genes G1-G18 (see Kendall et al., 1990). Genes G2, G3, G5, G7a, G9, G9a and G10 were also identified independently by Spies et al. (1989a, 1989b) where they were designated BAT2, BAT3, BAT4, BAT6, BAT7, BAT8 and BAT9 respectively.
When the experiments described in this thesis were started most of the known MHC genes had been isolated on single genomic clones or small contigs extending tens of kilobases which were unlinked to one another (Figure 1.5.a). The PFGE maps revealed that there were large gaps between these cloned regions, which theoretically could contain many additional genes. Recently it has been the aim of several laboratories to clone and analyse the DNA in these intervals, using a 'reverse genetics' approach to detect genes in genomic clones without any knowledge of the functions that those genes might encode. The application of reverse genetics techniques has resulted in a remarkable increase in our understanding of the molecular organisation of the MHC over the last three years (compare Figures 1.5.a and 1.5.b.).

Details of the molecular genetic organisation of the different MHC regions, as deduced from the application of the approaches described above, are described in the following sections.

1.3.2. The class I region

The class I genes are found in a region spanning approximately 2Mbp at the telomeric end of the MHC. The map order of the classical class I loci, HLA-A, -B and -C, as determined by recombination analysis has been confirmed by PFGE mapping. Thus, HLA-C is 130kb distal to HLA-B and just over 1000kb proximal to HLA-A (Carroll et al., 1987; Pontarrotti et al., 1988; Ragoussis et al., 1989).

Hybridisation of class I gene probes to Southern blots of total human genomic DNA revealed the presence of numerous other class I-related sequences (Orr and DeMars, 1983). Fourteen non-A-B-C class I gene sequences, accounting for all the cross-hybridising bands seen on Southern blots, were isolated from genomic libraries. 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 are expressed and give protein products and have been designated HLA-E (Koller et al., 1988), HLA-F (Geraghty et al., 1990) and HLA-G (Kovats et al., 1990). However, the functions of these non-classical class I genes is as yet unknown. Using a panel of mutant cell lines with deletions in different regions of the MHC, it was possible to demonstrate that the non-A-B-C sequences mapped to the class I region. Most of the
sequences mapped close to, or between, the \textit{HLA-A}, \textit{-B} and \textit{-C} genes, but from recombination analysis it was shown that four (including \textit{HLA-F} and \textit{-G}) mapped 8cM telomeric of \textit{HLA-A}, while a fifth mapped an additional 2cM telomeric (Koller et al., 1989).

Another intact non-\textit{A-B-C} class I gene, \textit{cda12}, has been described by Ragoussis et al. (1989), which maps within 50kb of \textit{HLA-A}. The relationship of this gene to those described by Koller et al. (1989) is currently unclear. In a PFGE mapping study using a class I probe at reduced stringency it was shown that all the class I-related genes were encompassed in a region spanning a total of 2Mbp (Ragoussis et al., 1989; Figure 1.5.b.). In the human genome, the average frequency of recombination between genes is such that a genetic distance of 1cM is roughly equivalent to a physical distance of 1Mbp. This relationship holds well in the proximal portion of the class I region, where the genetic distance between \textit{HLA-B} and \textit{HLA-A} is roughly 1cM, and the physical distance is just over 1Mbp. However, the more distal class I genes map within 1Mbp of \textit{HLA-A} and yet have a genetic separation of up to 10cM (Koller et al., 1989; Ragoussis et al., 1989). This provides evidence for an unusually high frequency of recombination between \textit{HLA-A} and the more distal class I genes.

1.3.3. 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 1Mbp (Hardy et al., 1986; Ragoussis et al., 1989; Dunham et al., 1989). The genes which encode the \(\alpha\) and \(\beta\) chains of the serologically defined class II antigens DP, DQ and DR are localised in discrete genetic subregions. In addition this region contains class II pseudogenes and genes which are apparently intact but for which no protein product has yet been identified.

The first class II cDNA clones isolated were for the \textit{DRA} gene. Polysomes from a B-lymphoblastoid cell line were selected with a monoclonal antibody against the \textit{DR\(\alpha\)} chain, and the mRNA from the antibody-bound polysomes was used as a template to make cDNA (Korman et al., 1982a). The cloning of \textit{DRA} facilitated the cloning of other class II \(\alpha\)-chain genes by hybridisation of the \textit{DRA} probe 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). Again, 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.

The DP subregion
The DP subregion spans 70kb at the proximal end of the class II region and contains two α-chain genes (DPA1 and DPA2) and two β-chain genes (DPB1 and DPB2) which are organised as shown in Figures 1.5 and 4.5. (Trowsdale et al., 1984; Servenius et al., 1984; Okada et al., 1985a). The expressed DP molecule has been demonstrated by transfection studies to be the product of the DPA1 and DPB1 genes (Okada et al., 1985a). The DPA2 and DPB2 genes, in contrast, are non-transcribed pseudogenes, containing frame-shift mutations and defective splice junctions which would prevent the expression of a functional protein product (Gustafsson et al., 1987).

The DQ subregion
The DQ subregion contains five genes, as shown in Figure 1.5.b. The DQA1/DQB1 gene pair and the DQA2/DQB2 (formerly DXα/DXβ) gene pair were initially isolated on two unlinked sets of overlapping genomic clones (Okada et al., 1985b; Auffray et al., 1984; Jonsson et al., 1987). The two gene pairs are highly related at the nucleotide sequence level. 15kb distal to the DQA2 gene is a truncated pseudogene, DQB3 (formerly DVβ), which apparently does not contain the exons encoding the signal sequence or β2 domain (Ando et al., 1989). Sequencing of the remaining exons revealed that DQB3 is more closely related to DQB1 and DQB2 than to other class II β-chain genes. The entire DQ subregion has now been cloned on overlapping cosmids, revealing that the DQA1/DQB1 and DQA2/DQB2 gene pairs are 75kb apart (Blanck and Strominger, 1988). The DQA1 and DQB1 genes are fully functional and encode the serologically defined DQ antigen. Although the DQA2 and DQB2 genes are intact and do not appear to contain any deleterious mutations which would prevent their expression at the RNA or
protein level, transcription of these loci has never been detected (Auffray et al., 1987).

The DR subregion
The DR subregion is the only class II subregion known to vary in gene number between individuals. All haplotypes carry a single α-chain gene, DRA, but the number of β-chain genes is variable. DR4 haplotypes, for example, carry four DRB genes, while DR1 and DR8 haplotypes are thought to have only one (Bohme et al., 1985; Andersson et al., 1987). The DR subregions of the DR2, DR3 and DR4 haplotypes have been analysed in detail by cosmid cloning although in none of these studies were all the DR genes linked on a single contig (Rollini et al., 1985; Spies et al., 1985; Andersson et al., 1987; Kawai et al., 1989). The three DRB genes, DRB1, DRB2 and DRB3, from the DR2 haplotype have been linked on overlapping cosmid clones (Kawai et al., 1989). The three genes were contained within a region spanning 80kb. They were not linked to the DRA gene. The DRB2 gene lacked an exon containing the 5' untranslated region, and did not give a cell-surface product when co-transfected with the DRA gene into mouse L-cells, and is therefore probably a pseudogene. The DRB1 and DRB3 genes, in contrast, were expressed with DRA in transfection studies to give functional class II molecules (Kawai et al., 1989). The DR subregions in the DR3 and DR4 haplotypes also contained two functional DRB genes, along with one or two pseudogenes (Rollini et al., 1985; Spies et al., 1985; Andersson et al., 1987).

The DN subregion
The DN subregion is currently defined by a single gene, DNA (formerly DZa). The DNA gene was isolated by screening a human cosmid library with a DRA gene probe at reduced stringency (Spielman et al., 1984). This gene has been shown by physical mapping to be between the DP subregion and the DOB gene (Inoko et al., 1989), but has not been accurately positioned within this interval or oriented on the chromosome. The nucleotide sequence of a DNA genomic clone revealed that the DNA gene has a similar organisation to that of other class II α-chain genes, and that it does not contain any mutations which would prevent transcription (Trowsdale and Kelly, 1985). Indeed, the DNA gene is transcribed in B-lymphocytes, giving a major transcript of 3.5kb
and a minor transcript of 1.1kb (Kelly, 1988). The major transcript is thought to be the result of inefficient 3' RNA processing directed by the unusual polyadenylation signal (ACTAAA) found in the DNA gene (Trowsdale and Kelly, 1985). The minor transcript however is processed and polyadenylated normally and does not appear to contain any deleterious mutations that would prevent translation; however, no protein product has been described (Young and Trowsdale, 1990). It has been speculated that a DNA gene product might pair with a DOB gene product to form a functional class II molecule, but this is unlikely because transcription of the DNA and DOB genes is not co-ordinately regulated (Tonnelle et al., 1985).

The DO subregion
The DO subregion is currently defined by a single gene, DOB. The DOB gene was first identified by screening a human cDNA library with a mixture of DRB and DPB probes at reduced stringency (Tonnelle et al., 1985). In an independent study DOB was isolated from a human genomic library using a probe for the mouse class II gene Ob (formerly Aβ2) at low stringency (Servenius et al., 1987). The DOB gene is transcribed at low levels in B-lymphocytes but no protein product has been reported. Like the DNA gene, DOB is isolated in the class II region and a closely physically linked α-chain gene has not yet been described. As mentioned above, it is unlikely that any DOB product would pair with an α-chain from the DP, DQ, DR or DN subregions because DOB gene expression is not co-ordinately regulated with that of the other class II loci. Specifically, DOB is transcribed only at low levels in B-lymphocytes, and transcription of DOB in fibroblasts is not inducible by γ-interferon (Tonnelle et al., 1985). Thus, it is possible that a DOA gene remains to be discovered, and that a DO molecule exists with a function distinct from the classical class II molecules. The DOB gene was mapped distal of the DQ subregion (Hardy et al., 1986) and has recently been accurately positioned 45kb proximal to DQB2 by cosmid walking (Blanck and Strominger, 1988; Figures 1.5. and 5.9.).

1.3.4. 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 1.1Mbp (Figure 1.5.; Dunham et al., 1987; Carroll et al., 1987; Ragoussis et al., 1989; Dunham et al., 1990).

With the advent of molecular cloning techniques it became possible to clone the class III complement genes which had been identified by classical studies. The genes for C4, C2 and factor B were all cloned by designing synthetic oligonucleotides based on the known protein sequences and using these to screen cDNA libraries. cDNA probes were then used to screen cosmid libraries, and overlapping clones were obtained which contained these genes (Carroll et al., 1984). The cloned complement gene subregion contained C2, Bf and two C4 genes (C4A and C4B) within 120kb of DNA. The C2 and Bf genes were separated by less than 500bp, while C4A and C4B were about 10kb apart. The products of these genes are serum glycoproteins which are components of the classical (C4 and C2) and alternative (factor B) complement cascades. These cascades ultimately result in the formation of a membrane attack complex, a transmembrane channel which causes lysis and death of the cells of invading micro-organisms (Reid, 1988). The classical pathway is activated by antibody bound to antigen and is therefore a major effector of the humoral immune response. The alternative pathway can be activated directly by the cell surface of the pathogen.

Prompted by linkage analysis in pedigrees affected by congenital adrenal hyperplasia caused by steroid 21-hydroxylase deficiency, which had mapped the defective locus to the class III region, it was soon demonstrated that the region encompassing the complement genes also contained two genes for steroid 21-hydroxylase (21A and 21B in Figure 1.5.), one just proximal of each of the C4 genes (Carrol et al., 1985; White et al., 1985). 21A is a pseudogene. Deleterious mutations in the functional steroid 21-hydroxylase gene were shown to be responsible for congenital adrenal hyperplasia (White et al., 1985). Deletions and duplications of one or other [21-hydroxylase/C4] gene unit are not uncommon in the population (Carroll and Alper, 1987).

The genes for the related lymphokines tumour necrosis factor (TNF) α and β (Tnfa and Tnfb) were localised to the MHC by analysis of MHC deletion mutants (Spies et al., 1986) and then to the class III region by
PFGE mapping (Inoko and Trowsdale, 1987; Dunham et al., 1987; Carroll et al., 1987).

PFGE mapping revealed that the distance between \textit{DRA} and the complement gene cluster was about 390kb, while the distance between the complement genes and \textit{HLA-B} was about 580kb (Dunham et al., 1987; Carroll et al., 1987; Ragoussis et al., 1989; Dunham et al., 1990). These intervals were clearly large enough to accommodate numerous genes in addition to the \textit{Tnf} genes. Intensive efforts to clone the entire class III region were initiated, and these have resulted in over 900kb of DNA being cloned in overlapping cosmids (Spies et al., 1989a, b; Sargent et al., 1989a, b; Kendall et al., 1990; Spies et al., 1990). The remaining interval between the proximal end of the cosmid contig and the \textit{DRA} gene has now been covered by yeast artificial chromosome clones (Ragoussis et al., 1991b).

Analysis of the cloned regions for coding sequences, using reverse genetics techniques, led to the discovery of at least 26 novel genes in the class III region (Figure 1.5.b.; Levi-Strauss et al., 1988; Spies et al., 1989a, b; Sargent et al., 1989a, b; Morel et al., 1989; Milner and Campbell, 1990; Kendall et al., 1990; Spies et al., 1990). The total number is uncertain because there have been two very recent independent reports of novel genes between \textit{21B} and \textit{DRA} and it is not yet clear whether both groups have identified the same genes (Kendall et al., 1990; Spies et al., 1990). In Figure 1.5.b. the data of Kendall et al. (1990) are shown because this report described seven novel genes whereas Spies et al. (1990) only described six. The density of genes in the class III region is remarkably high. In fact, one of the novel genes (\textit{OSG} in Figure 1.5.b.) actually overlaps, on the opposite strand, with the \textit{21B} gene (Morel et al., 1989).

Three of the novel genes were shown by nucleotide sequencing to be members of the heat shock protein \textit{hsp70} multigene family (Sargent et al., 1989a; Milner and Campbell, 1990). \textit{Hsp70-1} and \textit{hsp70-2} are very closely related, encoding identical protein products. \textit{Hsp70-hom} shares 90\% identity with \textit{hsp70-1} but unlike \textit{hsp70-1} and -2, expression of this gene is not heat-inducible.
The nucleotide sequences of a further four of these novel genes (RD, BAT2, BAT3 and OSG) have been published. The RD gene encodes a predicted protein product of 42kD, which contains an novel motif consisting of a reiterated arginine-aspartate dipeptide (Levi-Strauss et al., 1988). BAT2 and BAT3 (G2 and G3 in Figure 1.5.b.) encode proteins of 228 and 110kD respectively, both of which are unusually rich in proline (Banerji et al., 1990). The 'opposite strand gene', OSG, which overlaps with 21B is expressed, like 21B, in steroidogenic adrenal tissue (Morel et al., 1989). This intriguing finding is suggestive of a functional or regulatory relationship between 21B and OSG, but the partial sequence of OSG did not reveal what this might be. In addition to the sequence data from these human genes, a partial sequence has been determined for the mouse homologue of the B144 gene but this did not match any sequences in the nucleotide sequence databases and the function of this gene is not known (Tsuge et al., 1987). No sequence data have yet been published for the other novel genes.

1.4. Evolution of the MHC

1.4.1. Evolutionary relationships between class I and class II genes
Once the genes encoding class I and class II molecules had been cloned it became apparent from the conservation in gene organisation and identity at the nucleotide sequence and predicted amino acid sequence level that they were related. In particular, there was significant homology between the class II α2 domain, the class II β2 domain and the class I α3 domain. These domains were also related at the amino acid sequence level to the conserved 'antibody fold' domain found in members of the immunoglobulin supergene family (Korman et al., 1982b; Hood et al., 1986). The common ancestry of class I and class II molecules is also reflected in their related functions and their related structures (Figures 1.2. and 1.4.). It is not clear when in evolutionary time MHC molecules arose. However, class I and class II genes have now been described in amphibians, birds, 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; Hashimoto et al., 1990).
The ancestral class II and class I genes have clearly been duplicated many times to give the multiple loci seen in the class II and class I regions of the mammalian MHC. The organisation of the human class II region suggests that the subregions were generated from a primordial \( \alpha/\beta \) gene pair through a series of duplication and divergence events (Klein, 1986; Bodmer et al., 1986).

Comparisons of the nucleotide and amino acid sequences of the class II \( \alpha \)-chain genes and their protein products revealed that \( DPA1, DQA1, DRA \) and \( DNA \) are equally diverged from one another, suggesting that that they arose by duplication at roughly the same time (Figure 1.6.; Auffray et al., 1984). A similar analysis of the \( \beta \)-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 \( \beta \)-chain genes in evolutionary time (Tonnelle et al., 1985). Sequence comparisons between genes within subregions reveals that additional duplications have probably occurred much more recently (Figure 1.6.). For example, the \( DQA1 \) and \( DQA2 \) genes are 99% related in the exon encoding the \( \alpha2 \) domain (Auffray et al., 1987). The duplication which took place in the DP subregion probably occurred earlier than that in the DQ subregion because the \( DPA2 \) and \( DPB2 \) genes are significantly diverged from \( DPA1 \) and \( DPB1 \).

At the nucleotide level, the identity between \( DPA1 \) and \( DPA2 \) is 76%, while between \( DPB1 \) and \( DPB2 \) it is 86%. The increased divergence of the \( DPA2 \) gene compared to \( DPB2 \) suggests that \( DPA2 \) became a pseudogene before \( DPB2 \) (Gustafsson et al., 1987).

1.4.2. Organisation of the MHC region in different species

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; Denaro et al., 1985). In the same way the \( DQA \) and \( DQB \) genes are related to the mouse \( Aa \) and \( Ab \) genes respectively (Kaufman et al., 1984). The isolated mouse \( \beta \)-chain gene, \( Ob \) (formerly \( A\beta2 \)), is homologous to the isolated human \( \beta \)-chain gene, \( DOB \) (Larhammar et al., 1985; Tonnelle et al., 1985; Servenius et al., 1987). Finally, the isolated mouse \( \beta \)-chain pseudogene \( Pb \) (formerly \( A\beta3 \)) was found to be most related to the
Figure 1.6.
Evolutionary tree for the class II genes. HLA and Ig refer to the primordial MHC and immunoglobulin genes. ABC refers to the primordial class I gene. Dα and β refer to the primordial class II α- and β-chain genes. Subsequent events involve pairwise combinations of α and β genes except where indicated. Approximate divergence times are indicated in millions of years (from Bodmer et al., 1986). DZ is now known as DN; DX is now known as DQ2.
sequences of the two β-chain genes 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 by cosmid cloning and physical mapping it was revealed that the relative positions of the homologous subregions is conserved between the two species (Steinmetz et al., 1986; Hardy et al., 1986). These sequence comparison and gene mapping data have led to the hypothesis that the overall organisation of the class II region was established before the radiation of rodents and primates about 135 million years ago (Bodmer et al., 1986; 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 most likely occurred more recently. In contrast to the distinct conservation of subregions seen in the mouse and human class II regions, the class II genes of the chicken MHC are more closely related to one another than they are to DR, DQ or DP. Thus, the multiple chicken class II genes probably arose independently by duplication from a primordial class II gene pair after the separation of the mammalian and avian lineages about 300 million years ago (Kroemer et al., 1990).

As in the human MHC, the mouse class I and class II regions are separated by a class III region (Figure 1.7.). 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 not yet been mapped in the mouse. Both class III regions contain genes for steroid 21-hydroxylase and the complement components C4, C2 and factor B; the organisation of these closely linked genes is similar in the two species (Chaplin, 1985). The human RD gene has a homologue in the analogous position in the mouse class III region (Levi-Strauss et al., 1988). The B144, Tnfa, Tnfb and BAT1 genes have recently been linked on overlapping cosmids in mouse, and their order is conserved relative to the human class III region (Wroblewski et al., 1990). A mouse hsp70 gene has been localised to the class III region by recombination analysis but its position relative to other class III loci is not yet known (Gaskins et al., 1990).
Figure 1.7.

Schematic illustration of the organisation of the MHCs of man and mouse. The relative positions and approximate sizes of the class I, class II and class III regions are shown. The positions of some of the functional class I, class II and class III loci are also indicated.
The chicken MHC has a different organisation to the well-characterised mammalian MHCs in that the class II and class I 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 polymorphic cell surface antigens, the B-G molecules, whose function is unknown. No homologues of the B-G genes have been detected in mammalian genomic DNA by cross-hybridisation with chicken B-G probes, but this does not rule out the possibility that B-G homologues are present in mammals (and even in the MHC) because in general it has been found that mammalian and avian MHC genes have diverged too greatly at the nucleotide sequence level for cross-hybridisation to be detected. There is no distinct class III 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 the genes in the mammalian class III region (Kroemer et al., 1990). It will be of great interest to determine whether the chicken MHC is associated with class III genes, as this may help to resolve the controversy of whether these genes have become associated with the class I and class II genes completely 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., 1986).

The major difference in organisation between the human and mouse MHCs is in the class I region. The human MHC contains seventeen class I genes, while the mouse MHC contains between 26 and 33 class I genes depending on the haplotype (Flavell et al., 1986). As mentioned previously, the human class I genes all lie telomeric of the class III region (Figure 1.7.). Most mouse class I genes are also found in a region telomeric of the class III region. However, a pair of class I genes, K and K2, are present at the centromeric end of the class II region, about 70kb proximal to Pb (Figure 1.7.; Widera and Flavell, 1985; Steinmetz et al., 1986). These genes are proposed to have arisen by duplication in the Qa region of the mouse class I gene cluster, and to have subsequently moved to their current location by an intrachromosomal double cross-over event (Bodmer, 1981; Weiss et al., 1984). This organisational change is believed to have occurred since the radiation of rodents and
mammals because it has only been found in the MHCs of the closely related rat and mouse (Klein, 1986).

1.5. Clinical relevance of the class II region

The extreme polymorphism of the MHC genes and gene products has provided a powerful system of markers to test for association between the MHC and diseases. In population association studies, the frequency of a particular MHC allele is tested (using serological reagents, or, more recently, using RFLPs or sequence specific oligonucleotides) in a group of individuals with a particular disease and a group of healthy matched controls. The frequencies of each allele in the two groups are calculated and subjected to statistical tests to determine whether there is a significantly increased or decreased frequency of a given allele between the two groups. If a significant difference is found, that allele is considered to be associated with the disease.

This approach has revealed significant associations between the MHC and over 40 diseases (Table 1.2.; Tiwari and Terasaki, 1985). Many of the associations are strongest with class II region; these include insulin-dependent diabetes mellitus (IDDM), rheumatoid arthritis, narcolepsy and Hodgkin's lymphoma. The associations are never complete, reflecting the fact that these diseases usually have complex genetics and that the final development of the disease phenotype is probably dependent on a combination of environmental and genetic factors. Part of the genetic component is generally considered to be contributed by a 'disease susceptibility' gene in the class II region which predisposes an individual towards developing a certain disease (Bell and Todd, 1989).

Interpretations of the results of disease association studies must take in to account the phenomenon of linkage disequilibrium. Infrequent recombination between loci results in linkage disequilibrium between alleles at those loci in the population; that is, the frequency at which alleles at two loci are found together in the population is greater than would be expected from the individual frequency of the alleles. It is well documented that there is very strong linkage disequilibrium between genes in the MHC region (Klein, 1986). In the class II region,
<table>
<thead>
<tr>
<th>Condition</th>
<th>HLA Allele</th>
<th>Frequency (%)</th>
<th>Relative Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hodgkin’s disease</td>
<td>A1</td>
<td>40</td>
<td>32.0</td>
</tr>
<tr>
<td>Idiopathic hemochromatosis</td>
<td>A3</td>
<td>76</td>
<td>28.2</td>
</tr>
<tr>
<td>Behçet’s disease</td>
<td>B14</td>
<td>16</td>
<td>3.8</td>
</tr>
<tr>
<td>Congenital adrenal hyperplasia</td>
<td>B47</td>
<td>9</td>
<td>0.6</td>
</tr>
<tr>
<td>Ankylosing spondylitis</td>
<td>B27</td>
<td>90</td>
<td>9.4</td>
</tr>
<tr>
<td>Reiter's disease</td>
<td>B27</td>
<td>79</td>
<td>9.4</td>
</tr>
<tr>
<td>Acute anterior uveitis</td>
<td>B27</td>
<td>52</td>
<td>9.4</td>
</tr>
<tr>
<td>Subacute thyroiditis</td>
<td>B35</td>
<td>70</td>
<td>14.6</td>
</tr>
<tr>
<td>Psoriasis vulgaris</td>
<td>Cw6</td>
<td>87</td>
<td>33.1</td>
</tr>
<tr>
<td>Dermatitis herpetiformis</td>
<td>DR3</td>
<td>83</td>
<td>26.3</td>
</tr>
<tr>
<td>Celiac disease</td>
<td>DR3</td>
<td>79</td>
<td>26.3</td>
</tr>
<tr>
<td>IgA deficiency in blood donors</td>
<td>DR3</td>
<td>64</td>
<td>26.3</td>
</tr>
<tr>
<td>Sicca syndrome</td>
<td>DR7</td>
<td>Also increased</td>
<td></td>
</tr>
<tr>
<td>Idiopathic Addison’s disease</td>
<td>DR3</td>
<td>78</td>
<td>26.3</td>
</tr>
<tr>
<td>Graves’ disease</td>
<td>DR3</td>
<td>69</td>
<td>26.3</td>
</tr>
<tr>
<td>Insulin-dependent diabetes mellitus</td>
<td>DR3 and/or DR4</td>
<td>91</td>
<td>57.3</td>
</tr>
<tr>
<td>Myasthenia gravis</td>
<td>DR2</td>
<td>10</td>
<td>30.5</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>DR3</td>
<td>70</td>
<td>28.2</td>
</tr>
<tr>
<td>Idiopathic membranous nephropathy</td>
<td>DR3</td>
<td>75</td>
<td>20.0</td>
</tr>
<tr>
<td>Zn*-immunized mothers</td>
<td>DR3</td>
<td>95</td>
<td>15</td>
</tr>
<tr>
<td>Narcolepsy</td>
<td>DR2</td>
<td>100</td>
<td>22</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>DR2</td>
<td>59</td>
<td>25.8</td>
</tr>
<tr>
<td>Optic neuritis</td>
<td>DR2</td>
<td>46</td>
<td>25.8</td>
</tr>
<tr>
<td>C2 deficiency</td>
<td>DR2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goodpasture’s syndrome</td>
<td>DR2</td>
<td>88</td>
<td>32.0</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>DR4</td>
<td>50</td>
<td>19.4</td>
</tr>
<tr>
<td>Pemphigus (in Jews)</td>
<td>DR4</td>
<td>87</td>
<td>32.1</td>
</tr>
<tr>
<td>IgA nephropathy</td>
<td>DR4</td>
<td>49</td>
<td>19.5</td>
</tr>
<tr>
<td>Hydralazine-induced SLE</td>
<td>DR4</td>
<td>73</td>
<td>32.7</td>
</tr>
<tr>
<td>Postpartum thyroiditis</td>
<td>DR4</td>
<td>72</td>
<td>32.2</td>
</tr>
<tr>
<td>Hashimoto’s thyroiditis</td>
<td>DR5</td>
<td>19</td>
<td>6.9</td>
</tr>
<tr>
<td>Pernicious anemia</td>
<td>DR5</td>
<td>25</td>
<td>5.8</td>
</tr>
<tr>
<td>Juvenile rheumatoid arthritis</td>
<td>DRw8</td>
<td>23</td>
<td>7.5</td>
</tr>
<tr>
<td>Primary glomerulonephritis</td>
<td>C4B*2.9</td>
<td>25</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Table 1.2.
Associations between the MHC and diseases. The relative risk shows how many times more frequently the disease occurs amongst individuals with a particular MHC allele compared to those without that allele (from Klein, 1986).
alleles at the DR and DQ loci are particularly strongly associated. DP alleles in general are not strongly associated with the DQ/DR region, providing evidence for a hot spot of recombination in the proximal class II region, although in some haplotypes significant disequilibrium between DP and DR/DQ is seen (Rosenberg et al., 1989). The practical significance of linkage disequilibrium is that if a disease association is found with a particular class II allele, any gene (including known class II genes and, theoretically, novel genes) in linkage disequilibrium with that allele is a candidate for the disease susceptibility gene (Bell and Todd, 1989).

Many of the diseases which are associated with the class II region have an autoimmune pathology and consequently most efforts to explain these associations have focussed on the genes encoding the classical class II antigens. Since the products of these genes clearly play an important role in the control of the immune response (section 1.2.2.), the classical class II genes are intuitively excellent candidates for autoimmune disease susceptibility genes. The sequences of numerous class II alleles with which disease associations have been detected, or which are in linkage disequilibrium with the marker alleles, have been determined in an attempt to identify shared residues or epitopes in the encoded class II molecules which could explain why particular haplotypes predispose to disease (Bell and Todd, 1989). As a result of these studies, particular class II antigens have been strongly implicated in the mechanism of development of IDDM and rheumatoid arthritis (Todd et al., 1988; Nepom, 1990). Shared amino acids in these allelic products are proposed to give them common structural features which fail to delete potentially autoreactive T-cells in the thymus or which facilitate presentation of critical autoantigens in the periphery (Bell and Todd, 1989).

There is, however, still no proof that the class II genes implicated by these studies explain the association of any disease with the class II region. Thus it is possible that the true autoimmune disease susceptibility genes are as yet undiscovered class II genes or genes encoding accessory functions in antigen presentation (section 1.6.). Furthermore, not all class II-associated diseases have an autoimmune pathology. Almost 100% of individuals suffering from the sleep
disorder narcolepsy have DR2(Drw15) yet show no evidence of an autoimmune response (Aldrich, 1990). Other examples of non-autoimmune diseases in association with the class II region include Hodgkin's lymphoma, chronic lymphocytic leukaemia and acute non-lymphocytic leukaemia, all of which are associated with DP alleles (Bodmer et al., 1989a; Pawelec et al., 1989). In these cases the association with the class II region may be explained by the presence of novel genes with non-immunological functions. A precedent for this latter situation in the human MHC was the association of congenital adrenal hyperplasia with HLA-Bw47. The molecular basis of this disease was actually a deletion in the functional steroid 21-hydroxylase gene which was in linkage disequilibrium with HLA-Bw47 (White et al., 1985). Thus, a complete understanding of class II-disease associations will depend on characterising all of the genes in this region and determining their function.

1.6. Evidence for other genes in the class II region

1.6.1. Novel class II genes
Evidence for the existence of a novel human class II antigen has been reported by Carra and Accolla (1987), who isolated a monoclonal antibody which immunoprecipitated a class II molecule from B-lymphoblastoid cell lines lysates after the lysates had been cleared of DR, DQ and DP molecules. The immunoprecipitated molecules contained α- and β-chains as judged by SDS-polyacrylamide gel electrophoresis, but 2-D peptide mapping studies revealed that these were distinct from the DR, DQ and DP α- and β-chains present in the parental cell line. As yet, the genes encoding this novel antigen have not been identified.

It may be that the novel class II molecule described in this study is a product of the DNA locus with that of a hitherto undiscovered DNB locus, or a product of the DOB locus with that of a hitherto undiscovered DOA locus. As discussed in section 1.3.3. both the DNA and DOB genes are transcribed and the nucleotide sequences of the corresponding cDNA clones did not reveal and deleterious mutations which would prevent translation (Tonnelle et al., 1985; Young and
Trowsdale, 1990). DNA homologues have been described in whale and rabbit, and the rabbit gene is expressed, but the potential of these genes to encode a protein product has not yet been determined (Kulaga et al., 1987; Trowsdale et al., 1989). The mouse homologue of DOB, Ob, is also transcribed and potentially functional, and intuitively it seems likely that DOB, which clearly arose before the major mammalian radiation, would only be maintained in both mouse and man in a potentially functional form if it did indeed encode a protein product. It is therefore possible that the human class II region contains functional DNB or DOA genes. Alternatively, the class II molecule described by Carra and Accolla (1987) may be the product of a completely novel class II α/β gene pair. The approaches taken previously to isolate novel α- and β-chain genes involved the screening of genomic or cDNA libraries with class II gene probes at reduced stringency. These methods would not detect diverged class II genes. It should be mentioned that any novel class II gene would not necessarily be encoded within the class II region, although class II genes unlinked to the MHC region have not yet been described in any species.

1.6.2. Class I-modifying locus
Evidence for a non-class II gene potentially mapping in the class II region has come from studies of a mutant human B-lymphoblastoid cell line LBL 721.174 which has a defect in the presentation of antigen to T-cells by class I molecules (DeMars et al., 1985; Cerundolo et al., 1990). Class I molecules present short peptide fragments derived from the degradation of intracellular proteins, such as viral proteins, to cytotoxic T-lymphocytes (CTL), which then become activated to kill the infected cell. The class I molecules in LBL 721.174 are functionally normal because they could present exogenously added peptide fragments to CTL such that killing ensued. However, virally infected LBL 721.174 cells were not killed and the class I molecules were retained in the endoplasmic reticulum (which is where antigen binding to class I molecules is thought to occur) instead of progressing to the cell surface. This phenotype has been interpreted as showing that LBL 721.174 has a defect in the transport of intracellularly derived peptides from their site of generation in the cytosol to their site of class I binding in the ER (Cerundolo et al., 1990). Understanding the molecular basis of the defect in this mutant will provide important
clues as to the way in which class I molecules become complexed with antigen, about which little is known. LBL 721.174 has a deletion spanning from the DPB2 gene to the complement gene cluster, and the gene responsible for the antigen presentation defect is therefore likely to map within this interval (Cerundolo et al., 1990; DeMars et al., 1985). A strikingly similar phenotype has also been described in the rat, and in this case the gene responsible for the defect in class I antigen presentation could be mapped to the class II region of the rat MHC (Livingstone et al., 1989). Like the mouse class II region (section 1.4.2.), the rat class II region is homologous to the human class II region, and therefore the observation of Livingstone et al. (1989) provides additional evidence for the presence of a class I-modifying locus in the human class II region.

1.6.3. LMP antigens
Evidence for novel genes in the mouse MHC was obtained when an antiserum made between mice differing only in the MHC region was found to precipitate a large (~580kDa) multisubunit protein complex from a mouse macrophage cell line (Monaco and McDevitt, 1982). The complex was composed of a large number of noncovalently linked low molecular weight polypeptide (LMP) subunits which were biochemically, serologically and genetically distinct from the class I, class II and class III gene products. The subunits ranged in molecular weight from 12-35kD. Two of the subunits displayed electrophoretic polymorphism and both the polymorphisms mapped by recombination analysis within the mouse class II region. Lmp-7 was localised between Pb and Ab, while Lmp-2 was mapped slightly more accurately, between Pb and Ob (Monaco and McDevitt, 1986; Steinmetz et al., 1986). Neither gene has yet been cloned. The genes for the other fourteen LMP subunits could not be mapped in this study because their products were not polymorphic. A biochemically similar complex has also been described in human cells (Monaco and McDevitt, 1984). Therefore, given that the human and mouse class II regions are homologous, it is possible that the human class II region also encodes LMP subunits. The function of the LMP complex is unknown, although the fact that it is expressed in macrophages and lymphocytes and is inducible by γ-interferon, like class II molecules, has led to speculation that it may provide some accessory function in antigen presentation (Monaco and
McDevitt, 1986). The unusual properties of the LMP complex are also found in the eukaryotic multicatalytic proteinase (Rivett, 1989). This broad specificity non-lysosomal endopeptidase complex of about 600kD is composed of at least thirteen distinct subunits with molecular weights of between 20 and 35kD. An intriguing possibility is that the LMP complex is the same as, or closely related to, the high molecular weight proteinase, and generates peptides from intracellular proteins for presentation by MHC molecules (Parham, 1990).

1.7. Approaches to finding novel genes

There is substantial circumstantial evidence to support the hypothesis, on which the experiments described in this thesis are based, that there are previously undiscovered genes in the class II region of the human MHC. To summarise briefly the preceding sections, evidence that there may be additional class II genes in the class II region comes from (i) immunoprecipitation of non-DP-DQ-DR class II molecules from B-lymphoblastoid cell lines and (ii) the finding of two potentially functional class II genes, each without a 'partner'. Evidence for non-class II genes in the class II region comes from (i) studies of class I antigen presentation mutants in human and rat which implicate a class II-encoded locus and (ii) the mapping of genes encoding two subunits of a novel multisubunit complex to the mouse class II region. In addition, disease association studies can be interpreted in terms of previously undiscovered genes (both class II and non-class II) in the class II region.

In order to analyse a region of the genome of interest for coding sequences in the absence of detailed knowledge about the function of any gene which might be present, a 'reverse genetics' strategy can be applied (Orkin, 1986). Reverse genetics methods have provided spectacular successes in the cloning of the genes causing diseases which could not be analysed using more traditional approaches, such as Duchenne muscular dystrophy and cystic fibrosis (Monaco and Kunkel, 1987; Rommens et al., 1989). Over the last three years, this approach has also proved highly successful in the discovery of novel genes in the class III region of the human MHC, as described in section 1.3.4. Some
of the methods which can be used have already been alluded to, but the principles will now be described in more detail.

Once the genomic region of interest has been identified it is mapped and cloned. The most powerful technique for long-range physical mapping is currently pulsed field gel electrophoresis (Barlow and Lehrach, 1987), although irradiation-induced hybrids and naturally occurring chromosomal translocations and deletions have also proved extremely useful for determining the relative order of probes in a given region (Hastie et al., 1988; Cox et al., 1990). The region of interest is then cloned, using the previously mapped probes as starting points for the isolation of genomic clones. Genomic libraries in bacteriophage λ and cosmid vectors have previously proven successful, although the relatively small insert sizes in these systems can make the cloning of large regions of DNA extremely laborious. More recently however, yeast artificial chromosome (YAC) vectors have been developed in which hundreds of kilobases of genomic DNA can be cloned and propagated (Schlessinger, 1990). Jumping libraries and linking libraries provide additional sources of cloned genomic material (Poustka and Lehrach, 1986). The use of rare-cutter jumping libraries has the additional advantage that the cloned regions may contain CpG islands which are frequently found at the 5' ends of genes (see below and Chapter 3).

The methods used to search for genes in a cloned region are based on the known properties of transcribed sequences. For example, many genes are associated with CpG islands, short (1-2kb) regions of DNA which are unusually rich in unmethylated CpG dinucleotides (Bird, 1987). The function of these regions is not yet known, but they provide convenient diagnostic markers for expressed sequences, and can be detected because they contain sites for certain restriction endonucleases which otherwise cleave DNA very infrequently (Brown and Bird, 1986; Lindsay and Bird, 1987). Another diagnostic feature of coding sequences is that they are likely to be conserved in the genomes of other organisms. Cross-hybridisation of fragments from genomic clones to genomic DNA from other species has proved a successful approach to identifying genes (Monaco et al., 1986). Furthermore, transcribed regions can be detected by hybridising fragments from genomic clones
to northern blots or to cDNA libraries (Monaco et al., 1986). Very recently, 'exon-trapping' techniques have been developed to test for the presence of splice junctions in a region of cloned DNA (Duyk et al., 1990).

These approaches are theoretically applicable to any part of the human genome, including the class II region. In fact, as described in the preceding sections, many steps have already been taken towards characterising the class II region at the molecular level. In particular, physical maps of the region are available which can be used as a basis for further work, and numerous genes have been cloned which are suitable for use as probes in the construction of more detailed maps and the isolation of new genomic clones. The following chapters of this thesis describe the application of molecular mapping and cloning techniques in the class II region to determine the likely positions of genes, to obtain genomic clones covering these regions, to identify genes within these clones, and to characterise the novel genes thus found.
2. Materials and methods

2.1. Bacterial cell culture

Liquid cultures were grown by inoculating L-broth containing the appropriate antibiotic with a single bacterial colony and shaking vigorously overnight at 37°C. To grow bacterial colonies, cells were streaked onto the surface of L-agar plates containing the appropriate antibiotic and incubated with the plates inverted at 37°C overnight.

<table>
<thead>
<tr>
<th>L-broth</th>
<th>Bacto-tryptone</th>
<th>10g/litre</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bacto-yeast extract</td>
<td>5g/litre</td>
</tr>
<tr>
<td></td>
<td>NaCl</td>
<td>10g/litre</td>
</tr>
</tbody>
</table>

Sterilised by autoclaving.

<table>
<thead>
<tr>
<th>L-agar plates</th>
<th>1.5% (w/v) Bacto-agar in L-broth.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sterilised by autoclaving.</td>
</tr>
<tr>
<td></td>
<td>Solid media were melted by microwaving and cooled to 50°C before adding antibiotics.</td>
</tr>
</tbody>
</table>

**Antibiotics**

- **Ampicillin**
  - Anhydrous ampicillin was dissolved in double distilled water (DDW) at 50mg/ml, sterilised by filtration and used at a final concentration of 50μg/ml.

- **Tetracycline**
  - Tetracycline hydrochloride was dissolved in DDW at 12.5mg/ml, sterilised by filtration and used at a final concentration of 12.5μg/ml.

- **Kanamycin**
  - Kanamycin sulphate was dissolved in DDW at 10mg/ml, sterilised by filtration and used at a final concentration of 10μg/ml.

Antibiotics were stored at -20°C.
2.2. Screening of recombinant DNA libraries

2.2.1. Cosmid libraries

The cosmid library used in this study, a gift from Dimitri Kioussis, (National Institute of Medical Research, London) was made using genomic DNA from the human T-cell line HPB.ALL in the cosmid vector cos202 (ampicillin resistant). \(10^6\) recombinants were plated out and screened as follows. The titre of the cosmid library was first determined by preparing a \(10^{-2}\) dilution of the frozen library stock (10µl stock in 990µl L-broth) and plating out dilutions of this on L-agar plates containing ampicillin. These plates were incubated at 37°C overnight and the \(10^{-2}\) dilution was stored at 4°C. The following day the number of colonies obtained for each dilution was counted and used to calculate the volume of the \(10^{-2}\) dilution required to yield 250,000 colonies (the optimal number for a 20x20cm filter). Four 20x20cm Hybond-N membranes (Amersham) marked with an asymmetric pattern of dots were lowered onto the surface of four L-agar 245x245mm (Nunc) plates. The calculated volume of the \(10^{-2}\) dilution was then spotted on to the surface of the filter and spread evenly using a flamed glass spreader. These plates, the master plates, were incubated overnight at 37°C. The next day eight additional 245x245mm plates, the replica plates, were poured and overlaid with nylon membranes. To prepare duplicate replicas of the first master, the filter from the first master plate was lifted off the agar with Millipore forceps and placed colony side up on three thicknesses of Whatman 3mm paper. The wetted filter from the first replica plate was then overlaid on this filter, covered with further layers of Whatman and pressed on to the master. The master pattern of spots was marked on to the replica filter before the two were pulled apart and the replica was returned to its plate. The same master filter was then overlaid with a second replica filter and the process repeated. The first master filter was then returned to its plate. When duplicate replicas had been prepared from the four master filters, all twelve plates were incubated at 37°C until the colonies had regrown. The master plates were then sealed with Parafilm and stored at 4°C.

The filters from the replica plates were removed from the agar surface and processed by placing them successively (colony side up) on pads of
Whatman soaked in denaturing solution (7min) and neutralising solution (2x 3min). The filters were then rinsed by immersion in 2x standard saline citrate (SSC) and air dried before the DNA was fixed by baking for 2hr at 80°C or UV irradiation at 0.4 J/cm². The membranes were incubated for 1-2hr at 42°C in 500ml TENS buffer and the bacterial debris removed with a rubber policeman. After a final rinse in 2x SSC the filters were ready for hybridisation

Following hybridisation of the replica filters with the probe of interest, a positive hybridisation signal present in duplicate was identified on the autoradiographs and used to pinpoint on the appropriate master plate the position of the region containing the positive colony. Bacterial cells were removed from this region with a toothpick or flamed loop and transferred to 1ml L-broth. After vigorous mixing to disperse the cells, dilutions of this stock were made and plated out on 140mm L-agar plates containing ampicillin and incubated at 37°C overnight. A plate bearing well separated colonies was selected for the preparation of duplicate colony lifts. A nylon membrane of appropriate size was overlaid on the agar surface for 1min. During this time the position of the filter was marked by piercing with a needle and syringe containing ink. The filter was then removed and placed successively (colony side up) on pads of Whatman soaked in denaturing solution (7min) and neutralising solution (2x 3min). The filter was briefly rinsed in 2x SSC and the DNA fixed onto the surface by baking or UV irradiation. Meanwhile, a second membrane was overlaid on the surface of the plate and marked in the same position with the needle and ink. This replica was processed in the same way as the first. The secondary plate was incubated at 37°C until the colonies had regrown, then stored at 4°C. Following hybridisation of the secondary filters with the probe of interest the autoradiographs were used to identify the positions of individual positive colonies on the plates. These were picked and the cosmid DNA purified as described in section 2.5.1.

Denaturing solution

<table>
<thead>
<tr>
<th></th>
<th>1.5M NaCl</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5M NaOH</td>
</tr>
</tbody>
</table>

53
Neutralising solution
1.5M NaCl
0.5M Tris.Cl pH7.2
1mM EDTA

20x SSC
3M NaCl
0.3M sodium citrate

TENS buffer
50mM Tris.Cl pH8.0
1mM EDTA
1M NaCl
0.1% SDS

2.2.2. cDNA libraries
The cDNA libraries used in this study were constructed according to the method of Seed (1987) in the plasmid vector CDM8 or derivatives thereof. For each library, 2.5x10^5-1x10^6 clones, propagated in *E. coli* MC1061/p3, were plated out and screened in the same way as the cosmid library, using ampicillin + tetracycline selection. Once a single positive colony had been identified, plasmid DNA was purified as described in section 2.5.1. The CEM (T-cell line) cDNA library was a gift from Jenny Dunne (Lymphocyte Molecular Biology Laboratory, ICRF). The JY (B-cell line) and γ-interferon induced macrophage (U937) cDNA libraries were a gift from Dr. David Simmonds, Oxford.

2.2.3. Jumping libraries
The rare-cutter jumping libraries used in this study were a gift from Annemarie Poustka (German Cancer Research Centre, Heidelberg, FRG). Both libraries were constructed as described in Poustka and Lehrach (1988; see also Figure 5.5.). The NotI jumping library was constructed from NotI-cut human genomic DNA which was circularised around the marker plasmid and then re-cut with BamHI. The BssHIII jumping library was constructed from BssHIII-cut human genomic DNA which was circularised around the marker plasmid and then re-cut with BamHI+HindIII. In both cases the marker plasmid was pMLS-Mlu-Not, which carries the supF gene, and the vector was a modified form of the bacteriophage λ strain NM1151 which contains amber mutations that are suppressed in the presence of supF. Jumping
clones were propagated as temperature sensitive lysogens of phage λ recombinants in E. coli MC1061/p3 (ampicillin + tetracycline selection). The library was handled and screened in the same way as the cosmid library except that incubations were done at 30°C to maintain lysogenic growth. However, once an individual positive bacterial colony was detected by secondary screening the phage were induced to undergo lytic growth so that preparations of λ DNA could be made. A positive colony was inoculated into 5ml L-broth containing tetracycline and ampicillin and shaken at 30°C overnight. 1ml was diluted into 50ml L-broth containing antibiotics and shaken at 30°C for 90min. The lytic growth cycle was then induced by shaking the culture at 42°C for 20min. Phage DNA was prepared as described in section 2.6., starting with the step to pellet the lysed bacterial cells.

2.3. Subcloning of DNA fragments

DNA fragments were subcloned into the plasmid vector Bluescript (ampicillin resistant), which has a useful polylinker cloning site and carries the lacZ gene which facilitates colour selection of insert-carrying recombinants. Vector for subcloning was cleaved to completion (as judged by testing a sample on an agarose gel) with the appropriate enzyme(s) and then purified by phenol extraction and ethanol precipitation. Insert DNA was prepared by excising the desired fragment from a 1x TAE agarose gel and purifying the DNA using Geneclean (Bio 101). For the ligation reaction, vector and insert DNA were mixed in the ratio [1 vector terminus: 2 insert termini] and incubated overnight at 16°C with 1µl 10mM ATP, 1µl 10x ligation buffer, 1µl T4 DNA ligase (Biolabs) and DDW in a total volume of 10µl. The reaction was then ethanol precipitated and resuspended in 10µl DDW for transformation into the Bluescript host E. coli XL-1 Blue (Stratagene) by electroporation (section 2.4.2.).

10x Ligation buffer
500mM Tris.Cl pH8.0
100mM MgCl2
200mM dithiothreitol (DTT)
500µg/ml BSA
2.4. Transformation of bacterial cells with DNA by electroporation

2.4.1. Preparation of *E. coli* cells for electroporation

5ml L-broth were inoculated with a single colony of the desired *E. coli* strain and shaken at 37°C overnight. 2.5ml of this culture were diluted into 500ml L-broth in a 2 litre flask and shaken at 37°C until OD$_{600}$ was 0.5-0.6. The culture was then chilled in an ice-water bath for 15min, transferred to pre-chilled 500ml centrifuge bottles, and spun for 20min, 4000rpm, 2°C, in a Beckman J6B. The supernatant was poured off and the pellet resuspended in 5ml ice cold water. A further 500ml ice cold water were added before centrifuging at 10000rpm, 2°C, 20min. The supernatant was poured off quickly to avoid loss of the loose pellet. The cells were then resuspended and centrifuged as before. The supernatant was removed immediately and the pellet resuspended by swirling in the residual liquid. 40ml ice-cold 10% glycerol were added and the cells were pelleted at 10 000rpm, 10min, 2°C. Finally the cells were resuspended in an equal volume of 10% glycerol, aliquotted into Eppendorfs (50µl each), frozen on dry ice and stored at -70°C (Ausubel et al., 1987).

2.4.2. Introduction of DNA into cells

DNA in ligation buffer was ethanol precipitated and resuspended in water before transformation. The salts in ligation mixes lowered the efficiency of transformation because the resistance of the sample was reduced to sub-optimal levels.

DNA (typically 1-100ng) was added to 50µl electroporation-competent cells and the mixture was pipetted into the bottom of a pre-chilled electroporation cuvette. This was placed in the sample chamber of the electroporation apparatus, which was set to 2.5kV, 25µF with the pulse controller adjusted to 200Ω. The pulse was applied and the cuvette removed. 1ml SOC medium was immediately added and the mixture pipetted into a fresh Eppendorf and incubated at 37°C for 30-60min.
Aliquots of the transformation mix were plated out on L-agar plates containing the appropriate antibiotic. When Bluescript plasmids were used, the plates were spread with 100μl (for an 82mm plate) IPTG and 40μl X-gal 30min before use. The efficiency of transformation by electroporation was so high for supercoiled plasmids (typically 10⁸-10⁹ transformants/μg) that extensive dilutions of transformation mixes were made in SOC in order to obtain single colonies (Ausubel et al., 1987).

SOC Medium

<table>
<thead>
<tr>
<th>Ingredient</th>
<th>Concentration</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5% Bacto yeast extract</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2% Bacto-tryptone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5mM KCl</td>
<td>10mM NaCl</td>
<td></td>
</tr>
<tr>
<td>10mM MgCl₂</td>
<td>10mM MgSO₄</td>
<td></td>
</tr>
<tr>
<td>20mM glucose</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sterilised by autoclaving.

IPTG

25mg/ml isopropyl thiogalactoside in DDW, stored at -20°C.

X-Gal

25mg/ml bromo-chloro-indolyl-β-D-galactoside in dimethyl formamide, stored at -20°C.

2.5. Preparation of DNA from transformed bacterial cells

2.5.1. Small scale preparation of plasmid or cosmid DNA by alkaline lysis
This technique was used for the rapid small-scale purification of plasmid or cosmid DNA from bacterial cells (Sambrook et al., 1989).

5ml L-broth were inoculated with a single bacterial colony and shaken overnight at 37°C. 1.5ml of each culture were transferred to an Eppendorf tube and microfuged for 1min. The supernatant was removed by aspiration with a drawn-out Pasteur pipette. The cell pellet was resuspended by vortexing in 100μl GTE and the tubes were left at room temperature for 5min. The cells were lysed by adding 200μl freshly prepared 0.2M NaOH/1% SDS which was mixed by inverting the tube. After standing on ice for 5min the bacterial chromosomal DNA was
precipitated by adding 150μl KAc pH4.8, vortexing, and standing on ice for a further 5min. The chromosomal DNA was then pelleted by microfuging for 10min at 4°C, and the supernatant was transferred to a fresh Eppendorf tube. An equal volume of phenol/chloroform/iso-amyl alcohol (PCIA) was added and mixed thoroughly by vortexing, and the tubes were microfuged for 5min. The aqueous phase containing the plasmid DNA was transferred to a fresh Eppendorf tube. Two volumes of absolute ethanol were added and mixed before leaving at -20°C for 2hr. Following precipitation the DNA was pelleted by microfuging for 15min at 4°C. The supernatant was discarded before addition of 500μl 70% ethanol. The tubes were microfuged again and the ethanol removed by aspiration. The pellets were dried for 5min by vacuum desiccation and then resuspended in 20μl TE pH8.0 containing 100μg/ml DNase-free RNase. 2-5μl were analysed by restriction endonuclease digestion. This technique typically yielded 2μg cosmid DNA or 5μg plasmid DNA.

GTE  
50mM glucose  
10mM EDTA  
25mM Tris.Cl pH8.0

KAc pH4.8  
60ml 5M KAc  
11.5ml glacial acetic acid  
28.5ml DDW

PCIA  
Phenol (melted at 65°C) was mixed with chloroform and iso-amyl alcohol in the ratio 25:24:1 and buffered by equilibrating once with 1 volume 50mM Tris base, twice with 1 volume 50mM Tris.Cl pH8.0 and once with 1 volume TE pH8.0.

TE pH8.0  
10mM Tris.Cl pH8.0  
1mM EDTA pH8.0
DNase-free RNase Pancreatic RNase A was dissolved in 15mM NaCl, 10mM Tris.Cl pH7.5 at 10mg/ml, heated to 100°C for 15min and cooled slowly to room temperature before storing at -20°C.

2.5.2. Large scale preparation of plasmid or cosmid DNA by alkaline lysis
This technique was used to prepare milligram-scale quantities of high purity plasmid or cosmid DNA (Sambrook et al., 1989).

5ml L-broth were inoculated with a single bacterial colony and grown during the day (or overnight) at 37°C with vigorous shaking. This culture was then diluted into 400ml L-broth in a 2 litre flask (or split 2x 200ml in 1 litre flasks) and shaken at 37°C overnight. Bacterial cells were pelleted in a 500ml centrifuge bottle by centrifuging at 6000rpm for 10min and then resuspended thoroughly in 20ml GTE containing 4mg/ml lysozyme. This suspension was left at room temperature for 5min before adding 40ml 0.2M NaOH/0.1% SDS, mixing gently, and incubating for a further 5min on ice. 20ml 5M KAc pH4.8 were then added and mixed by vortexing. This mixture was incubated for 15min on ice and centrifuged for 15min at 8000rpm. The supernatant was filtered through a nylon tea-strainer into a 200ml glass bottle, and the nucleic acid precipitated by adding 0.6 volumes of propan-2-ol and incubating at -20°C for 2hr. The precipitate was pelleted by spinning at 2000rpm for 20min at 4°C and the pellet was resuspended in 5ml RNasing buffer. The solution was transferred to a 50ml Falcon tube before adding 10µl 10mg/ml RNase and incubating at 37°C for 15min. 200µl 20mg/ml proteinase K and 125µl 20% SDS were then added and the incubation continued at 37°C for a further 30min. An equal volume of PCIA was added and mixed thoroughly by vortexing. The phases were separated by centrifuging at 2000rpm for 15min, and the aqueous phase was transferred to a 30ml Corex tube. 0.1 volumes 3M NaAc pH5.2 and 2 volumes absolute ethanol were added and the mixture was stored at -20°C for 2hr. The DNA was pelleted by centrifugation at 10000rpm for 20min at 4°C and resuspended in 1ml A-50 buffer. The plasmid or cosmid DNA was then separated from contaminating bacterial DNA and RNA by fractionation through a
30ml A-50 biogel (Bio-Rad) column. Fractions contributing to the first OD260 peak were pooled in a 30ml Corex tube and ethanol precipitated. The DNA was pelleted by spinning at 10000rpm for 20min at 4°C, the supernatant was poured off and the pellet was washed in 5ml 70% ethanol before centrifuging again. The pellet was dried under vacuum and resuspended in 500μl TE pH8.0.

**RNasing buffer**

- 0.1M NaCl
- 5mM EDTA
- 0.1M Tris.Cl pH8.0

**Proteinase K**

Fungal proteinase K (BDH) was dissolved at 20mg/ml in 20mM Tris.Cl pH8.0, self-digested at 37°C for 1hr, and frozen at -20°C.

**Biogel A-50 buffer**

- 0.5M NaCl
- 25mM Tris.Cl pH8.0
- 1mM EDTA

### 2.6. Small scale preparation of bacteriophage λ DNA

100μl phage were mixed with 100μl 10mM CaCl₂, 10mM MgSO₄ and 100μl of a saturated culture of host bacterial cells and incubated for 20min at 37°C. This mixture was diluted in 50ml L-broth containing 10mM MgSO₄, 0.5% casamino acids and 0.2% maltose and shaken overnight in a 200ml conical flask at 37°C. 20ml aliquots were transferred to Falcon tubes and centrifuged at 4000rpm for 15min to pellet the debris of lysed bacterial cells. 15ml of supernatant were transferred to a fresh tube and incubated with 10μl 10mg/ml DNase for 15min at 37°C to degrade bacterial DNA. 5ml of AS buffer were then added and the mixture heated to 70°C for 15min to denature the phage capsids. The tube was cooled under the cold tap and 1.7ml KAc pH4.8 were added before centrifugation at 4000rpm for 15min. The supernatant was filtered through a tea strainer into a fresh Falcon tube, and 15ml propan-2-ol were added before centrifugation at 4000rpm for 15min. The supernatant was poured off and the pellet resuspended in 200μl 0.3M NaAc. The phage DNA was transferred to a microfuge tube,
treated with RNase for 15min at 37°C, and extracted with PCIA. The aqueous phase was transferred to a fresh microfuge tube and precipitated with 2 volumes absolute ethanol at -20°C. The DNA was pelleted by microfugation and washed with 70% ethanol. Following a second spin the supernatant was removed by aspiration, the pellet was dried for 1-2min under vacuum and the DNA was resuspended in 50-100µl DDW. 10µl were analysed by restriction enzyme digestion.

\[
\begin{align*}
\text{AS buffer} & : 0.3 \text{M Tris.Cl pH8.0} \\
& 0.15 \text{M EDTA} \\
& 1.5 \% \text{ SDS}
\end{align*}
\]

2.7. Preparation of eukaryotic DNA from cells in culture

Many of the human genomic DNA samples used to prepare Southern blots in this study were a gift from Susan Tonks (Tissue Antigen Laboratory, ICRF). Genomic DNA samples from other vertebrate species for the preparation of 'zoo' blots were obtained from Dr. Nigel Spurr (Clare Hall Laboratories, ICRF). Additional genomic DNA samples were prepared using the following protocol (Susan Tonks, personal communication).

Cells were grown in RPMI medium supplemented with 10% foetal calf serum (Gibco) at 37°C in 5% CO₂. For DNA preparation, 10⁸-10⁹ cells were pelleted at 3000rpm for 5min and resuspended in 10ml PBSA. This procedure was repeated. 90ml sucrose-triton were then added and the mixture centrifuged at 10000rpm for 10min. The supernatant was discarded and the pellet resuspended in 20ml 75mM NaCl, 24mM EDTA pH8.0. 40µl 10mg/ml RNase, 500µl 20% SDS and 200µl 20mg/ml proteinase K were added and the mixture was incubated at 37°C for 3hr. An equal volume of buffered phenol was added and mixed thoroughly by shaking. The phases were separated by spinning at 2000rpm for 5min and the aqueous phase was transferred to a fresh tube. The process was repeated using PCIA until the aqueous phase was clear. The aqueous phase was then extracted with chloroform/isoamyl alcohol (24:1) and transferred to a beaker. 0.1 volumes 0.5M NaCl and 2 volumes of absolute ethanol were then added and the precipitated DNA strands
were spooled onto a Pasteur pipette and transferred to an Eppendorf. 1ml TE was then added and the DNA left to dissolve at 4°C. The OD was assayed at 260 and 280nm to assay the ratio of DNA to protein. If \( \text{OD}_{260}/\text{OD}_{280} < 1.75 \), the sample was incubated again with proteinase K and re-extracted. A typical yield from 10^8 cells was 200\( \mu \)g DNA.

<table>
<thead>
<tr>
<th>PBSA</th>
<th>171mM NaCl</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.4mM KCl</td>
</tr>
<tr>
<td></td>
<td>10mM Na_2HPO_4</td>
</tr>
<tr>
<td></td>
<td>1.8mM KH_2PO_4</td>
</tr>
</tbody>
</table>

| Sucrose-Triton | 0.33M sucrose  |
|               | 10mM Tris.Cl pH7.5       |
|               | 5mM MgCl_2               |
|               | 1% Triton-X 100          |

2.8. Preparation of very high molecular weight DNA for pulsed field gel electrophoresis

DNA for pulsed field gel electrophoresis (PFGE) was prepared in the form of cells encased in blocks of low melting point agarose (LMPA) to avoid shearing (Dr. Denise Barlow, personal communication). Each LMPA block contained 3x10^6 cells, enough for three digests.

Approximately 10^8 cells were pelleted at 1000rpm for 5min, the supernatant was removed by aspiration and the cells resuspended in 10ml PBSA. The cells were counted using a haemocytometer, then pelleted as before and resuspended in a further 10ml PBSA. Finally, the cells were pelleted again and resuspended in 50\( \mu \)l PBSA per 3x10^6 cells. An equal volume of 1% LMPA in PBSA, melted and equilibrated at 42°C, was added to the cells and mixed. 100\( \mu \)l aliquots were pipetted into the slots of pre-chilled block formers (LKB) which had previously been soaked in 10% hydrogen peroxide for 30min to remove all traces of nucleases. The block formers were left on ice for 30min until the agarose had set. The blocks were then transferred to a Falcon tube containing 2.5ml 0.5M EDTA, 1% sodium laurylsarcosine and 1mg/ml proteinase K, and incubated at 50°C for 48hr. A second aliquot of
proteinase K was added about halfway through the incubation period. The tube was then topped up with TE pH8.0 and inverted several times to wash the blocks. The blocks were separated from the TE by passing through a tea strainer, then returned to the Falcon tube. This process was repeated three times with fresh TE. The TE was then replaced with TE containing 0.04 mg/ml PMSF (freshly prepared by dissolving 40 mg/ml in propan-2-ol) and incubated at 50°C for 30 min to inactivate any residual protease activity. This step was repeated and the blocks were then stored in 0.5 M EDTA pH8.0 at 4°C until required.

2.9. Preparation of RNA

Preparation and analysis of RNA was performed with the assistance of Dr. Ruth Lovering.

2.9.1. Isolation of total RNA from cells in culture

Cell lines were cultured in the Cell Production Unit, ICRF. RNA was isolated using RNase-free reagents, plasticware and glassware (Sambrook et al., 1989). Cells were pelleted by spinning at 2000 rpm, 5 min. The pellet volume was estimated and 5 volumes of guanidinium thiocyanate homogenisation buffer were added. Homogenisation was effected by drawing the lysate ten times into a syringe fitted with a 19-gauge needle. The homogenate was then layered onto a cushion of 5.7 M CsCl, 0.1 M EDTA in a polyallomer ultracentrifuge tube. The volume of the homogenate determined the size of the rotor used and the amount of CsCl required, as shown below.

<table>
<thead>
<tr>
<th>Rotor</th>
<th>Homogenate volume (ml)</th>
<th>CsCl (ml)</th>
<th>Spin time (hours)</th>
<th>Speed (rpm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW60</td>
<td>1.2</td>
<td>3.1</td>
<td>12</td>
<td>40 000</td>
</tr>
<tr>
<td>SW41</td>
<td>3.5</td>
<td>9.7</td>
<td>24</td>
<td>32 000</td>
</tr>
<tr>
<td>SW28</td>
<td>12.0</td>
<td>26.5</td>
<td>24</td>
<td>25 000</td>
</tr>
</tbody>
</table>

After centrifugation, the liquid in the tube except the last 500 µl was carefully removed with a pipette and the curved base of the tube, containing the RNA pellet, was cut off with a scalpel. The remaining supernatant in the base of the tube was carefully pipetted off and the
pellet was resuspended in 500μl 10mM Tris.Cl pH7.5, 5mM EDTA, 1% SDS. This was extracted twice with PCIA and the RNA was precipitated from the aqueous phase by adding 0.1 volumes 3M NaAc pH5.2 and 2 volumes absolute ethanol, and storing at -20°C overnight. The RNA was pelleted by microfuging at 4°C for 5min and the pellet was washed in 70% ethanol. After re-spinning the supernatant was removed and the pellet dried under vacuum. The RNA was resuspended in 200μl DDW.

<table>
<thead>
<tr>
<th>Homogenisation buffer</th>
<th>5M guanidinium thiocyanate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5mM sodium citrate</td>
</tr>
<tr>
<td></td>
<td>1% Sarkosyl</td>
</tr>
<tr>
<td></td>
<td>0.7% β-mercaptoethanol</td>
</tr>
</tbody>
</table>

To test for γ-interferon inducible gene expression in colon carcinoma cell lines, cells were incubated with 300 units of γ-interferon/ml for 36-48hr before RNA extraction (experiment performed by Adrian Kelly, this laboratory).

### 2.9.2. Isolation of poly(A)+ RNA

Polyadenylated RNA was purified 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 30min at room temperature. The mixture was then 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 and 2 volumes absolute ethanol. The mixture was frozen on dry ice to precipitate the RNA and then microfuged for 15min. The supernatant was removed and the pellet resuspended in low-salt buffer before storing the sample at -80°C.

### 2.10. Restriction endonuclease digestion of DNA

For routine mapping of plasmid, cosmid or bacteriophage DNA with restriction endonucleases, 100-500ng were typically digested in a volume of 20μl with a 2-5 fold excess of enzyme for 1-2hr. When
specific fragments were required for subcloning or probe preparation, the digest was scaled up accordingly.

For preparation of genomic Southern blots, 10μg genomic DNA were typically digested overnight in a volume of 50-100μl with a 5-fold excess of enzyme. A tenth of the digest was resolved on a minigel to monitor the extent of digestion. Incomplete digests were usually diluted into an equal volume of 1x restriction buffer and supplemented with more enzyme before continuing the incubation.

For preparation of PFGE blots, blocks (section 2.8) were washed free of EDTA storage buffer by rocking in 50ml TE at room temperature (2x 30min) followed by 30min in DDW. Blocks were then cut with a scalpel into thirds, each third containing enough DNA for one digest (10μg). These were pre-equilibrated in 1ml of the appropriate buffer before digestion with a 5-fold excess of enzyme in a total volume of 100μl.

All digests were carried out in the presence of 100mg/ml BSA. The volume of enzyme added was not allowed to exceed 10% of the total volume. Genomic digests were supplemented with 5mM spermidine if the salt concentration in the buffer exceeded 50mM. Digests were incubated at the optimum temperature as recommended by the manufacturer. Overnight digests involving enzymes from thermophilic bacteria (e.g. TaqI, BssHII) were covered with a layer of paraffin oil to prevent evaporation during incubation.

Four basic 10x buffers were prepared which were supplemented with additional salts if required, according to the manufacturers' recommendations.

10x L Buffer
100mM Tris.Cl pH7.5,
100mM MgCl₂, 10mM DTT

10x M Buffer
500mM NaCl, 100mMTris.Cl pH7.5,
100mM MgCl₂, 10mM DTT
2.11. Electrophoresis of DNA and RNA

2.11.1. Conventional agarose gel electrophoresis
All DNA samples were mixed with 0.2 volumes of the appropriate loading buffer before electrophoresis. Routine restriction endonuclease digests of plasmid, cosmid and bacteriophage DNA were resolved on 0.8% agarose gels in 0.5x TBE buffer containing 0.5μg/ml ethidium bromide using Bio-Rad Mini-Sub minigel apparatus (gel length 10cm). Digests of genomic DNA were resolved using Bio-Rad DNA Sub-Cell apparatus (gel length 20cm) in 1x TAE buffer, typically using 0.8% agarose gels which were ideal for most purposes. The agarose concentration was increased up to 1.2% for the resolution of smaller fragments. DNA size markers were HindIII-cut bacteriophage λ, BstEII-cut λ or HaeIII-cut φX174, depending on the size range of interest.

10x TBE

108g/l Tris base
55g/l Boric acid
20mM EDTA

10x TAE

48.4g/l Tris base
11.4ml/l glacial acetic acid
20mM EDTA

5x Loading buffer

50% glycerol
60mM EDTA
0.25% bromophenol blue
5x TBE or TAE
2.11.2. Pulsed field gradient gel electrophoresis
Pulsed field gradient gel electrophoresis (PFGE) was performed using the Pulsaphor apparatus (LKB). 250 ml of 0.8% agarose in 0.25x TBE were poured into the 20x20 cm casting frame supplied. The 16-well LKB comb was used to form the slots into which the digested blocks were inserted. The blocks were sealed into the slots with molten 1% LMPA. Size markers of yeast chromosomes (S. cerevisiae strain YP148, a gift from Denise Barlow, Genome Analysis Laboratory, ICRF) and concatemers of bacteriophage λ DNA were used. The gel tank contained 2.5 litres of 0.25x TBE buffer which was recirculated at 10°C. The electrodes were positioned in the standard double inhomogeneous field configuration as described in the LKB manual. The gel was run at 330V with a constant pulse time of between 45-60 sec for 22-26 hr, which resolved molecules of up to 900 kb.

2.11.3. Electrophoresis of RNA
RNA was resolved in formaldehyde agarose gels. 3 g agarose, 20 ml 10x MOPS buffer and 150 ml DDW were mixed and boiled in a microwave. The gel was cooled to 50°C and 24 ml 38% formaldehyde were added before pouring. RNA samples (10-15 µg total RNA or 1.5-3 µg poly(A)+ RNA) were desiccated under vacuum and resuspended in 5 µl deionised formamide, 1 µl 10x MOPS buffer, 1.6 µl 38% formaldehyde and 1.4 µl DDW. The samples were heated to 65°C for 5 min and cooled on ice. 1 µl loading dye (50% glycerol, 0.1% bromophenol blue) was added and the samples were loaded on the gel. Electrophoresis was carried out at 30-40 mA overnight.

2.12. Preparation of DNA and RNA blots
Nucleic acids were routinely transferred to nylon membranes (Hybond-N, Amersham) using blotting apparatus and protocols described in Sambrook et al. (1989).

2.12.1. Preparation of DNA blots
Following electrophoresis, gels which had not been run in the presence of ethidium bromide (e.g. pulsed field gels) were stained by immersing in running buffer containing 0.5 µg/ml ethidium bromide for 1 hr. The gel was destained by immersing in running buffer or DDW without
ethidium bromide for 30-60min. A photograph was then taken before the gel was soaked 0.15M HCl for 2x 10min to partially depurinate the DNA. (This treatment resulted in more efficient transfer of high molecular weight DNA molecules to the filter, because the depurinated sites are cleaved during the denaturation step, thus fragmenting the long molecules.) Following a rinse in DDW, the gel was soaked in denaturing solution for 2x 20min and then in neutralising solution for 2x 20min. The gel was then assembled in a capillary blotting apparatus containing 20x SSC and transfer of nucleic acid to the membrane was allowed to proceed overnight (genomic Southern) or for 48hr (PFGE gels). When blots of minigels containing cosmid, plasmid or phage DNA were being prepared, the acid treatment step was omitted and each denaturation or neutralisation step was reduced to 10min. Up to four blots were made from a single minigel by allowing transfer to proceed for 15min each time before changing the filter.

After transfer, filters were rinsed by immersing in 2x SSC and air dried. The DNA was then fixed to the membrane by baking for 2hr at 80°C or by UV cross linking at 0.4 J/cm².

Denaturing solution
0.5M NaOH
1.5M NaCl

Neutralising solution
0.5M Tris.Cl pH7.5
1.5M NaCl
1mM EDTA

2.12.2. Preparation of RNA blots
Following electrophoresis, northern gels were stained for 30min in 1µg/ml ethidium bromide and destained in water for 1hr before photography. Gels were then rinsed in 20x SSC and then blotted onto Hybond-N overnight in 20x SSC. After RNA transfer, filters were baked for 2hr at 80°C in a vacuum oven and UV cross linked at 0.4J/cm².

RNA loading was judged by hybridising northern blots with the probe pED1, a cDNA subclone of the esterase-D gene ESD, which is ubiquitously expressed in a cell-cycle independent fashion (Squire et al., 1986).
2.13. Preparation of DNA probes

Probe DNA fragments were purified using Geneclean (Bio101) from normal agarose gels run in 1x TAE buffer, or excised directly from low melting point agarose (LMPA) gels run in 0.5x TBE buffer.

Geneclean-purified DNA was labelled to high specific activity with [α-32P]dCTP by random hexamer priming (Feinberg and Vogelstein, 1983). For Genecleaned fragments, DDW was added to 10-50ng probe DNA to give a final volume of 33μl and boiled for 3 mins. The mixture was briefly cooled on ice before the addition of 10μl OLB, 2μl 10mg/ml BSA, 2.5μl [α-32P]dCTP (10μCi/μl) and 2.5μl Klenow fragment (8U/μl, BRL).

For probes isolated on LMPA gels, the required fragment was cut out with a scalpel over a UV transilluminator and the resulting gel slice boiled with 3 volumes DDW for 7min. After cooling to 37°C, aliquots containing 50ng DNA were transferred to fresh Eppendorfs and the volume adjusted to 33μl with DDW. One aliquot was then used for the labelling reaction as described above (Feinberg and Vogelstein, 1984).

The reaction was typically incubated for 6hr at room temperature. After this time, probe DNA was separated from unincorporated nucleotides by passing the reaction mix through a TES-equilibrated G-50 Sephadex column prepared in a Pasteur pipette plugged with glass wool. The first radioactive peak was collected and assayed. Labelled probes were boiled for 5min before addition to hybridisation bags (section 2.14.).

Probes containing repetitive DNA sequences were competed before use by adding 100μg sonicated human carrier DNA per 50ng labelled probe in a final concentration of 5x SSC, boiling for 5min, and incubating at 65°C for 100min before addition to the hybridisation bag.

<table>
<thead>
<tr>
<th>OLB</th>
<th>100μl Solution A</th>
</tr>
</thead>
<tbody>
<tr>
<td>250μl</td>
<td>Solution B</td>
</tr>
<tr>
<td>150μl</td>
<td>Solution C</td>
</tr>
</tbody>
</table>
Solution A
1.25M Tris.Cl pH7.5
0.125M MgCl₂
18μl β-mercaptoethanol
5μl each 0.1M dATP, dTTP, dGTP (Pharmacia)
(Solutions of nucleotides were prepared in
3mM Tris.Cl pH7.5, 0.2mM EDTA.)

Solution B
2M HEPES adjusted to pH7.4 with NaOH

Solution C
Hexadeoxyribonucleotides (Pharmacia)
suspended at 90 OD₂₆₀ units/ml in 10mM
Tris.Cl, 1mM EDTA.

TES
10mM Tris.Cl pH8.0
10mM EDTA
0.5% SDS

2.14. Hybridisation of blots

2.14.1. Hybridisation of DNA blots
Genomic blots were prehybridised at 65°C for 5-8hr in 6x SSC, 10%
dextran sulphate, 5x Denhardt's solution, 0.5% SDS and 100μg/ml heat
denatured sonicated salmon sperm DNA. Hybridisation was carried
out at 65°C overnight in fresh buffer with the addition of 10⁶ cpm probe
per ml (Sambrook et al., 1989).

Cosmid blot filters and recombinant DNA library filters were
hybridised in the same way except that the probe concentration was
reduced to 5x10⁵ cpm/ml.

In hybridisations using a human genomic or cDNA probe on human
DNA, filters were washed for 20min at 65°C in 2x SSC, 0.5% SDS
followed by 20min at 65°C in 0.1x SSC, 0.5% SDS. These washing
conditions were used unless otherwise stated. When a cross-species
hybridisation was being performed, lower washing stringencies were
used as described in the appropriate places in the text. A typical final
wash, determined empirically, was 6x SSC at 65°C. Filters were exposed
to Kodak XAR-5 autoradiography film at -80°C between intensifying screens.

100x Denhardt's solution
2% (w/v) BSA
2% (w/v) Ficoll
2% (w/v) Polyvinylpyrrolidine

2.14.2. Hybridisation of RNA blots
Northern blots were prehybridised at 42°C for 1hr in a buffer containing 50% deionised formamide, 6x SSPE, 5x Denhardt's solution and 100μg/ml sheared salmon sperm DNA. Hybridisation was carried out at 42°C overnight in fresh buffer with the addition of 10^6 cpm probe per ml.

When a human cDNA probe was hybridised to a human northern blot, filters were washed at 65°C for 20min in 2x SSPE, 0.5% SDS followed by 20min at 65°C in 0.1x SSPE, 0.5% SDS. When a human genomic fragment containing an unknown amount of exon was used as the probe, the washing stringency was reduced. A typical final wash, determined empirically, was 2x SSPE at 50°C.

20x SSPE
3.6M NaCl
0.2M sodium phosphate pH7.7
0.02M EDTA

2.15. DNA sequencing
DNA was sequenced by the chain-termination method (Sanger et al., 1977) using Sequenase protocols and reagents (USB). The Sequenase enzyme used was a modified bacteriophage T7 DNA polymerase. Sequencing was performed on double stranded DNA obtained from maxipreps or minipreps. DNA prepared by alkaline lysis minipreps was found to be clean enough for sequencing, but the RNA had to be removed first in order to accurately determine the concentration of DNA present. Oligonucleotide primers other than those for M13-based vectors were made in the Oligonucleotide Synthesis Laboratory, ICRF.
2.15.1. Sequencing reaction

2µg template DNA, 2µl 20mM EDTA and 2µl 2M NaOH were mixed in a microfuge tube with DDW to bring the final volume to 20µl. The mixture was incubated for 5min at room temperature before neutralising with 3µl 3M NaAc pH5.2 and 7µl DDW. 75µl absolute ethanol were then added and the DNA precipitated at -70°C for 15min. The DNA was pelleted by microfuging at 4°C for 10min, washed in 100µl 70% ethanol, respun, and the supernatant removed by aspiration with a drawn-out Pasteur pipette. The pellet was briefly dried under vacuum and resuspended in 7µl DDW and 2µl 5x reaction buffer. 0.5pmol primer were added to the resuspended denatured template and the two were annealed by incubating the mixture at 65°C for 2min and then cooling slowly to <35°C. The primer was extended by adding to the template/primer mix (10µl): 1µl 0.1M DTT, 2µl 1x labelling mix, 0.5µl [α-35S]dATP and 2µl Sequenase enzyme diluted 1:7 in TE pH8.0. The reaction was incubated at room temperature for 5min. 3.5µl were then transferred to four prewarmed tubes, each containing 2.5µl of one of the termination mixes. The reactions were incubated at 37°C for 5min, during which time dideoxynucleotides were incorporated into the extended DNA molecules. 4µl stop solution were then added to each tube. The samples were denatured at 75°C for 2min before resolving 2.5µl of each on a sequencing gel. The remainder of each sample was stored at -20°C.

To resolve compressions sometimes observed on the gel due to secondary structures in regions rich in dG and dC, the sequencing reaction was performed with the substitution of dITP for dGTP (and ddITP for ddGTP). dITP forms weaker secondary structures than dGTP and these are more readily denatured during electrophoresis, resulting in improved resolution in the compressed areas.

5x Reaction buffer

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>200mM Tris.Cl</td>
<td>pH7.5</td>
</tr>
<tr>
<td>100mM MgCl₂</td>
<td></td>
</tr>
<tr>
<td>250mM NaCl</td>
<td></td>
</tr>
</tbody>
</table>

72
5x Labelling mix 7.5μM each of dGTP, dCTP and dTTP
Diluted to 1x with DDW.

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

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

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

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

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

2.15.2. Denaturing polyacrylamide gel electrophoresis
Sequencing gel electrophoresis was performed using Koch-Light apparatus and gel plates. Before each run, the gel plates (40x20cm) were washed in warm soapy water, rinsed and dried. They were cleaned by wiping first with ethanol, then acetone, and finally DDW. The back plate was siliconised with Repelcote. 0.4mm spacers were positioned between the plates and the sandwich fastened together with bulldog clips. 50ml polyacrylamide gel mix were prepared and injected between the plates with a syringe. The comb was positioned and the gel was allowed to set. The gel was then assembled in the tank and warmed by electrophoresing in 1x STBE for at least 20min at 4kW. Excess urea was syringed from the loading area and 2.5μl of each sample were then loaded. A run time of about 100min resolved sequences near the
primer. Longer runs of up to 4hr allowed sequences over 200 nucleotides from the primer to be read.

Following the run, the back plate was carefully lifted off and the gel (on the front plate) was fixed in a tray containing 10% glacial acetic acid, 10% methanol for 15min. The gel was then drained, transferred to a sheet of Whatman 3mm paper, covered in Saran wrap and dried at 80°C for 30min on a Bio-Rad slab drier. Autoradiography was performed overnight using Kodak XAR-5 film at room temperature. Sequences were compiled using Intelligenetics GEL and SEQ software. Comparisons of nucleotide and predicted amino acid sequences to the databases were performed using the Intelligenetics IFIND program.

Gel mix
31.5ml 50% urea
6.7ml 40% acrylamide stock
5ml 10x STBE
DDW to 50ml
400μl 100mg/ml ammonium persulphate
60μl TEMED

40% acrylamide stock
38% (w/v) acrylamide
2% N,N’-bismethylene-acrylamide
Deionised, filtered and stored at 4°C.

10x STBE
Tris base 108g/litre
Boric acid 55g/litre
EDTA 9.3g/litre
3. Identification of potential CpG islands in the class II region by PFGE mapping

3.1. Strategy to identify potential sites of genes

The approach taken to identify the position of novel genes in the class II region was based on the observation that the 5' ends of genes are frequently associated with short stretches of CpG-rich DNA, known as CpG islands (Bird, 1987).

Vertebrate DNA is relatively depleted of C+G, and the dinucleotide CpG is particularly rare, being present at only 20% of the expected frequency. The majority of CpG dinucleotides are stably methylated at the 5 position on the cytosine ring (Bird, 1986). It has been observed, however, that about 1% of the genomic DNA in a wide variety of vertebrate species is non-methylated, as judged by cleavage of genomic DNA with the methylation sensitive restriction endonuclease HpaII (which cuts at CCGG but not CmCGG) (Cooper et al., 1983). Cloning and sequencing of a number of the HpaII fragments revealed that they were rich in C+G and that CpG was present at the expected frequency (Bird et al., 1985). Randomly selected clones were found to detect transcripts when used as probes on northern blots, and characterisation of one clone in detail revealed that the CpG rich region was located at the 5' end of two opposite strand divergent transcripts (Lavia et al., 1987). It had also been independently shown that non-methylated sequences characteristic of CpG islands were associated with the 5' ends of a number of vertebrate genes, such as the chicken α2(I) collagen gene (McKeon et al., 1982). A subsequent search of the gene databases revealed that all constitutively expressed 'housekeeping' genes and many tissue specific genes had CpG-rich islands at the 5' end (Bird, 1986; Gardiner-Garden and Frommer, 1987).

The unusual nucleotide composition of CpG islands led to the prediction that islands would be detectable in genomic DNA as containing clusters of sites for the infrequently cutting restriction endonucleases
used for constructing PFGE maps (Brown and Bird, 1986; Lindsay and Bird, 1987). These 'rare-cutter' enzymes have 6 or 8bp recognition sequences containing one or more CpG, which must be unmethylated for cleavage to take place. CpG islands are therefore likely to contain sites for these enzymes. Indeed, if it is assumed that CpG islands contain 65% G+C, are not depleted in CpG and are 1kb long, whereas bulk DNA is 40% G+C with CpG occurring at 25% of the expected level, then the theoretical number of sites for each enzyme in the human genome (3x10⁹bp) and the proportion of these sites in islands can be calculated (Table 3.1).

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Target sequence</th>
<th>Total sites in genome</th>
<th>% sites in islands</th>
<th>Sites per island Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>NotI [a]</td>
<td>GCGGCCGC</td>
<td>4100</td>
<td>89</td>
<td>0.12</td>
<td>0.3</td>
</tr>
<tr>
<td>BssHII [a]</td>
<td>GCGCGC</td>
<td>47000</td>
<td>74</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>EagI [a]</td>
<td>CGGCCG</td>
<td>47000</td>
<td>74</td>
<td>1.2</td>
<td>1.1</td>
</tr>
<tr>
<td>MluI [b]</td>
<td>ACGCGT</td>
<td>37000</td>
<td>21</td>
<td>0.3</td>
<td>0.03</td>
</tr>
<tr>
<td>NruI [b]</td>
<td>TCGCGA</td>
<td>37000</td>
<td>21</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 3.1. Distribution of selected rare-cutter sites in human DNA. For each enzyme (all of which contain two CpGs, underlined) the total number of sites, the percentage of sites in islands and the expected number of sites per island were calculated by Lindsay and Bird (1987). The observed number of sites for each enzyme was obtained from the nucleotide sequences of the CpG islands at a random selection of human genes in the databases (Bird, 1990). Class designation ([a] or [b]) is from Bird (1990).

The table shows that those rare-cutter enzymes with recognition sequences consisting entirely of C+G and containing two CpG dinucleotides are predicted to be highly diagnostic for CpG islands because the majority of sites for these enzymes should occur in islands (class [a] enzymes; Bird, 1990). Most importantly, it has been demonstrated in practice that sites for class [a] enzymes are indeed clustered in CpG islands.
islands and are associated with transcribed sequences (Brown and Bird, 1986; Lindsay and Bird, 1987; Bird 1990). The mapping and cloning of rare-cutter sites in a genomic region of interest has successfully led to the identification of novel genes in mouse and man (Rappold et al., 1987; Estivill et al., 1987). In contrast, rare-cutter enzymes with recognition sequences containing A+T in addition to C+G (class [b] enzymes, Table 3.1.), are predicted in theory and shown in practice to cut more frequently in inter-island DNA (Lindsay and Bird, 1987; Bird, 1990).

The feasibility of using PFGE to map the class II region has been demonstrated by Hardy et al. (1986) and more recently by Dunham et al. (1989) and Inoko et al. (1989). The aim of these studies was to use PFGE to construct detailed physical maps of the region from which the relative order of the different subregions and the distances between them could be established. In the present study it was decided to map the class II region by PFGE with the specific aim of locating class [a] rare-cutter sites which could mark the presence of genes. These sites could then be cloned and analysed in detail for evidence of transcribed sequences.

3.2. Choice of materials for construction of the PFGE map

3.2.1. Restriction endonucleases
A discussed above, class [a] rare-cutter enzymes are predicted from theory and shown in practice to cut far more frequently in CpG islands than in inter-island DNA (Table 3.1.) These were selected as the enzymes of choice for finding CpG islands. However, it was also necessary to use class [b] rare-cutters, which tend to cut between CpG islands, so that complementary, overlapping fragments were generated which could be used to link together the class [a] fragments and build up the map. The class [a] enzymes used in this study were BssHII, EagI and NotI. The class [b] enzymes used were MluI and NruI.

3.2.2. DNA
Most individuals in the population possess two distinct class II haplotypes, the long range maps of which may differ due to polymorphisms at the rare-cutter sites, through variation in DRB gene number, or
through differences in the amount of DNA between subregions (Dunham et al., 1989; Lawrance and Smith, 1990). To avoid possible difficulties in interpretation of data which could arise from using the DNA of an MHC heterozygote, PFGE blots were prepared from the DNA of the EBV-transformed B-lymphoblastoid cell line PGF, which is homozygous for the entire MHC as judged by tissue typing. The haplotype of this cell line is A3 B7 DRw15(2) Dw2 DQw6 DPw4.

3.2.3. PFGE apparatus
PGF DNA digested with rare-cutter enzymes was resolved by pulsed field gradient gel electrophoresis using LKB Pulsaphor apparatus with the electrodes in the double inhomogeneous field configuration (Figure 3.1.).

Figure 3.1. Schematic illustration of a double inhomogeneous pulsed field gel electrophoresis apparatus. Electrodes are represented by black dots. Arrows indicate the direction of motion during electrophoresis of a DNA molecule loaded in the central well under the influence of the electric field which pulses first in the north-south direction and then in the west-east direction. Fractionation of molecules is believed to be dependent on the greater ability of small molecules to re-orientate themselves when the field direction changes (Smith et al., 1987).
A constant pulse time of 45-60sec and a run time of 20-24hr was found to resolve molecules of up to 900kb. One disadvantage of this gel system is that electrophoresis results in tracks which are bowed rather than straight, which can complicate comparison of fragment sizes between the lanes. However, in practice it was found that only a small amount of distortion occurred during short runs (Figure 3.2.a.). Hybridisation of blots generated with this system resulted in sharp bands (Figure 3.2.b.) Blots generated using the LKB hexagonal electrode array, which resolves DNA in straight tracks, were found to give less well defined bands (data not shown).

3.2.4. Probes
The probes for the class II DPB1, DPA1, DOB, DQA2, DRB1 and DRA genes were those described in the report of the Tenth International Histocompatibility Workshop (Marcadet et al., 1989). The probe for the DNA gene was 8bα1, a 1.8kb PstI genomic fragment isolated from the cosmid JG8b (Trowsdale and Kelly, 1985).

3.3. Construction of the PFGE map

DNA from the cell line PGF was digested with rare-cutter enzymes, resolved by PFGE, transferred to nylon membranes and hybridised with probes for class II genes. The sizes of the fragments detected by each probe with each enzyme or combination of enzymes are shown in Table 3.2. Representative autoradiographs are shown in Figures 3.2.b., 3.3. and 3.4.
Figure 3.2.

(a) Ethidium bromide-stained gel containing PGF DNA digested with the enzymes shown and resolved by pulsed-field electrophoresis. Size markers are yeast chromosomes (YC). LM, limiting mobility.

(b) Autoradiographs of the blot of the same gel hybridised sequentially with the class II gene probes indicated.
Figure 3.3. Autoradiographs of a PFGE filter sequentially hybridised with probes for the class II genes DPB1, DNA and DOB. N, NotI; B, BssHII; M, MluI. The MluI+NotI double digest is partial. Size markers are concatemers of bacteriophage lambda DNA. LM, Limiting mobility.
Figure 3.4.

(a) Autoradiographs of a PFGE filter hybridised sequentially with DPB1 and DOB probes, illustrating the coincidence of BssHII and EagI sites within the class II region. B, BssHII; E, EagI.

(b) Autoradiographs of a PFGE filter hybridised sequentially with DOB and DQA2 probes, illustrating the strong and weak bands obtained with the DQA2 probe as discussed in the text. B, BssHII; E, EagI; M, MluI. Size markers are yeast chromosomes in each case.
Table 3.2. Sizes of rare-cutter fragments detected by MHC class II probes on PFGE blots. Fragment sizes are in kilobases. Probes are shown across the top and rare-cutter enzymes are shown down the side (B, BssHII; E, EagI; M, MluI; N, NotI). The DQA2 probe hybridised to two bands in each track: one strong (S) corresponding to the fragment carrying the DQA2 gene, and the other weaker (W), corresponding to the fragment carrying the cross-hybridising DQA1 gene (see text for further discussion). NT, not tested. Fragment sizes were deduced from more than one blot.

The genes of the DP subregion are known from family studies to be at the centromeric end of the class II region (Shaw et al., 1981). From cosmid walking it has been established that the configuration of the genes within the subregion is DPB2-DPA2-DPB1-DPA1 (Trowsdale et al., 1984). It has also been shown by deletion mapping that the cluster is oriented on the short arm of chromosome 6 with DPB2 toward the
centromere (Erlich et al., 1986). This information was used as a starting point for the construction of the PFGE map. Since the \textit{DPB1} and \textit{DPA1} genes are only 2kb apart (Kelly and Trowsdale, 1985), it was not surprising that the \textit{DPB1} and \textit{DPA1} probes detected similarly sized NotI, BssHII and EagI fragments (Table 3.2.). With MluI, however, the \textit{DPB1} probe detected a 130kb fragment while the \textit{DPA1} probe detected a 450kb fragment (Figure 3.2.b.). This result indicated that that there must be an MluI site in the genomic DNA between the \textit{DPA1} and \textit{DPB1} probes. These data allowed the immediate positioning of three MluI sites: the first between the \textit{DPA1} and \textit{DPB1} probes, the second 130kb proximal and the third 450kb distal (Figure 3.5.). The 250kb BssHII fragment detected by both \textit{DPB1} and \textit{DPA1} probes was cut by MluI to give a 130kb fragment detected by \textit{DPB1} and a 120kb fragment detected by \textit{DPA1} (Table 3.2.). The 370kb NotI fragment detected by both \textit{DPB1} and \textit{DPA1} probes was also cut down by MluI to give a 130kb fragment detected by \textit{DPB1} and a 240kb fragment detected by \textit{DPA1}. With additional data from EagI single and double digests (Table 3.2. and Figure 3.4.a.) it was possible to position BssHII, EagI and NotI sites on either side of the HLA-DP subregion as shown in Figure 3.5. It was already apparent that clusters of class [a] rare-cutter enzyme sites were present in the class II region. Cluster 1 contained sites for BssHII, EagI and NotI while cluster 2 contained sites for BssHII and EagI.

The \textit{DNA} gene probe hybridised to similarly sized fragments as the \textit{DPA1} probe in all single and double digest combinations tested. (Figure 3.3. and Table 3.2.). It was concluded that the \textit{DNA} gene mapped distal of the \textit{DPA1} gene, but within about 120kb as defined by the position of the distal BssHII site.

The \textit{DOB} gene probe detected a 450kb MluI fragment, like the \textit{DNA} and \textit{DPA1} probes (Figures 3.2. and 3.3.). However, the \textit{DOB} probe detected different NotI (>900kb), EagI (210kb) and BssHII (210kb) fragments, indicating that it must map distal of the \textit{DNA} gene but still within the same 450kb MluI fragment. The double digest data obtained with the \textit{DOB} probe allowed the positioning of EagI and BssHII sites coincident with the central NotI site. This defined a third cluster of class [a] rare-cutter sites (Cluster 3, Figure 3.5.)
The \textit{DQA2} probe hybridised to two fragments in each track, one strongly, corresponding to the fragment carrying the \textit{DQA2} gene, and one weakly, corresponding to the fragment carrying the \textit{DQA1} gene (Figure 3.4.b.). This pattern of hybridisation has also been reported in other studies with the same probe (Biro et al., 1989; Bontrop et al., 1989). With BssHII, EagI and MluI, the strongly hybridising band (\textit{DQA2}) was of a similar size to that detected with the \textit{DOB} probe, while the weaker band (\textit{DQA1}) was novel (Figure 3.4.b.). It was concluded that there was a fourth cluster of rare-cutter sites between the \textit{DQA2} and \textit{DQA1} genes (Cluster 4, Figure 3.5.). Because the \textit{DQA2} and \textit{DOB} probes detected similarly sized fragments in each single and double digest tested, it was not possible to determine the positions of these genes relative to one another. However, the relative positions of the \textit{DOB} and \textit{DQA2} genes has been established as cen-\textit{DOB}-\textit{DQA2}-tel for a number of different cell lines, including DR2 homozygotes, in other studies (Hardy et al., 1986; Dunham et al., 1989; Inoko et al., 1989; Blanck and Strominger, 1988). Probes for the \textit{DQB1}, \textit{DQB2} and \textit{DQB3} (formerly \textit{DVB}) genes were not used in the present study, since their close proximity to the \textit{DQA1} and \textit{DQB2} genes is well documented (Okada et al., 1985b; Jonsson et al., 1987; Blanck and Strominger, 1988; Ando et al., 1989; Inoko et al., 1989). The overall order of the genes in the DO and DQ subregions has been established as cen-\textit{DOB}-\textit{DQB2}-\textit{DQA2}-\textit{DQB3}-\textit{DQB1}-\textit{DQA1}-tel by pulsed field mapping in a DR2 homozygous cell line and by chromosome walking in a DR7 homozygous cell line (Inoko et al., 1989; Blanck and Strominger, 1988). This information was used in the construction of the map shown in Figure 3.5.

The \textit{DRB1} and \textit{DRA} gene probes both hybridised to similarly sized fragments as the weakly hybridising (\textit{DQA1}) bands detected by the \textit{DQA2} probe (Table 3.2.). It was not possible from these data to determine the relative positions of the \textit{DQA1}, \textit{DRB1} and \textit{DRA} genes, but it was clear that there were no further clusters of rare-cutter sites in the class II region. For the construction of the map shown in Figure 3.5., the positions of the genes in this region were assumed to be as previously determined for a DR2 homozygous cell line: cen-\textit{DQA1-DRB1-DRB2-DRB3-DRA}-tel (Inoko et al., 1989; Kawai et al., 1989).
Figure 3.5.
Physical map of the class II region of the cell line PGF compiled from PFGE mapping data and previously published results concerning the positions of genes within individual subregions, as discussed in the text. The positions of the four clusters of rare-cutter sites deduced from the PFGE data are shown. Apart from the genes in the DP subregion, the position indicated for each gene is approximate (see text). The extent of overlapping cosmid clones in the DP subregion, which could be used as a starting point for the cloning of adjacent clusters, is indicated.
The PFGE data were used to construct a physical map of the class II region of the PGF haplotype (Figure 3.5.) The novel feature of interest of this map is the presence of four clusters, as defined at the level of resolution of the PFGE technique, of class [a] rare-cutter enzyme sites, which as discussed previously are highly diagnostic for CpG islands. Each cluster contains one or more sites for EagI, BssHII and NotI. These clusters were considered likely to be the sites of novel genes and were selected as targets for cloning and further analysis.

3.4. Summary and discussion

The physical mapping data presented here reveal that the organisation of the class II region of the DR2-homozygous cell line PGF is very similar to that of other cell lines. The class II gene order deduced for PGF (centromere-DBP1-DPA1-DNA-[DOB, DQA2]-[DQA1, DRB, DRA]-telomere) is consistent with that reported for other cell lines in other studies. In another study in which DR2 haplotypes were mapped with some of the same enzymes (MluI, BssHII and NotI), the fragment sizes detected were in good agreement, within experimental error (Dunham et al., 1989). The fragment sizes detected in PGF DNA with DRA and DRB gene probes are consistent with the conclusion of Dunham et al. (1989) that DR2 haplotypes possess more DNA in the DR subregion than DR3, DR5 and DR6 haplotypes, but less than DR4 haplotypes.

The PFGE map of the PGF class II region has revealed the presence of four clusters of sites for rare-cutter enzymes which are highly diagnostic for CpG islands and which could mark the positions of novel genes (Figure 3.5.). The next step towards testing this hypothesis was to clone the regions containing the clusters of rare-cutter sites by walking from previously cloned class II genes. In general, the precise distance from each class II gene relative to the nearest rare-cutter sites could not be determined from the map. However, the DP subregion is anchored to the physical map because an MluI site occurs between the probes for the DPB1 and DPA1 genes, which map only 2kb apart (Kelly and Trowsdale, 1985). Cluster 1 maps 130kb centromeric of this site and cluster 2 maps 120kb telomeric. The DP subregion has been cloned on
overlapping cosmids, which extend 70kb centromeric and 20kb telo-
meric of the MluI site as shown in Figure 3.5. (Trowsdale et al., 1985).
This previously established cosmid walk was potentially a useful
starting point for chromosome walking towards clusters 1 and 2, since
cluster 1 was estimated to be about 60kb from the proximal end of the
DP region cosmid clones, while cluster 2 was about 100kb from the
distal end of the cloned region. In fact, the walk towards cluster 1 was
accelerated by the serendipitous discovery that the gene for the α2
chain of type XI collagen mapped between cluster 1 and the HLA-DP
subregion, as described in Chapter 4.
4. Mapping of the COL11A2 gene to the class II region

This chapter describes the mapping of the gene for the α2 chain of type XI collagen, a component of cartilage fibrils, to the centromeric end of the class II region of the MHC.

4.1. Introduction

The mechanical strength of cartilage is provided by a three-dimensional mesh of rigid, rod-like collagen fibres composed of type II, type IX and type XI collagen molecules (Mendler et al., 1989). Type II and type XI collagens are of the fibrillar class, containing a single uninterrupted triple helical domain as shown in Figure 4.1. (Miller and Gay, 1987). In cartilage collagen fibres, type II and type XI molecules are packed together in a ratio of about 8:1 in a supramolecular staggered array (Mendler et al., 1989).

Figure 4.1. Schematic diagram of a fibrillar collagen molecule. The mature molecule is 300nm long and is composed of three polypeptide chains, α1, α2 and α3. In fibrillar collagens the majority of the length of the molecule is accounted for by the long, uninterrupted triple helical domain (Miller and Gay, 1987).

Type XI collagen is a heterotrimer of α1, α2 and α3 polypeptide chains (Morris and Bachinger, 1987) which are encoded by three separate
genes. The α3(XI) chain is believed to be a heavily glycosylated variant of the α1 chain of type II collagen and would therefore be a product of the COL2A1 gene at 12q14.3 (Furuto and Miller, 1983; Law et al., 1986). The α1(XI) chain gene (COL11A1) and the α2(XI) chain gene (COL11A2) have been cloned and sequenced more recently (Bernard et al., 1988; Kimura et al., 1989). COL11A1 has been mapped to human chromosome 1, region p21 (Henry et al., 1988). COL11A2 was first mapped in the Human Cytogenetics Laboratory, ICRF, to human chromosome 6, region 21.3-22, by in situ hybridisation (Figure 4.2.; Hanson et al., 1989). The MHC has previously been shown to map to 6p21.3 (Spring et al., 1985).

The possibility that COL11A2 might map near the MHC was considered extremely interesting because of the involvement of cartilage collagen in a clinically important MHC-associated disease, rheumatoid arthritis (RA; Lotz and Vaughan, 1988). About 75% of RA patients have antibodies autoreactive to type II collagen, and type II collagen-reactive T-lymphocytes are found at high frequency in the synovial infiltrates of chronically affected individuals (Londei et al., 1989). The importance of an autoimmune response to cartilage collagen in the development of the arthritic condition has been demonstrated in rodents and primates where the disease can be induced with injections of type II collagen (Stuart et al., 1984). The disease can also be transferred by injection of antibodies or T-cells from affected individuals. The arthritogenic potential of type XI collagen has not yet been investigated in such detail because type XI collagen was discovered much more recently. However, it has now been reported that inoculation of mice with type XI collagen also induces the development of arthritis (Boissier et al., 1990). In view of these results, it was decided to map the COL11A2 gene in detail relative to the MHC. If the gene was found to be closely linked it was hoped to perform RFLP studies on diseased and normal individuals to test whether certain COL11A2 alleles were associated with the incidence of RA.
Figure 4.2.
Histogram showing distribution of silver grains over metaphase human chromosomes following in situ hybridisation with the COL11A2 probe. Analysis of 96 metaphase chromosome spreads revealed a clustering of grains over bands 6p21.3-22 (Hanson et al., 1989).
4.2. Mapping of COL11A2 using somatic cell hybrids

To confirm and refine the COL11A2 map position, it was decided to test for the presence of the gene in the DNA of a pair of complementary human/mouse somatic cell hybrids, MCP-6 and 56-47, which contain portions of chromosome 6 overlapping at band 6p21 as shown in Figure 4.3.a. MCP-6 contains the translocation chromosome t(6:X) (6qter-6p21::Xq13-Xqter) (Goodfellow et al., 1982) and 56-47 contains the translocation chromosome t(6:17)(6pter-6p21::17p13-17qter) along with other human chromosomes (Nagarajan et al., 1986). A Southern blot was prepared from EcoRI-digested DNA from the hybrids, and from total human and mouse genomic DNA as controls. Hybridisation of this blot with the COL11A2 probe, a 4.5kb EcoRI/BamHI fragment from the cosm id cosHcol.11 (Hanson et al., 1989), gave the result shown in Figure 4.3.b. The probe detected a fragment of about 11kb in total human, MCP-6 and 56-47 DNAs but not in mouse genomic DNA at the stringency used for washing (0.1x SSC, 65°C). It was therefore concluded that the COL11A2 gene must map to 6p21.1-3, the region shared by the hybrids (Figure 4.4.a).

Taken together, the results from in situ hybridisation (6p21.3-22, Figure 4.2.) and somatic cell hybrid mapping (6p21.1-3, Figure 4.3.) suggest that the most likely location of COL11A2 was in band 6p21.3, the map position of the MHC. It was therefore decided to attempt to link the COL11A2 locus with the MHC by PFGE mapping.

4.3. Mapping of COL11A2 by PFGE

When the COL11A2 gene probe was hybridised to PFGE blots it was found that it detected fragments similar in size to those detected by the DPB1 gene probe in all single and double digest combinations tested (Figure 4.4. and Table 4.1.).
Figure 4.3.
(a) Intact human chromosome 6 (showing the position of the MHC), and regions of chromosome 6 contained in the human/mouse hybrids 56-47 and MCP-6.
(b) Autoradiograph of Southern blot of EcoRI-digested DNA hybridised with the COL11A2 probe and washed to a final stringency of 0.1xSSC, 65°C. Lane 1, total human DNA; lane 2, total mouse DNA; lane 3, MCP-6; lane 4, 56-47.
Figure 4.4.

Autoradiographs of a PFGE filter hybridised sequentially with probes for the DPB1 and COL11A2 genes. The filter was stripped and checked for retention of signal by autoradiography between hybridisations. B, BssHII; M, MluI; N, NotI.
Table 4.1. Sizes of fragments detected by hybridising probes for the COL11A2 and DPB1 genes to PFGE blots of DNA from the cell line PGF cut with the enzymes indicated. (Figure 4.4. and Hanson et al., 1989).

It was concluded that the COL11A2 and DPB1 genes were in close physical proximity at the centromeric end of the class II region. The upper limit on the maximum distance between the two probes is given by the length of the shortest common band detected, a 130kb MluI fragment (Figure 4.4.). As shown in Chapter 3, the telomeric end of this 130kb MluI fragment falls between the DPA1 and DPB1 genes, while the centromeric end occurs within cluster 1 (Figure 3.5.).

4.4. Cosmid walking between the DP subregion and COL11A2.

The genes of the DP subregion have previously been isolated on overlapping cosmid clones (MANN 2.3, MANN 3.6 and MANN 2.2, Figure 4.5.; Trowsdale et al., 1985). These extend approximately 70kb distal of the MluI site between DPA1 and DPB1. The COL11A2 probe did not hybridise to blots of DNA from these cosmids, however (data not shown). The COL11A2 gene was therefore deduced to map centromeric of the previously cloned region but still within the same 130kb MluI fragment, as shown in Figure 4.5.
Figure 4.5.
Mapping of COL11A2 relative to the DP subregion. The positions of the genomic MluI sites (M), as deduced from pulsed field mapping, are indicated at the top. Below are shown restriction maps of overlapping cosmid clones used to accurately position the COL11A2 gene. Cosmids MANN 2.3, 2.2 and 3.6 have been previously described (Trowsdale et al., 1985). The position of the probe from the proximal end of MANN 2.3 used to screen the cosmid library is shown (WP). Arrows indicate the transcriptional orientation of each gene.
The DP-region cosm id walk was extended in the centromeric direction to confirm this and in so doing to link the \textit{COL11A2} locus to the DP genes.

A 4.5kb Clal/SalI fragment was subcloned into the Bluescript vector from the end of MANN 2.3, the most centromeric cosm id in the cloned region (Figure 4.5.). The Clal site was in the genomic insert of MANN 2.3 and the SalI site was at the cloning site of the cosm id vector, pTCF. The subcloned fragment, pCS2, was found to contain many repetitive sequences when used as a probe on Southern blots of human genomic DNA. pCS2 was therefore cleaved with Rsal to generate several smaller fragments, each of which was tested for the presence of repeats by hybridising back to genomic Southern blots. One 500bp Rsal fragment (WP in Figure 4.5.) was found to be single copy and was used to screen the cosm id library. Two new cosm ids, HPB.ALL 1 and HPB.ALL 8, were isolated with this probe and were shown to span the gap between MANN 2.3 and cosHcol.11, the cosm id clone carrying the \textit{COL11A2} gene (Figure 4.5.) Since the position and orientation of the \textit{COL11A2} gene within cosHcol.11 was already known (Dr. Kathy Cheah, personal communication), this result established that the \textit{COL11A2} gene was 45kb centromeric to \textit{DPB2} and oriented with the 3' end of the gene nearest the MHC.

4.5. Summary and discussion

Using mapping techniques of increasing resolution, the gene for the \(\alpha_2\) chain of type XI collagen has been localised to the centromeric end of the class II region of the MHC. This is consistent with two other reports localising \textit{COL11A2} to the short arm of chromosome 6, although neither of these studies mapped the gene relative to the MHC (Kimura et al., 1989; Law et al., 1990). This assignment is in keeping with the general trend for diverse map locations of human fibrillar collagen genes. The \textit{COL3A1} and \textit{COL5A2} loci are both found within the chromosomal region 2q24.3-31, but \textit{COL1A1}, \textit{COL1A2}, \textit{COL2A1}, \textit{COL11A1} and \textit{COL11A2} map to 17q21-22, 7q21.3-22.1, 12q14.3, 1p21 and 6p21.3 respectively (Vuorio and de Crombrugghe, 1990).
The finding that the COL11A2 locus maps only 45kb proximal to the DPB2 gene is intriguing, given the association of RA with the class II region, and the involvement of cartilage collagens described in section 4.1. However, the major association of RA with the class II region is seen with the DR subregion (Nepom, 1990) and in studies where a possible association of RA with the DP subregion has been investigated, no statistically significant difference in the frequency of DP alleles in patients and controls has been observed (Begovich et al., 1989; Stephens et al., 1989). Thus it is unlikely that the close linkage of the COL11A2 gene to the class II region is significant for the understanding of the association of classical RA with the class II region. However, another MHC-associated disease, pauciarticular juvenile rheumatoid arthritis (PJRA), does demonstrate an association with DP alleles; specifically, the incidence of DPw2 is increased in patients compared to controls (Odum et al., 1986; Begovich et al., 1989; Fugger et al., 1990). Anti-collagen antibodies have been observed in PJRA patients (Lotz and Vaughan, 1988). It remains possible therefore that the association of PJRA with the DP subregion might be at least partly explained by linkage disequilibrium between alleles of the DP genes and alleles of the closely physically linked COL11A2 locus. Consequently it was decided to search for RFLPs at the COL11A2 locus which could be used to extend the disease association studies. This work is described in Chapter 8.

In the course of the work described in this chapter, overlapping cosmid clones were isolated which extended approximately 130kb proximal of the MluI site between DPA1 and DPB1. The genomic insert of cosHcol.11 was mapped and shown not to contain an MluI site. The proximal end of the 130kb MluI fragment (Figure 4.5.) must therefore be located just proximal of cosHcol.11. Since this MluI site was known to map within the most centromeric of the clusters of rare-cutter sites (Cluster 1, Figure 3.5.), the cosmid walk was extended further to encompass this region, as described in Chapter 5.
5. Identification of novel genes associated with clusters of rare-cutter sites

This chapter describes the cloning of three of the clusters of rare-cutter sites identified by PFGE mapping of the class II region (Chapter 3), and the subsequent identification of five novel genes.

5.1. Introduction

The work described in Chapter 4, i.e. the positioning of the COL11A2 gene centromeric of the DP subregion, provided a useful step towards the cloning of the first cluster of rare-cutter sites identified by PFGE (Cluster 1, Figure 3.5.). As shown in Figure 4.5., the centromeric end of the 130kb MluI fragment which carries the COL11A2 and DPB1 genes must fall just beyond the proximal end of cosmid cosHcol.11. Since this MluI site occurs in cluster 1, the cosmid walk was extended in the centromeric direction in order to encompass the rare-cutter sites of cluster 1.

5.2. Cloning of cluster 1 by cosmid walking

A 7.5kb EcoRI fragment (E3, Figure 5.2.) was isolated from the centromeric end of cosHcol.11 and used as a probe to screen the cosmid library. Four new overlapping cosmids, HPB.ALL 25, HPB.ALL 31, HPB.ALL 33 and HPB.ALL 42, were isolated and mapped with restriction enzymes, including the rare-cutters. A complete map of the cloned region is presented in Figure 5.1. and a more detailed restriction map of the new cosmids is shown in Figure 5.2. As discussed above, the PFGE data indicated that there should be an MluI site just centromeric of the cosmid cosHcol.11. This was indeed found to be the case. In addition to the single MluI site, this region was found to contain one BssHII site, five EaiI sites and one NotI site (Figure 5.2.). To prove that
Figure 5.1.
Restriction maps of overlapping cosmid clones extending from the HLA-DP genes into the region centromeric of the COL11A2 gene which is predicted from pulsed-field mapping to contain cluster 1. The position of the walking probe E3 from the proximal end of cosHcol.11 which was used to isolate the new cosmid clones is shown in Figure 5.2.
Figure 5.2.
Restriction maps of new overlapping cosmid clones, showing the cluster of sites for the rare-cutter enzymes BssHII, EagI, MluI and NotI. Brackets indicate a portion of cosmid HPB.ALL 33 that was found to be rearranged. Also shown are the positions of various restriction enzyme fragments which were isolated for use as probes.
these were the cluster 1 sites as detected by PFGE, it was necessary to demonstrate that they were unmethylated in genomic DNA.

The methylation status of the BssHII, MluI and NotI sites in genomic DNA were determined by hybridising PFGE blots with probes mapping distal (COL11A2) and proximal (33X1, Figure 5.2.) of these sites as deduced from the cosmid map. The results are shown in Figure 5.3. and summarised in Table 5.1.

<table>
<thead>
<tr>
<th>PROBE</th>
<th>DIGEST</th>
<th>33X1</th>
<th>COL11A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BssHII</td>
<td>100kb</td>
<td>200kb</td>
<td></td>
</tr>
<tr>
<td>MluI</td>
<td>230kb</td>
<td>130kb</td>
<td></td>
</tr>
<tr>
<td>NotI</td>
<td>440kb</td>
<td>370kb</td>
<td></td>
</tr>
<tr>
<td>BssHII+MluI</td>
<td>100kb</td>
<td>130kb</td>
<td></td>
</tr>
<tr>
<td>BssHII+NotI</td>
<td>100kb</td>
<td>250kb</td>
<td></td>
</tr>
<tr>
<td>MluI+NotI</td>
<td>230kb</td>
<td>130kb</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1. Sizes of fragments detected by probes distal (COL11A2) and proximal (33X1) of the BssHII, MluI and NotI sites in the new cosmid clones.

The two probes detected differently sized fragments with BssHII, MluI and NotI, showing that these sites were cut and therefore unmethylated in genomic DNA.

To determine the methylation status of the EagI sites, Southern blots were prepared from genomic (PGF) DNA and cosmid (HPB.ALL 42) DNA cut with EagI and resolved by conventional gel electrophoresis. These blots were then hybridised with fragments adjacent to or spanning the EagI sites as deduced from the cosmid map. One result is shown in Figure 5.4. In cosmid DNA, which is unmethylated and therefore cut at all EagI sites, the probe 33X3 detected two EagI fragments because it spans the central EagI site. In PGF DNA, however,
Figure 5.3.
Determination of the methylation status of cloned BssHII, MluI and NotI sites in genomic DNA. At the top is shown a map of the rare-cutter sites in the overlapping cosmid clones. The autoradiographs show a pulsed field blot hybridised sequentially with probes mapping proximal (33X1) and distal (COL11A2) of the BssHII, MluI and NotI sites in the cosmid clones.
Figure 5.4.

Determination of the methylation status of cloned EagI sites in genomic DNA. At the top is shown a summary map of the rare-cutter sites in the overlapping cosmid clones. The autoradiographs show blots of cosmid DNA (HPB.ALL 42) and genomic DNA (PGF) cut with EagI and hybridised with the genomic fragment 33X3. Bars represent lambda/HindIII DNA size markers; from the top these are 23.1kb, 9.4kb, 6.6kb, 4.4kb, 2.3kb and 2.0kb.
the same probe detected a single fragment whose length was the sum of the two cosmid fragments. From these data it was deduced that the central EagI site was uncut and therefore methylated in genomic DNA while the two flanking sites were cut and therefore unmethylated. In the same way it was shown that all but one of the EagI sites were unmethylated in the genome.

The conclusion from these studies was that all but one of the rare-cutter sites cloned by extending the cosmid walk were unmethylated in genomic DNA and therefore that these must be the sites in cluster 1 as detected by PFGE mapping. The methylation data are summarised in Figure 5.13.a.

The fact that the unmethylated rare-cutter sites in cluster 1 were spread over 20kb, whereas a typical CpG island spans 1-2kb, suggested that there may be more than one gene in this region. To test this hypothesis, cosmid fragments close to the unmethylated rare-cutter sites were isolated and used to probe northern blots, zoo blots and cDNA libraries to obtain evidence for transcribed sequences as described in section 5.5.

5.3. Cloning of cluster 2 by chromosome jumping

The cloning of the unmethylated rare-cutter sites in cluster 1 made feasible the use of rare-cutter jumping libraries to clone adjacent clusters (Poustka and Lehrach, 1988). In this technique, the two ends of large genomic DNA fragments generated by cleavage with a rare-cutter enzyme are co-cloned, as shown in Figure 5.5. A probe adjacent to a site for that rare-cutter enzyme in the genome (i.e. at one end of a rare-cutter fragment) can be used to screen the library and obtain cloned DNA from the other end of the same fragment, which may be hundreds of kilobases away. It was decided to use fragments adjacent to, and distal of, the BssHII and NotI sites of cluster 1 as probes to screen BssHII and NotI jumping libraries respectively. In theory, this approach should facilitate the cloning of the nearest distal unmethylated BssHII site (i.e. in cluster 2) and the nearest distal unmethylated NotI site (i.e. in cluster 3).
Figure 5.5.
Construction of a rare-cutter jumping library (Poustka and Lehrach, 1988). Details of the enzymes and selection system used are given in the text (this chapter and Materials and Methods).
The probe used to screen the NotI jumping library was 31KN2, a 1.9kb Kpn1/NotI fragment mapping adjacent to, and distal of, the NotI site in cosmid HPB.ALL 31 (Figure 5.2.). Unfortunately, no positives were obtained with this probe. One explanation for this negative result was that by chance the nearest site for the second enzyme used to construct the library was too far from the NotI site, thereby generating a fragment too large to clone into the phage vector.

The probe used to screen the BssHII jumping library was jBK, a 3.7kb BssHII/KpnI fragment isolated from cosmid HPB.ALL 31. (Figure 5.2.). This probe detected five positives which were picked and the recombinant phage DNA purified. By restriction enzyme mapping it was shown that four of the positives were identical (clone λj2) while the fifth was different (clone λj3).

The jumping clone λj3 was shown by restriction enzyme mapping with Sall and BamHI to contain two genomic inserts, a 0.7kb Sall/BamHI fragment and a 2.0kb Sall/BamHI fragment. In each case the Sall site was derived from the poly linker of the marker plasmid pMLS-Mlu-Not while the BamHI site was presumably genomic. The inserts were subcloned into the Bluescript plasmid vector. The 2.0kb Sall/BamHI fragment was shown by hybridisation to derive from the starting probe jBK. The 0.7kb fragment, when hybridised to PFGE blots, detected similarly sized fragments as the COL11A2 and DPB1 genes; in particular it detected the 130kb MluI band (data not shown). This result was unexpected because a successful 'jump' to the nearest unmethylated distal BssHII site (in cluster 2) should have resulted in the cloning of a fragment which hybridised to the 450kb MluI band (Figure 3.5.). In order to work out where the 0.7kb insert was derived from, it was hybridised to a blot of DNA from the entire cosm id walk, spanning from DPA1 to cluster 1 (Figure 5.1.). The probe hybridised strongly to cosm id MANN 3.6 and more weakly to MANN 2.2 and MANN 2.3. This suggested that the 0.7kb insert could be derived from the DPB1 gene and was also cross-hybridising to the related DPB2 gene. To test this, a portion of the subcloned insert was sequenced using primers complementary to the Bluescript vector cloning site, and the partial nucleotide sequence obtained was found to be identical over 153 nucleotides (8195-8348 in Kelly and Trowsdale, 1985) to a region of the
second intron of the \textit{DPB1} gene, except for a single nucleotide mismatch at position 8270. These data confirmed that the 0.7kb insert of \( \lambda j3 \) was derived from the \textit{DPB1} gene. The sequenced region mapped near a BssHII site in the second exon of the \textit{DPB1} gene (position 7736 in Kelly and Trowsdale, 1985).

It was therefore concluded that \( \lambda j3 \) contained the ends of an unusual jump from the BssHII site in cluster 1 to the BssHII site in the \textit{DPB1} gene (Figure 5.8.). The BssHII site in the \textit{DPB1} gene must have been unmethylated, and hence cleaved, in the cell line from which the library was constructed, although in PFGE analysis of this region in PGF DNA (and in other published maps from other cell lines) the cleavage of this BssHII site has never been observed.

The jumping clone \( \lambda j2 \) was shown by restriction mapping with SalI, BamHI and HindIII to contain two genomic inserts, a 1.1kb SalI/HindIII fragment and a 2.0kb SalI/BamHI fragment (Figure 5.6.a.). The SalI site in each case was derived from the polylinker of the marker plasmid, while the BamHI and HindIII sites were presumably genomic. The 2.0kb SalI/BamHI insert was shown by hybridisation to be derived from the starting probe, jBK, and was probably identical to the 2.0kb BamHI/SalI insert identified in \( \lambda j3 \). The 1.1kb SalI/HindIII fragment, designated \( \lambda j2SH \), was hybridised to PFGE blots and was found to detect a 250kb BssHII fragment, a 450kb MluI fragment and a 370kb NotI fragment. This pattern of hybridisation would be expected for a probe mapping between \textit{DPA2} and cluster 2 (Figure 3.5.). \( \lambda j2SH \) was then used to screen the cosmid library and a positive clone, HPB.ALL 71, was isolated and characterised by restriction enzyme mapping (Figure 5.6.b.). \( \lambda j2SH \) hybridised just centromeric of the most centromeric BssHII site in HPB.ALL 71.

From the PFGE map, cluster 2 contained unmethylated sites for BssHII and EagI (Figure 3.5.). Cosmid HPB.ALL 71 was found to contain five BssHII site and two EagI sites spread over about 7kb (Figure 5.6.b.). To prove that the sites in HPB.ALL 71 mapped to cluster 2 (and not to cryptic rare-cutter sites, as was found for \( \lambda j3 \)) it was necessary to demonstrate that at least one of the BssHII sites and one of the EagI sites in HPB.ALL 71 was unmethylated in genomic DNA.
Figure 5.6.
(a) Restriction map of the jumping clone \( \lambda j2 \). The thick black line represents the vector arms. The marker plasmid is represented by a shaded box and the two genomic inserts by unfilled boxes. B, BamHI; H, HindIII; S, Sall.
(b) Restriction map of the cosmid HPB.ALL 71, which was isolated by screening the library with the 1.1kb HindIII/Sall genomic insert from \( \lambda j2 \). The two BssHII fragments, 71Bs3 and 71Bs1, which were subsequently used as probes to screen the cDNA library, are shown. The scale bar applies to both clones.
Figure 5.7.
Determination of the methylation status of cloned BssHII and Eagl sites in genomic DNA. The autoradiographs show blots of genomic DNA (PGF) cut with BssHII or Eagl and hybridised with the 3.9kb BssHII fragment 71Bs1 from cosmid HPB.ALL 71 (Figure 5.6.).
All sizes are in kilobases.
Figure 5.8.
Overview of the region covered by chromosome jumping. Both jumping clones were isolated with the probe jBK from cosmid HPB.ALL 31 as described in the text. The other end of jumping clone \( \lambda j3 \) mapped to the previously isolated cosmid MANN 3.6 and was shown by sequencing to be derived from the DPB1 gene. The other end of the jumping clone \( \lambda j2 \) was derived from cluster 2 and was used to isolate a new cosmid clone, HPB.ALL 71. B, BssHII site.
This was done by hybridising cosmid fragment 71Bsl (3.9kb BssHII fragment, Figure 5.6.b.) to blots of genomic (PGF) DNA cut with BssHII or EagI and resolved by conventional electrophoresis. Probe 71Bsl detected a 3.9kb BssHII fragment and a 1.6kb EagI fragment in genomic DNA (Figure 5.7.). Thus the two BssHII sites flanking 71Bsl and the two EagI sites within 71Bsl were shown to be unmethylated in genomic DNA (Figure 5.13.b.). This fact, coupled with the PFGE mapping data from λj2SH, led to the conclusion that HPB.ALL 71 must map to cluster 2 (Figure 5.8.).

To test for the presence of genes at cluster 2, fragments mapping between the BssHII sites in HPB.ALL 71 were isolated and used as probes for zooblot and cDNA libraries as described in section 5.5.

5.4. Identification of cluster 3 in previously isolated cosmid clones

Cluster 3 was cloned by taking advantage of a previously established cosmid walk around the DOB gene (Figure 5.9.; Blanck and Strominger, 1988). These cosmid clones extended 60kb centromeric of the DOB gene. From the PFGE map shown in Figure 3.5. it was apparent that cluster 3 mapped on the centromeric side of the DOB gene, although its precise position could not be determined. It was decided to test for the presence of rare-cutter sites in the cosmids mapping centromeric of the DOB gene.

Cosmid clones U10 and U15 (a gift from George Blanck) were digested with BssHII and NotI. It was found that cosmid U15, the most proximal clone in the walk of Blanck and Strominger (Figure 5.9.a.), contained two BssHII sites and two NotI sites as shown in Figure 5.9.b. The methylation status of these sites in genomic DNA was not determined at this stage. However, a cDNA clone subsequently isolated with a probe from U15 was hybridised to a PFGE filter of human DNA digested with BssHII and NotI in single and double digest combinations (Figure 5.10.). The cDNA probe detected a fragment of about 10kb in each track (the fragment of about 440kb in the NotI track of Figure 5.10. was a residual signal from a prior hybridisation).
Figure 5.9.
(a) Map of overlapping cosmid clones isolated by Blanck and Strominger (1988). The DXβ, DXα, DVβ, DQβ and DQα genes are now known as DQB2, DQA2, DQB3, DQB1 and DQA1 respectively.
(b) Detailed restriction map of cosmid U15 and the overlapping portion of cosmid U10. The position of the probe U15KN is shown.
Figure 5.10.
Determination of the methylation status of cloned BssHII and NotI sites in genomic DNA. The autoradiograph shows the result of hybridising the RING4 cDNA probe 2.1 to a PFGE filter of human genomic DNA (PGF) digested with BssHII, BssHII+NotI and NotI. Size markers are concatemers of bacteriophage lambda DNA.
This result indicated that both NotI sites and both BssHII sites in U15 are unmethylated in genomic DNA (Figure 5.13.c).

5.5. Analysis of cloned regions for coding sequences

5.5.1. Zoo blots

Transcribed DNA sequences are more highly conserved during evolution than non-coding regions, and genomic fragments which cross-hybridise with the genomic DNA of other species may indicate the presence of genes (Monaco et al., 1986). This principle was used to test for the presence of genes in clusters 1 and 2. Restriction fragments mapping close to the unmethylated rare-cutter sites were isolated from the cosmid clones and probed onto 'zoo-blots' of EcoRI-cut genomic DNA from different vertebrate species. A representative result is shown in Figure 5.11. A summary of the results obtained with all probes tested is given in Table 5.2.

<table>
<thead>
<tr>
<th>SPECIES</th>
<th>PROBE</th>
<th>33X1</th>
<th>31K1</th>
<th>jBK</th>
<th>71Bs3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhesus monkey</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Pig</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Rat</td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Mouse</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Whale</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Chicken</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.2. Summary of results obtained by probing cosmid fragments from clusters 1 and 2 onto zoo blots of EcoRI-cut genomic DNA from a variety of vertebrate species. Positions of probes are shown in Figures 5.2. and 5.6. (+), cross hybridisation detected; (-), no cross hybridisation detected. Blots were washed to a final stringency of 2x SSC, 65°C. A representative result is shown in Figure 5.11.
Figure 5.11.
Zoo blot analysis of genomic fragment 33X1. The autoradiograph shows the result of hybridising probe 33X1 to a Southern blot of EcoRI-cut genomic DNA from different vertebrate species. The final washing stringency was 2x SSC at 65°C.
The results obtained with the probes listed in Table 5.2. provide good evidence for the conservation of these sequences in the genomes of other organisms. These fragments may therefore contain exons of genes. To tested this hypothesis, the probes were subsequently hybridised to northern blots to obtain evidence for transcription.

5.5.2. RNA blots

The conserved fragments identified through zoo blotting were hybridised to northern blots of poly(A)+ RNA from a variety of cell lines to test whether or not they detected transcribed sequences. A representative result with the genomic fragment 33X1 is shown in Figure 5.12. 33X1 detected an RNA species of about 1.6kb in RNA from the cell lines K562 (erythroleukaemia) and U937 (histiocytic leukaemia) but not HL60 (promyelocytic leukaemia) or WEHI (mouse promyelocytic leukaemia). The RNA species detected by 33X1 is probably a polyadenylated mRNA because it is increased in abundance in poly (A)+ RNA (Figure 5.12.). Similar results were obtained using the probe 31K1 from cluster 1 and λj2SH from cluster 2 (data not shown). However, this was not conclusive evidence that these genomic fragments encoded the gene (or part of the gene) giving rise to the transcript; it remained possible that the transcribed locus was elsewhere and that the genomic fragments tested here contained pseudogenes or other sequences related to the true expressed locus which were cross-hybridising with the RNA species under the conditions used. More convincing evidence that these probes were really part of functional genes was obtained by the isolation of cDNA clones, as described below.

No hybridisation to northern blots was detected with the probe jBK. In practice, however, it was often difficult to obtain convincing results by hybridising northern blots with genomic fragments. Negative results were difficult to interpret because a given genomic fragment could contain a very small portion of an exon, which may not hybridise under the conditions used. Furthermore, a given gene may be not be expressed in the cell lines tested at a level detectable by the sensitivity of the northern blotting technique. A more satisfactory approach to determining whether a genomic fragment encoded transcribed sequences was to hybridise that fragment directly to cDNA libraries.
Figure 5.12.
Northern blot analysis of genomic fragment 33X1. The autoradiograph shows the result of hybridising probe 33X1 to a northern blot of total and poly(A)+ RNA from different cell lines as described in the text. The final washing stringency was 2x SSPE at 65°C.
5.5.3. cDNA libraries

cDNA clones were isolated by screening cDNA library filters with each of the probes 33X1, 31K1, 31KN2 (Cluster 1), 71Bs3, 71Bs1 (Cluster 2) and U15KN (Cluster 3). All of these genomic fragments detected positive colonies. In each case, DNA from positive colonies was purified and the longest clone was generally selected for detailed characterisation (Table 5.3.). However, four cluster 2 cDNA clones were characterised because these were isolated using two different genomic fragments, 71Bs3 and 71Bs1, and it was not initially certain that only one gene was present, although this was quickly established by restriction enzyme mapping of the clones. The cDNA clone inserts were hybridised back to pulsed field blots to confirm that the only location in the human genome of sequences related to that cDNA clone was within the MHC. Each cDNA probe was found to give the hybridisation pattern expected from the known location of the corresponding genomic probe. This eliminated the possibility that the genomic fragments contained pseudogenes which were detecting cDNAs derived from transcripts from genes elsewhere.

The cDNA clones were hybridised back onto the relevant cosmid clones to define the approximate positions of the cognate genes (Figure 5.13.). These genes were designated RING (Really Interesting New Gene) 1-5. RING1, RING2 and RING5, in cluster 1, were respectively 95kb, 90kb and 85kb proximal to the DPB2 gene. RING3, in cluster 2, was 110kb distal of the DPA1 gene. RING4, in cluster 3, was 25kb proximal of the DOB gene (Figure 5.16.)
Figure 5.13.
New genes in clusters 1, 2 and 3. The positions of the rare-cutter sites within each cluster, together with the methylation status of each site in genomic DNA, are shown (B, BssHII; E, EagI; M, MluI; N, NotI; o, site unmethylated; x, methylated; -, not tested). The positions of the cosmid fragments from each cluster used to probe the cDNA library are indicated. Boxed regions show the positions of the five novel genes, RING1-5, as defined by mapping the cDNA clones back onto the cosmids. The maps are oriented with the centromere at the left.
Table 5.3. cDNA clones isolated by screening cDNA libraries with genomic fragments from clusters 1, 2 and 3. The positions of the probes and the cognate cDNAs within each cluster are summarised in Figure 5.13.

5.6. Expression patterns of RING 1-5

Northern blots containing RNA from a range of cell lines were hybridised with the cDNA clones for the novel genes RING1-5 and washed at high stringency (0.1x SSPE, 65°C). Representative results are shown in Figure 5.14. and summarised in Table 5.4.

The RING1 cDNA probe CEM15 detected a transcript of about 1.6kb in poly(A)+ RNA samples from all cell lines tested. The transcript was expressed at particularly high levels in Molt4, a T-lymphoblastoid cell line (LCL), although expression was also significantly higher in Mann (B-LCL) than in K562 (erythroleukaemia) or U937 (macrophage).

The RING2 cDNA probe CEM21 detected a transcript of about 1.1kb in poly(A)+ RNA from T- and B-LCLs. No expression was detected in the macrophage cell line U937 or the erythroleukaemia line K562, even after prolonged autoradiographic exposure.
Figure 5.14. Northern blot analysis of the novel genes RING1-5. The autoradiographs show the results of hybridising blots of poly(A)+ (first four panels) or total (right hand panel) RNA from different cell lines with cDNA probes from the genes RING1-5. Probes are: RING1, CEM15; RING2, CEM21; RING3, CEM41; RING5, 8U5; RING4, 2.1. In the first four panels, T is T-LCL (Molt4), B is B-LCL (Mann), E is erythroleukaemia (K562), M is macrophage (U937). In the right hand panel B is B-LCL (Raji), T is T-LCL (Molt4); 1, 2 and 3 are colon carcinoma lines SW1222, SW620 and CC20 respectively in the absence (-) and presence (+) of gamma interferon. Approximate transcript sizes are shown in kilobases.
The RING3 cDNA probe CEM41 detected two large transcripts, a major species of about 3.5kb and a less abundant species of about 4.5kb, in all cell lines tested. The same transcripts were detected in γ-interferon induced and uninduced colon carcinoma cell lines SW1222, SW620 and CC20, indicating that RING3 expression is not γ-interferon inducible in these cell lines.

The RING4 cDNA probe 2.1 detected a transcript of about 2.8kb which was expressed at high levels in the B-LCL tested (Raji), and at lower levels in the T-LCL (Molt4). No RING4 expression was detected in resting colon carcinoma cell lines SW1222, SW620 and CC20, but the transcript was strongly induced in γ-interferon-treated cells. In addition, RING4 expression was found to be both α and γ-interferon inducible in cells of the fibrosarcoma line HT1080 (J. John and G. Stark, personal communication).

The RING5 cDNA probe γU5 detected two equally abundant RNA species of about 2.0 and 2.3kb in all poly(A)+ samples tested.

<table>
<thead>
<tr>
<th>GENE</th>
<th>cDNA PROBE</th>
<th>RNA SIZE (APPROX.)</th>
<th>T</th>
<th>B</th>
<th>M</th>
<th>E</th>
<th>H</th>
<th>IFN</th>
</tr>
</thead>
<tbody>
<tr>
<td>RING1</td>
<td>CEM15</td>
<td>1.6 kb</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>NT</td>
</tr>
<tr>
<td>RING2</td>
<td>CEM21</td>
<td>1.1 kb</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>NT</td>
<td>NT</td>
</tr>
<tr>
<td>RING3</td>
<td>CEM41</td>
<td>3.5+4.5 kb</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>RING4</td>
<td>2.1</td>
<td>2.8 kb</td>
<td>+</td>
<td>+</td>
<td>NT</td>
<td>NT</td>
<td>NT</td>
<td>+</td>
</tr>
<tr>
<td>RING5</td>
<td>γU5</td>
<td>2.0+2.3 kb</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>NT</td>
</tr>
</tbody>
</table>

Table 5.4. Summary of expression patterns of the genes RING1-5. +, expression detected; -, no expression detected. T, T-LCL; B, B-LCL; M, macrophage U937; E, erythroleukaemia K562; H, HeLa; IFN, α- and γ-interferon inducible (Figure 5.14.).
5.7. Refinement of the physical map of the class II region

Blanck and Strominger (1990) have recently published a cosmid walk which extends 120kb in the vicinity of the DNA gene (Figure 5.15.a). These cosmid clones were not linked to any other class II genes, and the precise position and orientation of the DNA gene therefore remained ambiguous. From the PFGE mapping data summarised in Figure 3.5. it was hypothesised that cluster 2 might fall within this cosmid contig. The probe 71Bsl (Figure 5.6.b.) from the cluster 2 cosmid HPB.ALL 71 was hybridised to DNA from the published cosmids (a gift from George Blanck). This probe cross-hybridised to the overlapping cosmid clones U22, O27 and HA14 (Figure 5.15.b). From the map shown in Figure 5.15.a., the cross-hybridising region was about 35kb from the DNA gene. When these cosmids were mapped with BssHII, it was found that they contained the same pattern of sites as HPB.ALL 71 (Figure 5.15.b.), providing additional evidence that cluster 2 is indeed contained within the cosmid walk. Since PFGE data show that the DNA gene is on the centromeric side of cluster 2 (Figure 3.5.) it follows that the DNA gene is about 35kb centromeric of RING3. Thus, the cosmid walk of Blanck and Strominger is oriented within the class II region such that the DNA gene has the same transcriptional orientation as DPA1 (i.e. the map shown in Figure 5.15.a. is oriented with the centromere towards the right). From the PFGE map it is estimated that the DNA gene is 75kb from the DPA1 gene.

The mapping of cluster 3 rare-cutter sites in the cosmid U15, which maps just centromeric to the DOB gene, anchors the cosmid walk of Blanck and Strominger (1988; Figure 5.9.a.) within the class II region relative to the PFGE map. It is estimated that the DOB gene is 160kb telomeric of the DNA gene.

A physical map of the class II region showing the positions of RING1-5 and the accurately determined positions of the class II genes relative to the clusters of rare-cutter sites is shown in Figure 5.16.
Figure 5.15.
(a) Map of cosmid clones in the region of the DNA gene (formerly DZA) from Blanck and Strominger (1990).
(b) Autoradiographs of cosmid DNA digested with Kpnl (i) or BssHII (ii) and hybridised with the 3.9kb BssHII fragment from the RING3 locus in cosmid HPB.ALL 71. '71' is cosmid HPB.ALL 71; other cosmids are shown in (a). Bars represent lambda/HindIII DNA size markers; from the top these are 23.1kb, 9.4kb, 6.6kb, 4.4kb, 2.3kb, 2.0kb and 0.56kb.
Figure 5.16.
Physical map of the class II region of the human MHC showing the positions of the COL11A2 gene and the novel genes RING1-5 with respect to the previously known class II genes and the clusters of rare-cutter sites. Gene orientations are shown by arrows where known. The orientations of RING1, RING3 and RING4 were determined during nucleotide sequencing (Chapter 6).
5.8. Summary and discussion

The data presented here provide evidence for five new genes in the class II region of the human MHC. *RING1, RING2* and *RING5* map centromeric of the DP subregion, *RING3* maps 35kb telomeric of *DNA* and *RING4* maps 25kb centromeric of *DOB* (Figure 5.16.). The expression patterns and transcript sizes of these genes are dissimilar to those characterised for the classical class II genes. In addition, sequence data from each cDNA clone (see Chapter 6) indicate that none is related to the class II genes. Thus the discovery of *RING3* and *RING4* in the middle of the class II region demonstrates for the first time that the classical class II genes are interspersed with non-class II sequences. This finding has implications for the understanding of phenotypes associated with the class II region, since genes closely physically linked to the known class II genes are candidates for involvement in class II-associated diseases (section 1.5.). Furthermore, both *RING3* and *RING4* map in the interval believed to contain the gene responsible for a defect in antigen presentation by class I molecules (Cerundolo et al., 1990; section 1.6.2.). Elucidation of the function of these novel genes is clearly of importance. As a first step towards this goal, nucleotide sequencing of *RING1-5* cDNA clones was undertaken (Chapter 6).

The finding of non-class II genes in the class II region also has implications for the understanding of the evolution of the region. The class II region of the MHC is thought to have evolved from a single primordial α/β gene pair through a series of duplication and divergence events (Klein, 1986; Bodmer et al., 1986). The similarity of the overall organisation of the classical class II genes of mouse and man have led to the hypothesis that these duplication and divergence events occurred before the radiation of rodents and primates (section 1.4.). To test whether the *RING3* and *RING4* genes have become inserted into the middle of the human class II region after the divergence of rodents and primates, or whether their position was established beforehand, the presence of the novel genes in the mouse class II region was tested. This work is described in Chapter 7.
The strategy used to detect novel genes, i.e. the identification of CpG islands, is likely to influence the type of gene discovered. In a search of the nucleotide sequence databases, CpG-rich regions were found to be associated with all ubiquitously expressed 'housekeeping' genes, but only some tissue-specific genes (Bird, 1986; Gardiner-Garden and Frommer, 1986). Consistent with this, expression of RING1-5 was detected in most cell lines tested (Figure 5.14. and Table 5.4.).

It is likely that there are further genes associated with the clusters of rare-cutter sites. Cosmid U15, for example, contains two NotI sites, both of which are unmethylated in genomic DNA. Given that most NotI sites are found in CpG islands (Table 3.1.), it would be surprising if there were no further genes in this region. Experiments to test this are currently underway in the laboratory.

In the present study, the methods used would specifically detect those genes associated with unmethylated CpG islands. CpG islands generally remain stably unmethylated in all tissues of the body, regardless of the site of expression of the associated gene (Bird, 1987). However, it has recently been reported that in some cell lines a proportion of CpG islands become irreversibly methylated and would therefore not be detectable in genomic DNA (Antequera et al., 1990). Such islands would, however, be detectable by clustering of rare-cutter sites in cloned DNA and could be tested for in the recently described yeast artificial chromosome clones which span the entire class II region (Ragoussis et al., 1991b). Such clones could also be used to test for the presence of non-island-associated genes. The recently developed 'exon trapping' technique, to test for the presence of splice junctions in a cloned region, is a potentially powerful method to identify genes regardless of whether or not they are associated with CpG islands (Duyk et al., 1990).
6. Characterisation of novel genes by nucleotide sequencing

6.1. Introduction

One method of obtaining valuable preliminary information about the possible functions of a gene is to determine the nucleotide sequence of that gene. A gene encoding a protein product should contain an open reading frame (ORF), the predicted amino acid sequence of which can be used to search protein sequence databases. Important clues as to the function of the protein may be obtained from significant matches (Doolittle, 1986). Since the RNA species detected by the probes for the novel genes RING1-5 were all found at increased frequency in poly(A)+ selected RNA samples compared to total RNA (Figure 5.14. and data not shown), it was likely that the transcripts of these genes were polyadenylated, and therefore encoded proteins. This chapter describes the complete nucleotide sequencing of cDNA clones for the RING1, RING3 and RING4 genes and the partial nucleotide sequencing of the RING2 and RING5 genes.

6.2. Nucleotide sequence of RING1

The nucleotide sequence of RING1 was obtained from the cDNA clone CEM15. The insert was estimated to be about 1.5kb long, and contained an internal XhoI site. The CDM8 vector contained XhoI sites either side of the cloning site so that two insert fragments were liberated when CEM15 was digested with XhoI. These were subcloned into the Bluescript vector to give two constructs: pXB, containing the ~0.65kb XhoI insert fragment, and pXT, containing the ~0.85kb XhoI insert fragment. The two ends of each subclone were sequenced using oligonucleotide primers complementary to the Bluescript cloning site. Thereafter, the novel sequence data obtained were used to design new complementary oligonucleotides which were used as primers to extend
the sequence further. Finally, primers were designed to sequence across the internal XhoI site of the CEM15 insert.

The CEM15 insert was 1439 nucleotides in length. The first 1116 of these encoded a single uninterrupted ORF of 372 amino acids, ending with a stop codon. The ORF was followed by a 3' untranslated region of 323 nucleotides, terminating in 6 A residues which were potentially the start of a poly(A) tail. The sequence AACAAA, which resembles the consensus poly(A) addition signal AATAAA, started 25 nucleotides upstream of the first of the 6 A residues. The AACAAA variant seems to be relatively inefficient in initiating 3' processing of RNAs, since in an in vitro 3' processing assay it was found to direct message cleavage and polyadenylation with only 4% of the efficiency of the consensus AATAAA (Sheets et al., 1990). Nevertheless, in a search of the nucleotide sequence databases, the AACAAA variant was detected in 0.8% of vertebrate cDNAs, and does therefore seem to be a naturally occurring poly(A)-addition signal, albeit rare (Sheets et al., 1990). It may be that additional nucleotides in the vicinity of the AACAAA sequence increase the efficiency of 3' processing of RING1 mRNAs in vivo. It is also possible that the 6 A residues are not part of the poly(A) tail, in which case the cDNA clone is truncated at the 3' end before the true tail.

The ORF began at the first codon of the CEM15 insert and there was no upstream in-frame stop codon. The CEM15 cDNA clone was therefore probably truncated at the 5' end. Since the length of the cDNA insert (1.44kb without a poly(A) tail) corresponded well to the transcript size detected in poly(A)+ RNA (1.6kb with a poly(A) tail), it was concluded that the CEM15 cDNA insert was probably missing only a few nucleotides at the 5' end. It was therefore decided to obtain the sequence of the corresponding genomic region to see whether a starting methionine residue and an in-frame upstream stop codon were present.

The two subclones of CEM15, pXT and pXB, were hybridised back to cosmid DNA to determine the orientation of the RING1 gene. pXT hybridised to the 4.1kb XhoI fragment (33X3) of cosmid HPB.ALL 33 (Figure 5.2.) while pXB hybridised to the adjacent 1.7kb XhoI fragment, 33X1 (data not shown). From the sequence data it was known that pXT...
encoded the 5' end of RING1, while pXB encoded the 3' end. Thus, the RING1 gene must be oriented with the 3' end towards the centromere. The 4.1kb genomic XhoI fragment, 33X3, containing the 5' portion of the RING1 gene, was subcloned from the cosmid HPB.ALL 33 into the Bluescript vector. A part of the resulting construct, p33X3, was sequenced using a primer complementary to the 5' end of the CEM15 insert and extending further in the 5' direction. The sequence so obtained revealed the presence of a methionine codon just 5 codons upstream of the first codon in the CEM15 insert (Figure 6.1.). The methionine codon was in a sequence context of CCATAATGG, corresponding well to the general initiation consensus of CCACCATGG (Kozak, 1986). An in-frame termination codon was found a further 21 nucleotides upstream. This methionine was judged likely to be the initiating codon of the RING1 gene product, and for the purpose of the following discussion it will be assumed that the sequence shown in Figure 6.1., with an ORF of 377 amino acids, is correct. However, because this sequence was obtained from a genomic clone rather than a cDNA clone, there is no evidence that it is present in a mature RING1 transcript. This region could be part of an intron in the RING1 gene, and subsequently spliced out during transcript processing.

The predicted protein product of the RING1 gene was rich in glycine in the C-terminal two-thirds (27%, compared to a mean of 7% in human proteins; Doolittle, 1986). The N-terminal third, in comparison, was relatively depleted of glycine. The protein contained 10 cysteine residues, 7 of which were clustered together in the N-terminal region. Another feature of interest was the motif KRPR, starting at residue 172. The same sequence is found in the nuclear localisation signal of polyoma virus large-T antigen (Richardson et al., 1986).

The amino acid sequence of the predicted protein product of the RING1 gene was used to search the OWL(9.1) protein sequence database (collaboration with Dr. Paul Freemont, Protein Structure Laboratory, ICRF). When the entire RING1 sequence was used as the query sequence, weak identity was detected with a number of database sequences, all of which were related to RING1 solely on the basis of their high glycine content. However, when the glycine-rich C-terminal two-
Figure 6.1.

Nucleotide sequence of the RING1 gene. Nucleotides after the vertical line were obtained from the cDNA clone CEM15. Nucleotides before the line were obtained from a genomic subclone as described in the text. The clustered cysteine and histidine residues are circled, and the potential nuclear localisation signal is underlined.
<table>
<thead>
<tr>
<th>RAG</th>
<th>KSISCCICE</th>
<th>HILADPV</th>
<th>ETNCKhMVFCRVCILRCLKVMGS</th>
<th>YCPSCRYPCFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAD18</td>
<td>TLLRCCHICK</td>
<td>DFLKVPV</td>
<td>LT PGCHIFCSLCHRTHLNNQPN</td>
<td>CPLCLFEFRE</td>
</tr>
<tr>
<td>IE110</td>
<td>EGDVCAWCT</td>
<td>DEIAPHLRC</td>
<td>DTFPCMRFCIPCMK TWMQLRN</td>
<td>TCPLCNAKLYV</td>
</tr>
<tr>
<td>VZ61</td>
<td>SDNCTICLM</td>
<td>STVSDLG</td>
<td>KTMPCLHDRCFVPCIR AWTSTSV</td>
<td>QCPLCRCPVQS</td>
</tr>
<tr>
<td>CG30</td>
<td>VKLQCNICFSVAEIKYFIPIDRLTIIPVLELDTCKQLCSMCIRKIRKRV</td>
<td>PCPLCRVESLH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RET</td>
<td>QETTCPVCL</td>
<td>QYFAEPM</td>
<td>ML DCNHICCAACLARCWGTAET</td>
<td>NVSCPQCRETFPQ</td>
</tr>
<tr>
<td>RPT-1</td>
<td>EEVTPICL</td>
<td>ELLKEPV</td>
<td>SA DCNHSFCRACITLYESNRNTDGKGNCPVCRVPYPF</td>
<td></td>
</tr>
<tr>
<td>RING1</td>
<td>SELECTPICL</td>
<td>.DMLKNTM</td>
<td>TTKECILHRGCDICVTALRSGNK</td>
<td>ECPCTRKKLVS</td>
</tr>
</tbody>
</table>

**Figure 6.2.**
Alignment of the amino acid sequences of the homologues of the RING1 gene product. Boxed residues are conserved, or conservatively substituted, between proteins and constitute the RING1 motif described in the text. The identities of the homologous genes are shown in Table 6.1. and described in the text.
thirds of the RING1 amino acid sequence were eliminated and the 139 N-terminal amino acids were used as the query sequence, a much more significant match was detected with seven other proteins (Table 6.1.).

<table>
<thead>
<tr>
<th>GENE NAME</th>
<th>ORGANISM</th>
<th>TOTAL AMINO ACIDS</th>
<th>MOTIF START</th>
</tr>
</thead>
<tbody>
<tr>
<td>RING1</td>
<td>Human</td>
<td>377</td>
<td>19</td>
</tr>
<tr>
<td>RAG-1</td>
<td>Human</td>
<td>1043</td>
<td>290</td>
</tr>
<tr>
<td>rfp (ret)</td>
<td>Human</td>
<td>513</td>
<td>16</td>
</tr>
<tr>
<td>rpt-1</td>
<td>Mouse</td>
<td>353</td>
<td>15</td>
</tr>
<tr>
<td>RAD18</td>
<td>S. cerevisiae</td>
<td>487</td>
<td>29</td>
</tr>
<tr>
<td>IE110</td>
<td>Herpes simplex</td>
<td>775</td>
<td>116</td>
</tr>
<tr>
<td>VZ61</td>
<td>Varicella-zoster virus</td>
<td>467</td>
<td>19</td>
</tr>
<tr>
<td>CG30</td>
<td>Baculovirus</td>
<td>264</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 6.1. Homologues of RING1 detected by searching the protein databases with the N-terminal 139 amino acids of the putative RING1 product. The total number of amino acids in the protein product of each homologue, and the position of the first cysteine residue of the conserved motif, are shown.

When these proteins were aligned to maximise identity between them, it was apparent that they all shared a previously undescribed cysteine-histidine motif, with a consensus sequence of:

\[-C-X-(I,V)-C-X_{11-30}-C-X-H-X-(F,I,L)-C-(I,L,M)-X_{10-18}-C-P-X-C-\]

(Figure 6.2.). In each protein, the motif was always contained within the N-terminal third, and frequently started very close to the N-terminus (Table 6.1.).

The motif is reminiscent of those found in zinc-binding transcription factors, in which cysteine residues, or a combination of cysteine and
histidine residues, co-ordinate zinc ions, thereby stabilising the structure of a sequence specific nucleic-acid binding domain (Berg, 1990). One class of zinc-binding transcription factors includes the well-studied archetypal 'zinc-finger' protein TFIIIA from *Xenopus laevis*, which has zinc-dependent sequence-specific DNA and RNA binding activity. It is now known that there are large families of TFIIIA-related genes in the genomes of both vertebrates and invertebrates which include developmentally important loci such as *Kruppel* in *Drosophila melanogaster* (Dressler and Gruss, 1988). The consensus amino acid sequence of the zinc-binding domain of these proteins is:

\[-(\text{Tyr,F})-\text{X-C-Y-C}-\]

A single protein may contain several of these domains. The pair of cysteine residues and the pair of histidine residues within each domain co-ordinate a single zinc ion, which is thought to induce folding of the domain to form an \(\alpha\)-helix which interacts with DNA (Lee et al., 1989b; Berg, 1990).

A second well-characterised class of transcription factors, the steroid-hormone receptors, also have zinc-dependent sequence-specific DNA binding domains. Each receptor molecule contains one zinc-binding domain with a consensus sequence of:

\[-\text{C-X-C-X-C-X-C-X-C-X-C-S}^{-}\]

The first two pairs of cysteine residues co-ordinate one zinc ion while the next two pairs co-ordinate a second ion. The structure thus created is believed to contain a short \(\alpha\)-helical region around the fourth cysteine residue which makes sequence-specific contacts with the DNA double helix (Hard et al., 1990).

A number of other distinct cysteine-histidine motifs have now been recognised in other proteins and although these are generally less well characterised some of them are known to be responsible for a zinc-dependent nucleic acid-binding function (Berg, 1990). It is therefore very tempting to speculate that the RING1 motif may also be involved in metal-dependent nucleic acid binding.
The known functions of some of the *RING1* homologues detected in the databases are highly suggestive of a role for the motif in DNA binding and, in some cases, transcriptional activation. The *IE110* gene of herpes simplex virus type 1 encodes a nuclear phosphoprotein which is required for transcriptional activation of later virus genes (Perry et al., 1986). The mouse *rpl-1* gene is expressed in resting CD4+ T-lymphocytes and encodes a nuclear protein which down-regulates expression of the interleukin-2 receptor α chain gene (Patarca et al., 1988). Normal expression of the human *rfp* gene is up-regulated during spermatogenesis, but the 5' end of this gene, including the region encoding the cysteine-histidine motif, has also been identified in a fusion gene, *ret*, which can transform NIH 3T3 cells (Takahashi and Cooper, 1987; Takahashi et al., 1988). The yeast *RAD18* gene is required for postreplication repair of DNA damaged by UV light or other agents. The human *RAG-1* gene activates DNA recombination at V(D)J sequences found at the T-cell receptor and immunoglobulin gene loci, a process which leads to the generation of diversity in the protein products of these genes (Schatz et al., 1989). The functions of products of the varicella-zoster virus gene *VZ61* and the baculovirus gene *CG30* are unknown (Davison and Scott, 1986; Thiem and Miller, 1989). Despite the tantalising suggestion of a DNA binding function for these gene products, there is as yet no evidence that any of the *RING1* homologues bind DNA or zinc, or even that the cysteine-histidine motif is necessary for the function of these proteins. However, based on the speculation that the phylogenetically conserved cysteine and histidine residues of the *RING1* motif are involved in co-ordinating divalent metal ions and creating a DNA-binding domain, a hypothetical structure for such a complex has been proposed and is presented in Figure 6.3. (Dr. Paul Freemont, personal communication).

One approach to establishing the biological function of the *RING1* motif involves the PCR-based cloning of human genomic sequences which bind to the cysteine-histidine-rich region (Kinzler and Vogelstein, 1989). This method is currently being used (collaboration with Dr. Ruth Lovering) to determine whether the motif has a sequence specific DNA binding activity. The possible role of the *RING1* gene product in MHC-associated phenotypes is discussed in Chapter 9.
Figure 6.3.
(a) Schematic illustration of the co-ordination of zinc ions by two TFIIIA-type zinc fingers.
(b) Schematic illustration of the hypothetical co-ordination of metal ions by the RING1-related motif.
RING2 CEM21 5' end

GAC GCG GCC GTG TCC GGC GTG TTT GTG CCT ACC TCA GGA GCG AGG GGC ATC GGC CGA 54
A R S G V F C P T S G A G S G I G R 18

GCG GTC AGT GTA CGC CTG GCC GGA GAG GGG GCC ACC GTA GCT GCC TGC GAC CTG 108
A V S V R L A G E G A T V A A C D L 36

GAC CGG GCA GCG GCA CAG GAG ACG GTG CGG CTG CTG GCC GGC CGG CCA GGG AGC AAG 162
D R A A A Q E T V R L L G G P G S X 54

GAG GGG CCC CCC CGA GGG AAC CAT GCT GCC TTC CAG GCT GAC GTG TCT GAG GCC 216
E G P P R G N H A A F Q A D V S E A 72

AGG GCC GCC AGG TGC CTG CTG GAA CAA GTG CAG GCC TTC CTG CTG CAG CAC 270
R A A R C L L E Q V Q A C F S R P P 90

TCT GTC GTT GTG TCC GTG TGC GGC ATC ACC CAG GAAT GAG TTT CTG CTG CAC ATG 324
S V V V S C A G I T Q D E F L L H M 108

GAA CCT CCG GAG TTT GCC CCC TTA GCT GGG AGG AGT TGG AGG AGG GCT GTG 429
E P R E F G P L A G R S W R R A V 143

Figure 6.4.
Nucleotide and predicted amino acid sequence of the 5' end of the RING2 cDNA clone CEM21. The sole methionine codon within the open reading frame is shown in bold type. The termination codons in the other two reading frames are underlined.
6.3. Partial nucleotide sequence of RING2

The two ends of the RING2 cDNA clone CEM21 were sequenced using primers complementary to the cDNA vector CDM8 either side of the cloning site. One end (primed with the oligonucleotide CDM8-F) clearly contained a long tract of A residues, characteristic of a poly(A) tail, but could not be read further, presumably due to secondary structure created by the homopolymeric tail. The nucleotide sequence at the other end of CEM21, presumably near the 5' end of the gene, was obtained by priming the sequencing reaction with the oligonucleotide CDM8-B. This sequence was extended further in the 3' direction by priming the reaction with an oligonucleotide containing nucleotides 86-107 (Figure 6.4.). A total of 429 nucleotides were sequenced in this way (Figure 6.4.). This region potentially encoded a single long ORF of 143 amino acid residues, including a single methionine at position 108. This methionine codon was not in the optimal initiation consensus sequence (CCACCATGG; Kozak, 1986). Both other reading frames contained termination codons throughout the region sequenced. It is not possible at this stage to tell whether the ORF detected at the 5' end of the cDNA clone is that encoding the RING2 protein product (in which case the CEM21 cDNA clone is truncated at the 5' end), or whether it occurs by chance in the 3' untranslated region of the RING2 gene. This question will be resolved with the solution of the complete nucleotide sequence of CEM21. The complete amino acid sequence of the ORF was used to search the PIR and Swiss-Protein sequence databases but no significant match was found.

6.4. Nucleotide sequence of RING3

Four RING3 cDNA clones were isolated (Figure 5.13.). CEM32 and CEM35 were isolated by screening the library with the genomic BssHIII fragment 71Bs1, while CEM41 and CEM44 were isolated by screening the library with an adjacent BssHIII fragment, 71Bs3. Theoretically the two genomic probes could encode part of two separate genes. However, when the cDNA clones were mapped with restriction enzymes it was apparent that they were derived from the same gene because they all
shared several sites in common (Figure 6.5.). Together, the four cDNAs provided a useful nested set of clones within the RING3 gene. The nucleotide sequence of the two ends of each cDNA was determined, using CDM8-F and -B primers, in order to obtain the sequence of RING3 at several points.

The sequence data from the truncated 3' ends of the CEM44 and CEM41 cDNA clones revealed potential ORFs whose predicted amino acid sequences were used to search the databases. Both sequences identified the product of fsh, a Drosophila homeotic gene, as the top match. Because of the striking homology between fsh and these partial sequences from RING3, and the extremely interesting phenotypes of mutations in the fsh locus (see below) it was decided to determine the complete nucleotide sequence of the longest RING3 cDNA clone, CEM32.

The nucleotide sequence CEM32 was obtained by sonicating the CEM32 insert and blunt-end ligating the resulting short fragments into M13. The M13 clones were then sequenced and aligned to build up the sequence of the CEM32 insert (collaboration with Dr. Stephan Beck). The nucleotide sequence of the 4004-nucleotide CEM32 insert is shown in Figure 6.6. The sequence at the 5' end of the CEM32 insert exactly matched that previously determined at the 5' end of the CEM35 insert, beginning at the same nucleotide. Therefore it is very likely that this is the true 5' end of the RING3 transcript. The CEM32 insert contained a single long ORF. The first potential methionine residue within this ORF was not in a good context for initiation of translation, while the second was in the context CCACAATGG, which differs by only one nucleotide from the consensus of CCACCATGG (Kozak, 1986). The second methionine was therefore considered most likely to be the initiating codon of the RING3 protein. Use of the second methionine would give a protein product of 754 amino acids and a 5' untranslated region of 1153 nucleotides (Figure 6.6.). The ORF was followed by a 3' untranslated region of 589 nucleotides which was truncated before the poly(A) tail but which contained a sequence exactly matching the consensus poly(A) addition signal (AATAAA) just before the end (starting at nucleotide 3991).
Figure 6.5.
Restriction maps and orientations of the four RING3 cDNA clones. 'F' and 'B' refer to the oligonucleotide primers which were used to obtain the nucleotide sequence at the end of each cDNA clone as described in the text.
Figure 6.6.

Complete nucleotide sequence of the RING3 cDNA clone CEM32. The predicted amino acid sequence is shown, starting at the methionine residue in the best context for initiation. An additional upstream methionine codon is boxed and the in-frame upstream termination codon is underlined. The potential nuclear localisation signal within the predicted amino acid sequence is boxed.
The predicted protein product contained a potential nuclear localisation signal, KKKRK, starting at residue 508. In addition it contained regions which were enriched in one particular amino acid. Thus, a highly acidic glutamate-rich region was present between residues 445-468, a basic lysine-rich region was present between residues 496-516 and a serine-rich region was present between residues 729-753. A further feature of interest was an internal duplication in the amino acid sequence (Figure 6.6.). Two stretches of 26 amino acids starting at positions 50 and 322 shared 19 residues in common (73% identity).

When the amino acid sequence of the predicted protein product of the RING3 gene was used to search the protein sequence databases a highly significant match was obtained with the 1106-amino acid product of the 5.9kb transcript of the Drosophila melanogaster gene female sterile homeotic (fsh; Haynes et al., 1989). The amino acid sequence identity between the two proteins is shown in Figure 6.7. This alignment reveals three domains which are strikingly conserved. In one stretch of 120 amino acids (positions 22-142 in RING3), 91 residues (76%) were identical in the two proteins and a further 9 were highly conservative substitutions. The human protein seems to have two large regions deleted relative to the Drosophila protein, which accounts for its smaller size. Interestingly, the two regions which have appear to been deleted from RING3 contained proposed transmembrane regions in fsh. It had previously been suggested that the fsh protein may be an integral membrane protein (Haynes et al., 1989). However, the absence of these regions in the human protein raises the possibility that they are not transmembrane domains in fsh. It may be significant that many of the stretches of amino acids which were proposed to be membrane-spanning domains were rich in alanine, the hydrophobicity of which is controversial (Haynes et al., 1989).

A further similarity between RING3 and fsh is that both give two major transcripts. The RING3 transcripts are 3.5kb and 4.5kb while the fsh transcripts are 5.9kb and 7.6kb. The fsh transcripts arise through alternative RNA processing and the predicted product of the 7.6kb transcript contains the same 1106 amino acids encoded by the 5.9kb transcript with an additional 946 residues at the C-terminal end (Haynes et al., 1989). The CEM32 insert is thus related to both fsh gene
Alignment of the amino acid sequences of the human RING3 and Drosophila 5.9-fsh gene products. Residues which are identical in the two proteins are indicated by colons. Dashes indicate gaps which have been introduced to maximize the homology. The alignment starts with residue 1 of the RING3 protein and residue 8 of the fsh protein.
products but seems to be analogous to the 5.9kb transcript because the
ORF terminates at the same position (Figure 6.7.). It is not clear at this
stage whether CEM32 is derived from the 3.5kb or 4.5kb $RING3$
transcript. It will be necessary to characterise additional cDNA clones
and the genomic structure of the $RING3$ locus in order to determine
the relationship between the $RING3$ transcripts and the $fsh$ transcripts.

A role for the $fsh$ locus in Drosophila development has been suggested
by the observation that $fsh$ expression is required both maternally
during oogenesis and then later in embryogenesis for normal embry-
onic pattern formation to occur. Furthermore, certain mutant $fsh$
alleles interact synergistically with mutant alleles at other develop-
mentally important loci (such as the homeotic gene Ultrabithorax) to
increase the frequency of abnormalities in the segmentation pattern
of the embryo (Haynes et al., 1989). It has also been directly shown, using
immunofluorescence-tagged antibodies against the product of the
Kruppel gene (another developmentally important locus) that Kruppel
expression is altered in $fsh$-mutant embryos (Huang and Dawid, 1990).
These observations suggest that the $fsh$ gene product(s) plays a role in
the complex interaction between gene products which is known to be
critical for normal pattern formation in the developing Drosophila
embryo. It is not yet known how this interaction is mediated.

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 exciting possibility that $RING3$ may be
critical in human embryonic development. It will be of great interest to
test the expression of $RING3$ in embryonic tissues.

The region which is duplicated within the predicted protein product of
the $RING3$ gene (Figure 6.6.) shows homology to a duplicated domain
in the human CCG1 protein (Sekiguchi et al., 1988). CCG1 was
identified in a transfection assay as the gene which complements
temperature-sensitive mutant baby hamster kidney cell lines which are
arrested in the G1 phase of the cell cycle. The mechanism by which the
CCG1 gene product overcomes the block in the cell cycle is unknown,
and the function of the duplicated domain therefore remains obscure. The homology between the CCG1 and RING3 gene products is confined to the region of internal homology. This duplicated motif may define a novel protein domain.

6.5. Nucleotide sequence of RING4

The complete nucleotide sequence of the ~2.6kb RING4 cDNA clone 2.1 was determined by Ian Mockridge and Adrian Kelly in this laboratory. The cDNA clone contained a single long ORF which started at the 5' end. A methionine codon occurred at nucleotide 84 but this was not preceded by an in-frame upstream stop codon. A portion of the cosmid clone U15 containing the 5' end of the RING4 gene was sequenced (collaboration with Dr. Stephan Beck, ICRF), and the ORF was found to extend a further 180 nucleotides in the 5' direction before a stop codon was encountered. The sequence presented in Figure 6.8. is a composite of the cDNA sequence and the genomic sequence. As with RING1, there is as yet no evidence that this genomic sequence is part of the transcribed region, but at present it will be assumed that the sequence shown in Figure 6.8. is correct. The composite ORF contained two potential initiating methionine residues (arrowed in Figure 6.8.), one immediately following the upstream in-frame stop codon in the genomic sequence and a second at amino acid position 61 (nucleotide 211) in the cDNA clone. Use of the first methionine would give a protein containing 808 amino acids, while use of the second would give a product of 748 amino acids. The ORF was followed by a 3' untranslated region of 370 nucleotides and a poly(A) tail. Starting 19 nucleotides upstream of the tract of A residues was the sequence AATAAAA, which is identical to the consensus signal for 3' processing of mRNAs (Sheets et al., 1990).

The N-terminal two-thirds of the predicted protein product of the RING4 gene was highly hydrophobic in character. The potential membrane-spanning regions indicated in Figure 6.8. were revealed in a hydropathicity analysis by Dr. Michael Sternberg (Figure 6.9.). The C-terminal third, in contrast, was more hydrophilic. When the predicted amino acid sequence of the RING4 product was used to search the pro-
Figure 6.8.

Nucleotide sequence of the RING4 gene. Nucleotide sequences after the vertical line were obtained from the cDNA clone 2.1. Sequences before the vertical line were obtained from a genomic subclone. Two potential initiating methionine residues are arrowed. Three consensus N-linked glycosylation sites are indicated by squares. Potential transmembrane regions in the ORF are underlined. The polyadenylation sequence in the 3' untranslated region is underlined. The region encoding the ATP-binding domain is boxed.
Figure 6.9.
Assignment of potential membrane-spanning regions in the predicted protein product of the RING4 gene. The plot was generated using a computer program written by Dr. Mike Sternberg, ICRF, which is based on the algorithm of Rao and Argos (1986) to predict the location of transmembrane regions. Short horizontal lines indicate the positions of putative membrane-spanning domains.
tein sequence data bases, significant homology was found between a stretch of amino acids within the C-terminal region (boxed in Figure 6.8.) and the ATP binding site of members of the 'ABC' (ATP-binding cassette) superfamily of energy-dependent transport proteins (Figure 6.10.), which includes P-glycoprotein, the plasma-membrane-associated pump responsible for the multi-drug resistance phenotype found in some human tumour cells (Juranka et al., 1989; Hyde et al., 1990). The physiological substrate of P-glycoprotein is uncertain, but other members of the ABC superfamily are each specialised for the transport of a particular substrate. Known substrates include sugars, inorganic ions, amino acids, peptides and proteins.

In general, ABC transporters have a similar overall organisation to P-glycoprotein, shown in Figure 6.11., with two highly hydrophobic integral membrane domains and two ATP-binding domains (Hyde et al., 1990; Juranka et al., 1989). The four domains may be present in one polypeptide encoded by a single gene, or they may be encoded by more than one gene so that the functional transporter is a multi-subunit complex. P-glycoprotein, for example, is encoded by a single gene, while the oligopeptide permease of bacteria is the product of four separate genes. The RING4 product which contains one hydrophobic domain and one ATP-binding domain falls between these two extremes, resembling one-half of a functional ABC transporter molecule, like the hlyB haemolysin transporter of E. coli (Hyde et al., 1990; Felmlee et al., 1985). If the RING4 product is part of an ABC transporter, the functional molecule may be a RING4-homodimer, or a heterodimer with a RING4-related gene product.

The possibility that the RING4 product is part of an ABC transporter may be particularly relevant for the understanding of the class II-associated defect in antigen presentation by class I molecules as discussed in section 1.6.2. Studies of mutant cell lines in man and rat have implicated a role for a class II-encoded function in normal class I antigen presentation in these species (Livingstone et al., 1989; Cerundolo et al., 1990). The RING4 gene maps within the interval known to contain this function in the human mutant B-cell line LBL 721.174. This cell line expresses normal class I molecules but these are unable to present intracellular viral antigens and do not progress from
RING4 GLTFTLRPGEVTALVGFNGSKSTVAALQN VAAVGQPQVFGRSLQENIAYGLTQK DEAGSQLSGGQRQAVALARALKPCVLILDDATSALDNSQ

     --SLEVKK-QTL-----SS-C-----VQ--ER LGI-S----IL-DC-IA-----DNSR GDK-T-------K-RI-I-----V-Q-HI-L--E------TE-E
ste6  NVSLNFSA-QF-FI--KS------LSN--LR ITV-E-RCTL-NDTLRK--LL-S-DS GTG-VT-------Q-R--I----FE-DTPF-F-E-V----IVHR
     NMN-DMFC-QTLEG1-ES-T----LVL--TK ISV-E-K-LL-NGTIRD-LT---QDE RIDTLT-----A-RLCI-----L--SKI----EC----SV-S
     oppF  -V-LR-YE--TLGV--ES-C-----F-RAIIG IQMIF-D-LASLNFRTMIGI11AEIFL NRYPHEF---ECNRIGI-----LE-KLI-C----V----VSIQ
cyAB  NVSLRIA-----VGV--RS------LTR-I-R LGV-L--STL-N--VRD-----LTRPGA G-N-VG--------RIGI-----HR-R--------E-----YE-E

Figure 6.10.
Alignment of the amino acid sequences of three conserved regions in the ATP-binding fold (boxed domain in Figure 6.8.) of the RING4 product and related proteins. Identical residues are indicated by a dash. HmDr, human multidrug resistance protein; Ste6, yeast mating factor transporter; OppD and OppF, S. typhimurium oligopeptide permease; HylB, E.coli haemolysin transporter; CyaB, Bordetella pertussis adenylate cyclase transporter (Trowsdale et al., 1990).
Figure 6.11.
Model of the transmembrane topology of P-glycoprotein, illustrating the structure of the members of the ABC superfamily of transporters. Each circle represents an amino acid; branched structures represent N-linked carbohydrate. The tandemly duplicated structure is clearly evident. The predicted product of the RING4 gene contains one membrane-associated domain and one ATP-binding domain and would therefore have a structure similar to one half of P-glycoprotein.
the endoplasmic reticulum (ER), which is thought to be the site of class I/peptide interaction. The defect can be relieved by the addition of exogenous viral peptides, which induce folding of the class I molecules, association with β-2 microglobulin, and transport of the complex to the cell surface where normal antigen presentation occurs. This observation has been interpreted as showing that LBL 721.174 is defective in the transport of intracellularly derived peptides from their site of generation in the cytosol to their site of binding to class I, in the ER (Cerundolo et al., 1990). It is known that some members of the ABC family are specialised for the transport of peptides and proteins; thus, the product of the STE6 gene in *S. cerevisiae* is responsible for export of the 12-amino acid a mating factor (McGrath and Varshavsky, 1989), the product of the *E. coli hlyB* gene transports the protein haemolysin (Mr 107kD; Felmlee et al., 1985), and the product of the *opp* (oligopeptide permease) operon of *S. typhimurium* transports peptides of up to 5 amino acids in length (Hiles et al., 1987). These members of the ABC superfamily are specialised for the transmembrane transport of polypeptides which do not have signal sequences by a mechanism independent of the normal secretory pathway. An attractive hypothesis is that the RING4 product is part of a signal-independent peptide transporter which is associated with the ER membrane and which pumps peptides from the cytosol into the lumen of the ER, where binding to class I molecules then occurs (Trowsdale et al., 1990). The role of the RING4 product in class I antigen presentation is currently being tested by transfecting the RING4 gene into mutant cell lines and testing for complementation of the defect (collaboration with Dr. Alain Townsend).

The RING4 gene is an exciting candidate for explaining some class II-associated autoimmune diseases. It can be speculated that different allelic forms of this gene could have a profound influence on the immune response by determining which peptides ultimately become associated with class I molecules and presented to cytotoxic T-lymphocytes. Thus, in theory, the RING4 protein could promote the presentation of certain autoantigens to autoreactive T-cells, and participate in the initiation of an autoimmune response.
6.6. Partial nucleotide sequence of \textit{RING5}

The two ends of the insert of the 2.3kb \textit{RING5} cDNA clone γU5 were sequenced using the two synthetic oligonucleotide primers (CDM8-F and CDM8-B) complementary to the CDM8 vector on either side of the cloning site. The sequence primed by CDM8-F extended for 162 nucleotides and contained a single ORF of 54 amino acid residues (Figure 6.12.) This ORF contained a stretch of hydrophobic amino acid residues followed by a region unusually rich in histidine residues. The partial amino acid sequence of the putative \textit{RING5} protein product was found to be very similar to the N-terminal region of the predicted product of a mouse gene, \textit{KE4} (St.-Jacques et al., 1990). When the \textit{RING5} and \textit{KE4} gene sequences were aligned, 145 out of 162 nucleotides (89.5%) were identical. At the amino acid level, 45 out of 54 (83%) residues were identical (Figure 6.12.). The γU5 sequence primed by the CDM8-B oligonucleotide contained stop codons in all three reading frames. When this sequence was compared with the sequence of the \textit{KE4} cDNA, significant identity was detected with the 3' untranslated region of \textit{KE4}. It was therefore concluded that \textit{RING5} and \textit{KE4} are homologues of the same gene. Consistent with this conclusion are the observations that the \textit{RING5} cDNA clone γU5 detects cross-hybridising sequences 45kb proximal of the mouse \textit{Pb} gene, which is known to be the map position of \textit{KE4} (see Chapter 7 for further details), and that \textit{KE4} and \textit{RING5} were both found to be expressed in all tissues tested (St.-Jacques et al., 1990; Table 5.4.). It is interesting however that \textit{RING5} appears to give rise to two major RNA species, while \textit{KE4} produces only one (Figure 5.14.; St.-Jacques et al., 1990). Furthermore, the \textit{KE4} transcript was estimated to be 2.8kb in length, while the \textit{RING5} transcripts are 2.0 and 2.3kb. The explanation for these differences will await the complete sequencing of cDNA clones for the two forms of \textit{RING5} and comparison of these sequences to \textit{KE4}. The protein coding region of the 2.7kb \textit{KE4} cDNA which was sequenced in the study of St.-Jacques et al. (1990) spanned only 1.3kb of the total and it is therefore possible that the differences in length observed between the mouse and human transcripts occur in the untranslated regions.
Figure 6.12.

Partial nucleotide sequence and predicted amino acid sequence of γU5, the RING5 cDNA clone. The nucleotide sequence of the homologous mouse gene KE4 is shown above the human sequence, and the predicted amino acid sequence is shown below (St.-Jacques et al., 1990). Amino acids and nucleotides which are identical in the two species are indicated by a dash. The first amino acid of the partial RING5 sequence shown here aligns with the eighth amino acid in the predicted product of the KE4 gene.
The KE4 gene was identified in a chromosome walk around the H-2K subregion of the mouse MHC in an attempt to clone the gene responsible for the embryonic lethal mutation \( t^{w5} \) (Abe et al., 1988; Artzt et al., 1988). Five genes, designated KE1-5, were identified in the cloned region and four of these, including KE4, were found to be expressed in 5.5 day-old mouse embryos, the time at which the \( t^{w5} \) mutation causes embryonic death. As a step towards characterising the candidate genes for the \( t^{w5} \) phenotype, the complete nucleotide sequence of KE4 was obtained by St.-Jacques et al. (1990). The predicted protein product contained 436 amino acids and had two domains which were extremely rich in histidine residues. The first of these was partly contained within the RING5 sequence (Figure 6.12.). Within these domains, histidine was typically found as every alternate residue. The KE4 product also contained three stretches of hydrophobic residues characteristic of transmembrane domains and a further stretch at the N-terminus which was characteristic of a signal sequence. The KE4 product may therefore be an integral membrane protein. Its function at this stage remains obscure but its role in development, and particularly in the \( t^{w5} \) phenotype, could be tested by the construction of transgenic mice.

6.7. Summary

The sequence data described in this chapter have given intriguing insights into the possible functions of four of the five novel genes. The RING1 gene encodes a protein containing a novel cysteine-histidine motif which is conserved in a range of species including yeast and which may interact with DNA. The RING3 gene is homologous to the developmentally important fsh locus of D. melanogaster. Partial sequencing of the RING5 gene revealed an ORF with homology to the KE4 gene of the mouse, which maps to a region known to contain a developmental lethal. Finally, the RING4 gene product appears to be an ABC-type transmembrane pump, related to bacterial peptide transporters. Experiments to gain further insights into the functions of the products of these genes and suggestions for their possible roles in explaining MHC-associated phenotypes are discussed in Chapter 9.

155
7. Comparative mapping of novel genes in the MHC region of mouse and man

This chapter describes the mapping of sequences homologous to the human genes RING1-5 and COL11A2 in the mouse MHC, and the mapping of sequences homologous to three novel mouse genes, KE1-3, in the human MHC.

7.1. Introduction

The organisation of genes in the human and mouse class II regions is very similar. The finding that the subregions are essentially homologous, and that the order of the subregions with respect to one another is conserved (section 1.4.) has led to the proposal that the organisation of the class II region was established before the radiation of primates and rodents (Klein, 1986; Bodmer et al., 1986). Modifications of this organisation, such as the duplication of the human DQA1 and DQB1 gene pair to give the DQA2 and DQB2 genes, and the deletion in the mouse of sequences homologous to the human genes DPA1 or DPA2, are thought to have occurred more recently (Bodmer et al., 1986). It was therefore of interest to test for the presence of homologues of the human genes RING3 and RING4 in the mouse class II region in order to determine whether these genes were also present in the ancestral class II region or whether they have more recently become inserted into the human MHC.

The region immediately centromeric of the class II region, containing the genes RING1, RING2 and RING5, is also interesting from an evolutionary point of view. The major difference in overall organisation of the mouse and human MHCs is that the mouse has a pair of class I genes, K and K2, mapping about 70kb proximal of the most proximal class II gene, Pb (Figures 1.7. and 7.1.). From sequence data and restriction mapping studies the K and K2 genes are thought to have arisen by duplication in the Qa region of the mouse class I gene.
Figure 7.1.
Molecular map of the proximal end of the mouse MHC.
The main diagram shows the 600kb cosmid walk of Steinmetz et al. (1986) which links the class II genes (Pb, Ob, Ab, Aa, Eb, Eb2 and Ea) with the class I gene pair (K and K2) in the BALB/c mouse (H-2d haplotype). Above this map is shown the 190kb region cloned from the Iu5 haplotype in an independent study by Uehara et al. (1987). Also indicated are the positions of the five novel genes, KE1-5, which were found within this cloned region by Abe et al. (1988). Arrows show the transcriptional orientation of the genes, where known. Both maps are to the same scale.
cluster and to have subsequently moved to their current location by an
intrachromosomal double cross-over event (Bodmer, 1981; Weiss et al.,
1984). This organisation is also found in the rat, but seems to be unique
to rodents because every other species examined to date, including
man, has all the class I genes grouped together in a single region
mapping further towards the telomere (Klein, 1986). The report of five
novel genes, KE1-5, in the immediate vicinity of K and K2, raised the
possibility that these genes may also have been moved from the class I
region (Abe et al., 1988; Figure 7.1.). In this case, it would be predicted
that the human homologues of the mouse KE genes should map to the
human class I region. It has been already been shown in Chapter 6 that
RING5, which maps 85kb centromeric of DPB2, is the human
homologue of the mouse KE4 gene, which maps 45kb centromeric of Pb.
This result suggests that the KE4 gene has not come from the class I
region with K and K2 but rather that its position at the centromeric end
of the class II region was established before the radiation of rodents and
primates. To clarify the situation for the other KE genes, and to
determine whether RING1 and RING2 were homologous to any of the
KE genes, it was decided to determine the positions of the KE genes in
the human MHC and of the RING genes in the mouse MHC.

7.2. Determination of the positions of sequences homologous to
the human genes RING1-5 and COL11A2 in the mouse MHC

The entire class II region and the H-2K subregion of the BALB/c (H-2\(^d\))
mouse has been cloned in a single contig of overlapping cosmid clones
(Figure 7.1.; Steinmetz et al., 1986). To determine whether or not the
human genes RING1-5 and COL11A2 have homologues in the mouse
MHC, cDNA probes from RING1-5 and a genomic probe from
COL11A2 were hybridised to blots of DNA from these mouse cosmid.
It was found that all six of the human genes cross-hybridised with
specific mouse cosmid clones (Figure 7.2.).

The RING1, RING2 and RING5 cDNA probes (CEM15, CEM21 and \(\gamma\)U5
respectively) all cross-hybridised with two overlapping mouse cosmid,
II5.8 and II6.4 (Figure 7.2.). The RING2 and RING5 probes detected
similarly sized HindIII fragments. The region of overlap between the
Figure 7.2.
Detection of sequences cross-hybridising to the human genes RING1-5 and COL11A2 in the mouse MHC. The autoradiographs show the results of hybridising probes from each of these genes to blots of HindIII-cut (first five panels) or EcoRI-cut (last panel) DNA from cosmid clones in the mouse class II region (Figure 7.1.). Probes were: RING1, CEM15 cDNA; RING2, CEM21 cDNA; RING5, λU5 cDNA; COL11A2, 4.5kb BamHI/EcoRI genomic fragment; RING3, CEM41 cDNA; RING4, 2.1 cDNA. The final washing stringency was 6x SSC, 65°C for the RING5, COL11A2 and RING4 probes and 2x SSC, 65°C for the RING1, RING2 and RING3 probes. Bars at left represent lambda/HindIII DNA size markers; from top these are 23.1kb, 9.4kb, 6.6kb, 4.4kb, 2.3kb, 2.0kb and 0.56kb.
two cosmids is about 40kb (Figure 7.1.). Detailed mapping of II5.8 and II6.4 with the restriction enzymes KpnI and SalI allowed the relative position of the putative mouse homologues of \textit{RING1}, \textit{RING2} and \textit{RING5} to be determined, according to detailed cosmid maps kindly provided by Dr. Michael Steinmetz (data not shown). Thus the putative mouse homologues, as defined by cross-hybridisation, of \textit{RING1}, \textit{RING2} and \textit{RING5} were found to be respectively about 60kb, 50kb and 45kb proximal to the \textit{Pb} gene, and therefore they have the same relative position in mouse as in human (Figure 7.7.). This result is consistent with \textit{RING5} being the human homologue of the mouse \textit{KE4} gene as determined by nucleotide sequencing (section 6.6.), since \textit{KE4} is known to map 45kb proximal of \textit{Pb} (Figure 7.1.).

The \textit{RING3} cDNA probe CEM41 cross-hybridised with two overlapping mouse cosmids, II2.27a and II4.24. This places the putative mouse homologue of \textit{RING3} between 30 and 40kb distal of the \textit{Pb} gene (Figures 7.2. and 7.1.).

The \textit{RING4} cDNA probe 2.1 cross-hybridised with mouse cosmid II5.9, which places the mouse homologue of \textit{RING4} between 10 and 40kb proximal of the \textit{Ob} gene (Figures 7.2. and 7.1.). This result has recently been confirmed by Deverson et al. (1990), who have described a gene in this region in both mouse and rat which is related to \textit{RING4} at the nucleotide sequence level.

The human \textit{COL11A2} probe, a genomic 4.5kb BamHI/EcoRI fragment, cross-hybridised with mouse cosmid II10.6 (Figure 7.2.; the bands seen in the '5.8' and '6.4' tracks are due to prior hybridisation of the same filter with the \textit{RING1} probe). Cosmid II10.6 also contains the \textit{Pb} gene (Figure 7.1.). Detailed restriction enzyme mapping of II10.6 revealed that the region cross-hybridising to the \textit{COL11A2} probe was about 10kb proximal of the \textit{Pb} gene (data not shown). This region has now been sequenced in the mouse, and the presence of the mouse \textit{COL11A2} gene confirmed (Dr. Kathy Cheah, personal communication; L. Stubbs and K. Cheah, manuscript in preparation).

It was apparent from these results that the positions of \textit{RING1-5} and \textit{COL11A2} with respect to one another and to the characterised class II
genes are strikingly conserved in mouse and man, as summarised in Figure 7.7. It was apparent that the only genes in the proximal region of the human MHC which did not have a homologue in the mouse MHC were DNA and DPA1/DPA2. Previous attempts to search for novel class II α-chain genes in the mouse class II region by screening mouse genomic libraries at low stringency with mouse α-chain gene probes had not revealed homologues of DNA or DPA1/DPA2 although it is possible that such genes may have diverged too greatly at the nucleotide sequence level to be detected in this way. It was therefore decided to hybridise the human DNA and DPA1 gene probes directly to DNA from the mouse class II region cosmids at low stringency. Washing at 55°C in 6x SSC revealed a single cross-hybridising band in cosmid II2.27a with the human DNA probe (data not shown). This result suggests that there are indeed sequences related to DNA between Pb and RING3 in the mouse (Figure 7.7.). No cross-hybridising bands were detected in the region near Pb using the human DPA1 probe under the same conditions. It remains possible that there is a mouse homologue of DPA1 but that it is highly diverged.

7.3. Determination of the positions of sequences homologous to three mouse genes, KE3-5, in the human MHC

Probes for the mouse genes KE2, KE3, KE4 and KE5 were a gift from Dr. Kuniya Abe. These probes were genomic fragments from mouse cosmid clones as described in Table 7.1. Each has been previously reported to detect sequences in human genomic DNA (Abe et al., 1988). The conditions for cross-hybridisation to sequences in human genomic DNA were determined for each mouse probe in the present study by hybridising them to human Southern blots at 65°C and washing at gradually increasing stringency (Table 7.1.).

It was shown in section 6.6. that the RING5 gene is the human homologue of the mouse KE4 gene. It was therefore hypothesised that the human homologues of the other KE genes may also map near RING5. The KE gene probes were hybridised to blots of DNA from overlapping human cosmids in this region, using the conditions determined for detection of cross-hybridising sequences.
<table>
<thead>
<tr>
<th>MOUSE GENE</th>
<th>PROBE IDENTITY</th>
<th>CROSS-HYBRIDISATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>KE2</td>
<td>No.5; 1.7kb BamHI fragment</td>
<td>6x SSC, 65°C</td>
</tr>
<tr>
<td>KE3</td>
<td>No.7; 1.8kb PstI fragment</td>
<td>2x SSC, 65°C</td>
</tr>
<tr>
<td>KE4</td>
<td>No.11; 4.9kb BamHI fragment</td>
<td>2x SSC, 65°C</td>
</tr>
<tr>
<td>KE5</td>
<td>No.14; 3.9kb PstI fragment</td>
<td>6x SSC, 65°C</td>
</tr>
</tbody>
</table>

Table 7.1. Description of genomic probes for the mouse genes KE2-5. The second column shows the probe number, as assigned by Abe et al. (1988), and the size of the probe. The third column shows the washing conditions under which cross-hybridisation of each probe to human genomic DNA was observed.

The KE5 probe cross-hybridised to fragments in cosmids HPB.ALL 31 and cosHcol.11, as shown in Figure 7.3. The large EcoRI band detected in HPB.ALL 31 is shorter than that detected in cosHcol.11 because the HPB.ALL 31 cosmid insert terminates within this EcoRI fragment (Figure 5.2.). The results are consistent with the presence of a homologue of KE5 mapping within the ~12kb EcoRI fragment just proximal to the COL11A2 gene.

The KE4 probe cross-hybridised to two adjacent EcoRI fragments in the region of overlap between cosmids HPB.ALL 31 and HPB.ALL 33 (Figures 7.3. and 5.2.). These fragments encode the RING5 and RING2 genes. In addition, the KE4 probe was found to cross-hybridise to cDNA clones for both the RING5 and RING2 genes (data not shown). This was surprising, given that nucleotide sequencing had revealed that RING2 was unrelated to KE4, while RING5 and KE4 were homologous. However, it has recently been reported that the KE4 probe, a 4.9kb genomic BamHI fragment, detects two RNA species of 1.5kb and 2.8kb on mouse northern blots and encodes part of two physically close but distinct genes (St.-Jacques et al., 1990). The 2.8kb transcript is the product of the originally described KE4 gene (Abe et al., 1988; St.-Jacques et al., 1990). Since it has already been demonstrated that RING5 is the human homologue of the true KE4 gene, it is most likely that RING2 is
Figure 7.3.
Detection of sequences cross-hybridising to the mouse genes KE4 and KE5 in the human MHC. The autoradiographs show the results of hybridising mouse genomic probes for the genes KE4 and KE5 to blots of EcoRI-digested DNA from overlapping human cosmids at the proximal end of the human class II region (Figure 5.1.). 33 is cosmid HPB.ALL 33; 31 is cosmid HPB.ALL 31; Col is cosmid cosHcol.11; 1 is cosmid HPB.ALL 1; 8 is cosmid HPB.ALL 8; 2.3 is cosmid MANN 2.3. The final washing stringency was 2x SSC, 65°C for the KE4 probe and 6x SSC, 65°C for the KE5 probe.
the homologue of the second (uncharacterised) mouse gene encoded by the KE4 probe. This explanation is consistent with the results described in section 7.2., where mouse sequences cross-hybridising to RING2 and RING5 were found to be in close proximity.

Probes for the mouse genes KE2 and KE3, which map centromeric of the K and K2 genes, did not cross-hybridise with the human cosmid clones. The KE3 probe was used to screen the human cosmid library and to isolate a novel clone, HPB.ALL 51 (Figure 7.4.). This cosmid was mapped with restriction enzymes and found to contain a 2.6kb BamHI fragment, designated B51, which cross-hybridised with the KE3 probe. B51 was used as a probe to map the putative human homologue of the KE3 gene. B51 was first hybridised to Southern blots of DNA from the somatic cell hybrids, MCP-6 and 56-47, as described for the mapping of COL11A2 in section 4.2. The B51 probe detected a large EcoRI fragment of about 15kb in both hybrids, indicating that B51 maps to 6p21 (data not shown). B51 was then mapped relative to the MHC by PFGE. It was found that B51 detected similarly sized bands in human DNA cut with NotI, MluI, MluI+NotI, NruI and NruI+NotI as the probe 33X1 (Figure 7.5. and Table 7.2.). Apart from these shared bands, 33X1 detected a 100kb BssHII fragment which was not cleaved by NruI, while B51 detected a 50kb BssHII fragment which was cut by NruI to yield a very small fragment of about 10-15kb in length (Figure 7.5. and Table 7.2.).

These results facilitated the accurate mapping of B51, the putative human homologue of the mouse KE3 gene, between 80-95kb centromeric of the RING1 gene (Figure 7.6.). Consistent with the genomic mapping results, the cosmid HPB.ALL 51 was found to contain a single NruI site and a single BssHII site separated by about 15kb (Figure 7.4). Furthermore, the B51 probe mapped between these two sites as judged by hybridising B51 to blots of cosmid DNA (data not shown).

In conclusion, the positions of the human sequences which cross-hybridise with probes for the mouse genes KE3, KE4 and KE5 are in the same order as their mouse counterparts (Figure 7.7.).
**Figure 7.4.**
Restriction site map of the human cosmid HPB.ALL 51, which was isolated by screening the library with a probe from the mouse gene KE3. Within HPB.ALL 51, the KE3 probe cross-hybridised to a 2.6kb BamHI fragment, B51, which mapped between the BssHII site and the most distal XhoI site as shown.
Figure 7.5.
Pulsed-field mapping of B51, the putative human homologue of the mouse KE3 gene. The first two panels show a PFGE filter hybridised sequentially with 33X1 (Figure 5.2.) and B51. Lane 1, human genomic DNA digested with NotI; lane 2, BssHII+NotI; lane 3, BssHII; lane 4, BssHII+MluI; lane 5, MluI; lane 6, MluI+NotI. The third panel shows a different filter hybridised with B51. Lane 7, NruI; lane 8, BssHII; lane 9, BssHII+NruI; lane 10, NruI+NotI; lane 11, NotI. Fragment sizes are indicated in kilobases. LM, limiting mobility. Size markers in the second panel are concatemers of bacteriophage lambda DNA.
Figure 7.6.
Physical map of the region proximal to cluster 1 as determined by pulsed-field mapping. The positions of the genes RING1, RING2 and RING5 and of the probe B51, the putative human homologue of the mouse gene KE3, are shown. The position of the cosmid HPB.ALL 51, which was isolated using the KE3 probe, is also indicated. B, BssHII; M, MluI; N, NotI; R, Nrul.
### Table 7.2. Sizes of fragments detected on PFGE blots with the probes B51 and 33X1.

<table>
<thead>
<tr>
<th>DIGEST</th>
<th>B51 (kb)</th>
<th>33X1 (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BssHII</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>MluI</td>
<td>230</td>
<td>230</td>
</tr>
<tr>
<td>NotI</td>
<td>440</td>
<td>440</td>
</tr>
<tr>
<td>NruI</td>
<td>&gt;900</td>
<td>&gt;900</td>
</tr>
<tr>
<td>BssHII+MluI</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>BssHII+NotI</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>MluI+NotI</td>
<td>230</td>
<td>230</td>
</tr>
<tr>
<td>NruI+NotI</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>NruI+BssHII</td>
<td>10-15</td>
<td>100</td>
</tr>
</tbody>
</table>

### 7.4. Summary and discussion

This chapter has described the comparative mapping of the proximal MHC regions of man and mouse. At this stage, the human homologues of the mouse KE3 and KE5 genes and the mouse homologues of the human RING1, RING2, RING3 and DNA genes have been defined only by cross-hybridisation. It will be necessary to identify and sequence the genes within these cross-hybridising regions to formally prove homology, as has already been done for the RING4, RING5/KE4, and COL11A2 genes (Deverson et al., 1990; Monaco et al., 1990; Chapter 6, this study; Dr. Lisa Stubbs, manuscript in preparation).

The comparative map of the proximal region of the mouse and human MHCs is probably the most detailed for any part of the mammalian genome (Figure 7.7.). The relative positions of the putative homologues of RING1-5, COL11A2 and KE3-5 are strikingly similar in the two species. This degree of conservation strongly suggests that the establishment of the molecular genetic organisation of the entire proximal MHC region predates the radiation of rodents and primates.
Figure 7.7.
Comparative map of human and mouse class II regions. Homologous sequences are linked by dashed lines. Open boxes represent class II genes; hatched boxes represent class I genes; black boxes represent non-class I-class II genes. The positions of the human homologues of the mouse KE3 and KE5 genes, and the positions of the mouse homologues of the human genes RING1-3 and COL11A2 have been defined only by cross-hybridisation as discussed in the text. The sizes of the genes are not to scale.
The organisational differences which are apparent in this region in mouse and man have almost certainly occurred more recently. The most obvious of these differences is the presence of the two class I genes, K and K2, between KE3 and RING1 in the mouse. The fact that the gene organisation in this region is otherwise conserved indicates that that the K gene pair has been inserted into the gap between KE3 and RING1 in the mouse. The maximum size of the translocated region is given by the distance between KE3 and RING1 in the mouse, which is estimated to be 60kb. Other modifications of the organisation of the class II region, such as the duplication of the human DQA1 and DQB1 gene pair to give DQA2 and DQB2, and the deletion from the mouse class II region of any sequences homologous to DPA1 or DPA2, have clearly occurred without disturbing the closely linked non-class II genes.

It is noticeable that the distances between genes is often smaller in the mouse, in keeping with the observation that the overall size of the mouse class II region is about 300kb, compared to 900kb in man. Thus, the human COL11A2 gene is 45kb proximal to the DPB2 gene while the mouse COL11A2 gene is only about 10kb proximal to the Pb gene. Similarly, RING3 is 110kb distal to the HLA-DP subregion in man but in mouse the region cross-hybridising with RING3 is only 30-40kb distal of Pb. It is striking, however, that the distance between RING3 and RING4 is virtually identical in both mouse and man (about 100kb, Figure 7.7.). This interval may well be the same size in the two species because it contains conserved genes.

It is intriguing that a homologue of the RING4 gene is conserved in the class II region of mouse and rat (this study; Deverson et al., 1990; Monaco et al., 1990). The nucleotide sequence and map location of the human RING4 gene has led to the suggestion that it encodes an ATP-driven transporter which pumps peptides from the cytoplasm into the lumen of the endoplasmic reticulum where binding to class I molecules then occurs (section 6.5.; Trowsdale et al., 1990). It is possible that the location within the MHC of a gene encoding a product which is potentially so intimately involved with the function of class I antigens provides a selective advantage, with RING4 and class I gene sequences co-evolving within the same gene complex where they
would be rarely separated at meiosis. From this point of view it will be interesting to test for the presence of a homologue of \textit{RING}4, (and of the other \textit{RING} and \textit{KE} genes) in the MHCs of more distantly related species, such as the chicken.

The mouse homologues of the human \textit{RING}1, \textit{RING}2 and \textit{RING}5 genes map within the region of the mouse MHC believed to contain the \textit{tw}5 embryonic lethal gene (Artzt et al., 1988). \textit{RING}5 is the human homologue of the mouse \textit{KE}4 gene, which is already being considered as a candidate for \textit{tw}5 (St.-Jacques et al., 1990). The mouse \textit{RING}1 and \textit{RING}2 genes are additional candidates for the \textit{tw}5 gene. It will be of particular relevance to determine whether the mouse homologues of \textit{RING}1 and \textit{RING}2 are expressed in early embryogenesis.
8. RFLP analysis of novel genes

8.1. Introduction

As described in section 1.5., many diseases are known to show an association at the population level with the class II region. Theoretically the basis of the association could be an influence of the class II allele with which the association was originally detected or of any gene in linkage disequilibrium with the original marker. To determine whether any alleles of the novel genes (\textit{COL11A2} and \textit{RING1-5}) are in linkage disequilibrium with any class II alleles, and whether these alleles are more closely associated than the class II alleles with any diseases, it is first necessary to define polymorphisms in the novel genes with which the relevant studies can be performed.

8.2. Approach to identifying RFLPs

The genomic DNA samples used in this study were from random Caucasian donors and were a gift from Susan Tonks, Tissue Antigen Laboratory, ICRF. Genomic DNA from five individuals, giving a sample size of ten chromosomes, was typically digested with each enzyme used. Digested DNA was resolved on 0.8% agarose gels in 1x TAE buffer. Gels were run to maximise the resolution of larger fragments, and it is therefore possible that some small fragments may have run off the end of the gel during electrophoresis. On the filters of MspI- and TaqI-cut DNA, the maximum fragment size retained was about 1.5kb. On the other filters, the maximum fragment size retained was about 1.2kb. All hybridisations were performed at 65°C and washed to 0.1x SSC at 65°C to avoid cross-hybridisation with any related sequences. In practice, some weaker bands could occasionally be detected after long exposures of hybridised filters. These were not recorded in the present analysis.
8.3. RFLP analysis of COL11A2

The probe used was a 4.5kb BamHI/EcoRI genomic fragment subcloned from the cosmid cosHcol.11 (Hanson et al., 1989). The results obtained with this probe are shown in Table 8.1.

<table>
<thead>
<tr>
<th>ENZYME</th>
<th>CHROMOSOMES TESTED</th>
<th>FRAGMENTS</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>EcoRI</td>
<td>30</td>
<td>11kb</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9kb</td>
<td>0.47</td>
</tr>
<tr>
<td>BglII</td>
<td>10</td>
<td>9.5kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.0kb</td>
<td>1.0</td>
</tr>
<tr>
<td>MspI</td>
<td>10</td>
<td>none</td>
<td></td>
</tr>
<tr>
<td>TaqI</td>
<td>10</td>
<td>4.5kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.0kb</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 8.1. Fragment sizes detected with the genomic COL11A2 probe in Southern blots of human DNA cut with the enzymes indicated.

The COL11A2 probe detected a frequent RFLP in EcoRI-digested DNA samples. Of fifteen samples tested, two contained a single cross-hybridising fragment of about 11kb and one contained a single cross-hybridising fragment of about 9kb. These individuals were presumably homozygous for the 11kb fragment and the 9kb fragment respectively. In the other twelve samples the COL11A2 probe detected two fragments, one of 11kb and one of 9kb. These individuals were presumably heterozygotes. In the sample tested here the frequency of the two fragments was 0.53 (11kb) and 0.47 (9kb). Cheah et al. (1990a) have reported a similar observation, and have demonstrated by family studies that the polymorphic bands are allelic. In addition, BamHI and StuI RFLPs have been demonstrated near the COL11A2 gene (Priestly et al., 1990; Cheah et al., 1990b). The potential applications of COL11A2 RFLPs are discussed in Chapter 9.
No bands were detected in MspI-digested samples, probably because the cross-hybridising fragments were small and had run off the end of the gel during electrophoresis. However, 50 MspI-cut samples were tested in the study of Cheah et al. (1990a), and no polymorphism was detected. No TaqI or BglII polymorphism was detected in the small sample used here.

8.4. RFLP analysis of RING1

The probes used were 33X1, a 1.7kb genomic XhoI fragment from cosmid HPB.ALL 33 which encodes the 3' end of the RING1 gene (Figure 5.2.), and CEM15, the RING1 cDNA clone (Figure 5.13.). Results obtained with these probes are listed in Tables 8.2. and 8.3. On the basis of these data it was concluded that the RING1 gene did not contain any high-frequency polymorphisms for the ten restriction endonucleases tested.

<table>
<thead>
<tr>
<th>ENZYME</th>
<th>CHROMOSOMES TESTED</th>
<th>FRAGMENTS</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>BamHI</td>
<td>10</td>
<td>4.3kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.3kb</td>
<td>1.0</td>
</tr>
<tr>
<td>BglII</td>
<td>8</td>
<td>4.3kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.0kb</td>
<td>1.0</td>
</tr>
<tr>
<td>DraI</td>
<td>10</td>
<td>5.8kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.8kb</td>
<td>1.0</td>
</tr>
<tr>
<td>EcoRI</td>
<td>10</td>
<td>~15kb</td>
<td>1.0</td>
</tr>
<tr>
<td>EcoRV</td>
<td>10</td>
<td>&gt;20kb</td>
<td>1.0</td>
</tr>
<tr>
<td>HindIII</td>
<td>10</td>
<td>&gt;20kb</td>
<td>1.0</td>
</tr>
<tr>
<td>PstI</td>
<td>10</td>
<td>5.5kb</td>
<td>1.0</td>
</tr>
<tr>
<td>PvuII</td>
<td>10</td>
<td>7.7kb</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 8.2. Results obtained by hybridising the RING1 cDNA probe CEM15 to Southern blots of human DNA cut with the enzymes shown.
Table 8.3. Fragment sizes detected on Southern blots of human genomic DNA with the genomic probe 33X1 which encodes the 3' end of the RING1 gene.

8.5. RFLP analysis of RING2

The RING2 probe was the 800bp genomic fragment 31K1 (Figure 5.2.). The results obtained with this probe are shown in Table 8.4. From these data it was concluded that the part of the RING2 locus covered by the probe 31K1 did not contain any high-frequency polymorphisms for the four enzymes tested. It would be informative to repeat this study with a probe covering a larger region of the genome, such as the RING2 cDNA, CEM21.

Table 8.4. Results obtained by hybridising the RING2 genomic probe 31K1 to Southern blots of human genomic DNA cut with the enzymes shown.
8.6. RFLP analysis of \textit{RING3}

The \textit{RING3} probe was the cDNA clone CEM41. The results obtained with this probe are shown in Table 8.5.

<table>
<thead>
<tr>
<th>ENZYME</th>
<th>TESTED</th>
<th>FRAGMENTS</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>BamHI</td>
<td>10</td>
<td>&gt;20kb</td>
<td>1.0</td>
</tr>
<tr>
<td>BglII</td>
<td>8</td>
<td>&gt;20kb</td>
<td>1.0</td>
</tr>
<tr>
<td>DraI</td>
<td>10</td>
<td>5.8kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.7kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.8kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.6kb</td>
<td>1.0</td>
</tr>
<tr>
<td>EcoRI</td>
<td>10</td>
<td>~15kb</td>
<td>1.0</td>
</tr>
<tr>
<td>EcoRV</td>
<td>10</td>
<td>&gt;20kb</td>
<td>1.0</td>
</tr>
<tr>
<td>HindIII</td>
<td>10</td>
<td>5.1kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.0kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.5kb</td>
<td>1.0</td>
</tr>
<tr>
<td>MspI</td>
<td>10</td>
<td>11.0kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.5kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.2kb</td>
<td>1.0</td>
</tr>
<tr>
<td>PstI</td>
<td>10</td>
<td>3.9kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.9kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.3kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0kb</td>
<td>1.0</td>
</tr>
<tr>
<td>PvuII</td>
<td>10</td>
<td>4.5kb</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.3kb</td>
<td>0.1</td>
</tr>
<tr>
<td>TaqI</td>
<td>10</td>
<td>2.6kb</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.0kb</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.6kb</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 8.5. Results obtained by hybridising the \textit{RING3} cDNA probe CEM41 to Southern blots of human genomic DNA cut with the enzymes shown.
In four of the five PvuII-cut genomic DNA samples, a fragment of 4.5kb was detected by the CEM41 probe. In the fifth sample, a fainter band of 4.5kb and an additional band at 4.3kb was observed. This individual was presumably heterozygous for a PvuII polymorphism. In TaqI digests, the CEM41 probe detected a 1.6kb fragment and a 2.0kb fragment in all five samples. One sample had an additional fragment of 2.6kb. This individual was presumably heterozygous for a TaqI polymorphism. In the small sample tested here, the 4.3kb PvuII fragment and the 2.6kb TaqI fragment were each present on one out of ten chromosomes (frequency 0.1). The two polymorphic fragments were not found in the same individual. The RING3 locus did not contain any high-frequency polymorphisms for the other eight enzymes tested.

8.7. RFLP analysis of RING4

The RING4 probe was the cDNA clone 2.1. The results obtained with this probe are shown in Table 8.6. It was concluded that the RING4 locus, as defined by the 2.1 cDNA probe, did not contain any high-frequency polymorphisms for the nine restriction endonucleases tested. However, it has recently been reported that HAM1, the rat homologue of the RING4 gene, is polymorphic (Deverson et al., 1990). Further studies on RING4 using additional enzymes and more DNA samples may reveal variation at this locus.
<table>
<thead>
<tr>
<th>ENZYME</th>
<th>TESTED</th>
<th>FRAGMENTS</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>BamHI</td>
<td>10</td>
<td>~18kb</td>
<td>1.0</td>
</tr>
<tr>
<td>BglII</td>
<td>8</td>
<td>6.0kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.2kb</td>
<td>1.0</td>
</tr>
<tr>
<td>EcoRI</td>
<td>10</td>
<td>11kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.4kb</td>
<td>1.0</td>
</tr>
<tr>
<td>EcoRV</td>
<td>10</td>
<td>&gt;20kb</td>
<td>1.0</td>
</tr>
<tr>
<td>HindIII</td>
<td>10</td>
<td>5.7kb</td>
<td>1.0</td>
</tr>
<tr>
<td>MspI</td>
<td>10</td>
<td>2.1kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.8kb</td>
<td>1.0</td>
</tr>
<tr>
<td>PstI</td>
<td>10</td>
<td>9.7kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.6kb</td>
<td>1.0</td>
</tr>
<tr>
<td>PvuII</td>
<td>10</td>
<td>1.8kb</td>
<td>1.0</td>
</tr>
<tr>
<td>TaqI</td>
<td>10</td>
<td>1.8kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.7kb</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 8.6. Results obtained by hybridising the RING4 probe 2.1 to Southern blots of genomic DNA cut with the enzymes shown.

8.8. RFLP analysis of B51

B51 encodes the putative human homologue of the mouse KE3 gene, as described in Chapter 7. This probe, a 2.6kb BamHI fragment isolated from the cosmid HPB.ALL 51, has been mapped to a position 80-95kb centromeric of RING1 (Figure 7.6.). The results obtained with this probe are shown in Table 8.7.

B51 detected a two-allele polymorphism with BamHI. Three of the five genomic DNA samples tested contained a single hybridising fragment of about 15kb. The two other samples contained the ~15kb fragment and an additional 2.6kb fragment. These were presumably heterozygous for the polymorphism. In the small number of chromosomes tested the frequency of the two alleles was 0.8 (~15kb) and 0.2 (2.6kb). No bands were detected in MspI-digested human DNA, presumably
because the B51 probe does not hybridise to any MspI fragments greater than 1.5kb, the minimum fragment size retained on this blot. No high-frequency polymorphisms were detected with the other eight enzymes tested. Potential applications of the B51 RFLPs are discussed in section 9.3.

<table>
<thead>
<tr>
<th>ENZYME</th>
<th>CHROMOSOMES TESTED</th>
<th>FRAGMENTS</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>BamHI</td>
<td>10</td>
<td>~15kb</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.6kb</td>
<td>0.2</td>
</tr>
<tr>
<td>BglII</td>
<td>16</td>
<td>5.4kb</td>
<td>1.0</td>
</tr>
<tr>
<td>Drai</td>
<td>10</td>
<td>5.8kb</td>
<td>1.0</td>
</tr>
<tr>
<td>EcoRI</td>
<td>18</td>
<td>~15kb</td>
<td>1.0</td>
</tr>
<tr>
<td>EcoRV</td>
<td>10</td>
<td>11kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.9kb</td>
<td>1.0</td>
</tr>
<tr>
<td>HindIII</td>
<td>10</td>
<td>10kb</td>
<td>1.0</td>
</tr>
<tr>
<td>MspI</td>
<td>10</td>
<td>none</td>
<td></td>
</tr>
<tr>
<td>PstI</td>
<td>10</td>
<td>1.9kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.5kb</td>
<td>1.0</td>
</tr>
<tr>
<td>PvuII</td>
<td>10</td>
<td>1.8kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.65kb</td>
<td>1.0</td>
</tr>
<tr>
<td>TaqI</td>
<td>10</td>
<td>1.8kb</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.5kb</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 8.7. Results obtained by hybridising the B51 probe to Southern blots of human genomic DNA cut with the enzymes shown.

8.9. Summary and discussion

RFLPs were identified at the COL11A2, RING3 and B51 loci, and their potential uses are discussed in the following chapter. In contrast, the RING1, RING2 and RING4 probes did not detect polymorphisms with the enzymes tested in the small number of individuals examined. It remains possible that RFLPs could be detected with these enzymes in a
much larger study, although any such polymorphisms would probably be very rare in the population.

A more thorough approach to identifying polymorphisms at or near these loci would involve the use of many more restriction endonucleases and longer probes, such as whole cosmids with repetitive sequences competed out. In addition, it may be advisable to resolve each digest on two different gels run to maximise the resolution of long or short fragments. In practice, however, this traditional approach to searching for polymorphism is very time consuming and is ultimately dependent on the presence or absence of a site for a specific restriction endonuclease, which probably accounts only for a small fraction of the natural variation in the human genome. One alternative strategy to identify polymorphism is to search for dinucleotide repeat sequences which often display high frequency variation in the number of repeat units present at a given locus. The variation in repeat number can be detected by PCR. In a recent study of such loci in mice it was found that 88% were polymorphic between strains (Love et al., 1990).
9. Concluding discussion

The preceding chapters have described the identification and preliminary characterisation of five novel genes, RING1-5, in and around the class II region of the human MHC. In this chapter the general significance of this work in the context of recent progress in MHC mapping will be discussed along with suggestions for future experiments.

9.1. Advances in MHC mapping

When this study was initiated, the molecular map of the MHC region was dominated by the 'classical' class I, class II and class III genes (Figure 1.5.a.). Over the last three years however the powerful techniques of reverse genetics have led to the discovery of numerous novel genes, including those described in this study, within the MHC region (Figure 1.5.b.).

The greatest progress has been made in the class III region. 890kb of the class III region has now been cloned in overlapping cosmids and this interval contains at least 36 genes (section 1.3.4.). Progress in the class II region has been slower, partly because it was thought that it was unlikely that there would be non-class II genes amongst the known class II genes, given that the region has probably evolved through multiple gene duplication events. It might be expected that such a process would not occur if other genes were in the immediate vicinity since these could be disrupted with deleterious consequences. The α- and β-globin gene complexes, both of which are thought to have evolved through gene duplication, are apparently devoid of other genes. However, it is now becoming apparent that the evolutionary history of the class II region has not precluded the presence of unrelated genes. Evidence has been presented in this thesis for two non-class II genes, RING3 (homologous to the Drosophila developmental gene fsh) and RING4 (a member of the ABC family of
transmembrane transporters) at CpG islands between the DNA and DOB genes in the middle of the class II region. Another study of the human class II region has provided evidence for five genes, Y1-5, between DNA and DOB (Spies et al., 1990). From sequence data Y3 is equivalent to RING4, and from mapping data Y4 is probably equivalent to RING3. Two genes related to RING4 have recently been described in the rat class II region (Devenser et al., 1990). Finally, seven novel genes have been identified in the equivalent interval in the mouse class II region (Monaco et al., 1990). Sequence data have revealed that three of these are highly diverged class II genes which were not previously identified by cross-hybridisation with class II gene probes (Dr. John Monaco, personal communication; Monaco et al., 1990).

The region immediately centromeric of the class II region has been characterised in man for the first time in the present study. This region contains a cluster of unrelated genes: a fibrillar collagen gene (COL11A2), a gene encoding protein containing a novel cysteine-histidine motif (RING1), a histidine-rich transmembrane protein gene (RING5), a partially characterised gene (RING2), and sequences homologous to two mouse genes (KE3 and KE5). No class H-like sequences were found in this region, and this cluster of genes most likely defines the centromeric end of the class II region and thus the proximal boundary of the whole human MHC.

In contrast to the detailed map of the centromeric end of the MHC region, much less is known about the class I region. Although considerable progress has been made in characterising and mapping the numerous class I-related genes and pseudogenes, there are very large gaps between these sequences whose genetic content is unknown. There is evidence that clinically important genes remain to be discovered in the class I region. Family studies in pedigrees with inherited forms of idiopathic haemochromatosis, spinocerebellar ataxia and juvenile myoclonic epilepsy have indicated that these diseases may be caused by genes mapping in, or close to, the class I region (Edwards et al., 1986; Spence et al., 1989). A combination of genetic and physical mapping of the class I region has provided evidence for very high recombination frequencies telomeric of the HLA-A gene (section 1.3.2.). Thus, disease genes which have been mapped many cM from HLA-A
in family studies may in fact be in close physical proximity to HLA-A. Recently, YAC clones covering portions of the class I region have been described and these should facilitate the search for novel genes in the class I region using the same approaches which have been successful in the class III and class II regions (Chimini et al., 1990).

It may be many years before the full genetic content of the MHC is known, partly because the potential total number of genes is very large, given the overall size of the MHC, and partly because the reverse genetic approach to identifying novel transcribed regions may miss some genes. Ultimately, the identification of every gene in the MHC may not be possible until the entire region has been sequenced.

The intensive efforts to map, clone and sequence MHC genes have made the MHC one of the best studied regions of the human genome (Stephens et al., 1990). The detailed maps which are available for the MHC region and the impressive progress made towards cloning the entire region make the MHC an obvious starting point for large scale genomic sequencing of the type envisioned by the Human Genome Project. It is interesting in this context to consider the size of the MHC as a proportion of the total genome: taking the haploid size of the human genome to be 3000Mbp, the size of chromosome 6 to be 165Mbp and the size of the MHC to be 4Mbp, the MHC represents about 2.5% of the length of chromosome 6 and about 1/750th of the length of the human genome.

9.2. Potential role of novel genes in class II-associated phenotypes

Regardless of whether the function of a gene is known, some progress can be made in assessing whether that gene is associated with certain diseases by identifying RFLPs in the gene and using these in population association studies as described in section 8.1. However, knowledge of the function of a gene is clearly helpful in formulating hypotheses as to what specific phenotype that gene may influence. In the case of COL11A2, localised to the centromeric end of the class II region in this study, much was already known about the function of the gene.
(Chapter 4). In contrast, nothing was known about the functions of RING1-5. Partial or complete nucleotide sequences of RING1-5 were obtained as a first step towards overcoming this obstacle (Chapter 6). In four cases out of five the predicted amino acid sequence provided clues about the possible function of the gene product and allowed hypotheses to be made about their possible role in MHC-associated diseases.

9.2.1. COL11A2
As discussed in Chapter 4 the characterisation of RFLPs at the COL11A2 locus (section 8.3.) is potentially of interest in ascertaining whether or not alleles of COL11A2 are associated with pauciarticular juvenile rheumatoid arthritis (PJRA). PJRA has previously been shown to display a weak but significant association with the DP region (Odum et al., 1986; Begovich et al., 1989; Fugger et al., 1990) which theoretically may be explained by an involvement of an allele at a closely linked gene which is in linkage disequilibrium with the DP region. The association of COL11A2 RFLPs with RFLPs at the DP loci is being determined to test whether COL11A2 is in linkage disequilibrium with the DP subregion and whether it has a stronger association than DP with PJRA.

9.2.2. RING1
The predicted protein product of the RING1 gene contained a novel cysteine-histidine motif reminiscent of the metal-dependent nucleic acid-binding motifs found in transcription factors (section 6.2.). To test the hypothesis that the RING1 motif is a DNA-binding domain, the N-terminal third of the RING1 gene is being expressed in bacterial cells in order to prepare large quantities for use in DNA binding assays (collaboration with Dr. Ruth Lovering).

Proof that RING1 encodes a transcription factor could be of interest considering the association of the class II region with certain cancers. One of the RING1-related proteins, containing the same cysteine-histidine motif, was identified in a transformation assay (Takahashi and Cooper, 1987). Hodgkin's lymphoma, chronic lymphocytic leukaemia and acute non-lymphocytic leukaemia all show association with certain DP alleles (Bodmer et al., 1989a; Pawelec et al., 1989). Given that RING1 is in close physical proximity to the DP subregion it is
possible that these associations are explained by alleles of \textit{RING1} in linkage disequilibrium with DP.

A possible function for \textit{RING1} in the regulation of gene expression is also of interest given that \textit{RING1} has a mouse homologue which maps within the region known to contain the mouse developmental lethal mutation \textit{t}w\textsubscript{5} (section 7.2.; Artzt et al., 1988). Many of the genes which cause embryonic abnormalities in \textit{Drosophila} are transcription factors (Dressler and Gruss, 1988). In this respect it will be of interest to determine whether expression of \textit{RING1} is developmentally regulated.

\textbf{9.2.3. \textit{RING2}}

The partial nucleotide sequence of \textit{RING2} (section 6.3.) did not allow any conclusions to be drawn about the possible function of this gene. The main point of interest is that a sequence homologous to \textit{RING2} in the mouse maps to the region known to contain the developmental lethal mutation \textit{t}w\textsubscript{5} (section 7.2.; Artzt et al., 1988). This gene is therefore another candidate for \textit{t}w\textsubscript{5}.

\textbf{9.2.4. \textit{RING3}}

The nucleotide and predicted amino acid sequence of the \textit{RING3} gene reveals that it is the human homologue of the \textit{Drosophila} developmental homeotic gene \textit{fsh} (section 6.4.). Genetic experiments are underway in \textit{Drosophila} in an attempt to determine how the \textit{fsh} protein interacts with other known developmental genes (Huang and Dawid, 1990; Dr. Igor Dawid, personal communication). Given the striking conservation of the amino acid sequence between the human and \textit{Drosophila} proteins it is likely that the \textit{RING3} gene product may play a role in mammalian embryonic development, as has been observed for mammalian homologues of other developmentally important \textit{Drosophila} genes. In general these homologues have been studied in the mouse using transgenic techniques to determine their precise developmental and spatial expression patterns. More immediately, it would be of importance to determine by northern blotting whether \textit{RING3} was expressed in embryonic tissues.

Evidence that \textit{RING3} is a developmentally important gene could be of significance in understanding the association of the MHC with habitual
spontaneous abortion. There are numerous reports that women who suffer from habitual abortions have a significantly higher frequency of certain MHC alleles (Christiansen et al., 1989) or show an elevated frequency of shared MHC alleles with their partners (Tiwari and Terasaki, 1985). One interpretation of these findings is that certain MHC haplotypes carry recessive or dominant mutations in a developmentally important locus such as RING3 (or a human equivalent of 1w5) which cause embryonic abnormalities, and consequently abortion. A dominant mutation carried on a maternal haplotype would be detected as an association with the MHC in affected women. A recessive mutation carried on the same haplotype in both partners would be detected as an elevated sharing of haplotypes amongst affected couples.

9.2.5. RING4
The predicted amino acid sequence of the RING4 gene product revealed that it is a member of the ABC superfamily of transmembrane pumps (section 6.5.). This finding is of particular significance given that the RING4 gene maps to the region known to encode a locus required for normal antigen presentation by class I molecules (section 1.6.2.). It has been proposed that a mutant cell line unable to present intracellular antigens by class I molecules has a defect in the mechanism by which peptides were transported from the cytoplasm into the ER (Cerundolo et al., 1990). It has also been hypothesised that peptide transport across the ER membrane could be facilitated by a system like the bacterial oligopeptide permease, which is also an ABC transporter (Townsend and Bodmer, 1989). Clearly the finding of RING4, an ABC transporter gene, in the region which is known to be deleted in the mutant cell line is highly suggestive of a role for the RING4 gene product in peptide transport (Parham, 1990; Trowsdale et al., 1990). In another mutant cell line with a similar defect in class I antigen presentation which is thought to be caused by a point mutation, transcription of RING4 could not be detected (Spies et al., 1990). The RING4 gene is also being transfected into mutant cell lines to attempt to complement the defect in class I antigen presentation (collaboration with Dr. Alain Townsend).
Proof that the \textit{RING4} gene product is a transporter which delivers cytoplasmic peptide antigens to class I molecules in the lumen of the ER may be of significance in understanding the association of the class II region with autoimmune diseases. It is conceivable that a peptide transporter could exert selectivity on the type of antigen delivered to the lumen of the ER by preferentially transporting peptides of a certain length or with certain sequence characteristics (Elliott et al., 1990). It could be envisaged that polymorphic forms of the transporter might result in a different selection of peptides, perhaps including critical autoantigenic epitopes from self proteins, being made available for binding to and presentation by class I molecules. In this respect it is interesting that polymorphisms, both at the functional level and the RFLP level, have been identified at the locus equivalent to \textit{RING4} in the rat (Livingstone et al., 1989; Deverson et al., 1990) although the human \textit{RING4} locus was not polymorphic with the restriction enzymes tested (section 8.7.). It will be of great interest to determine whether \textit{RING4} is polymorphic with enzymes other than those tested in the present study, as the association of \textit{RING4} RFLPs with susceptibility to autoimmune disease can then be investigated.

\textbf{9.2.6. \textit{RING5}}

The partial nucleotide and predicted amino acid sequence of the \textit{RING5} gene revealed that it is homologous to the mouse \textit{KE4} gene (section 6.6.). The mouse gene has been completely sequenced and the predicted protein product had three potential transmembrane regions and two domains in which histidine was typically found every second residue (St.-Jacques et al., 1990). Although the \textit{KE4} gene product does not have any homology with other proteins which could help shed light on its function, it is potentially interesting because (like the mouse homologues of \textit{RING1} and \textit{RING2}, see above) it maps to the region known to contain the mouse developmental lethal gene \textit{t\textsuperscript{w5}} (Artzt et al., 1988). Furthermore it is expressed in early embryogenesis, the time at which the \textit{t\textsuperscript{w5}} mutation is known to act. Since this mutation has been characterised in the mouse, further work on the \textit{KE4/RING5} locus would probably be most productive in this system. The construction of transgenic mice homozygous for mutant forms of \textit{KE4} could reveal whether \textit{KE4} was the \textit{t\textsuperscript{w5}} gene.
9.3. Additional applications of probes generated in this study

The probes generated during the work described in this thesis provide additional accurately mapped markers for the class II region and the area immediately centromeric of the MHC. These probes have so far been used to characterise X-irradiated fusion hybrids carrying fragments of chromosome 6 (Ragoussis et al., 1991a) and yeast artificial chromosome clones covering the class II region (Ragoussis et al., 1991b).

The RFLPs detected with the B51 probe (section 8.8.) are of interest because B51 is a novel marker mapping over 200kb centromeric of DPB1, the most proximal class II gene in which polymorphism has previously been studied in detail. The association of B51 RFLPs with pauciarticular juvenile arthritis and Hodgkin's lymphoma could be tested to determine whether B51 is more closely associated with these diseases than the DP subregion. This could provide evidence for new disease susceptibility factors mapping centromeric of the class II region.

The RFLPs characterised at the COL11A2 locus may be useful for assessing the role of this gene in a heterogeneous group of joint disorders described as osteoarthritis or chondrodysplasia, which are characterised by progressive degeneration or abnormal development of articular cartilage. In families where the disease shows a clear pattern of inheritance, linkage analyses can be performed using polymorphic probes for candidate genes, such as cartilage collagens. To date, the COL2A1 locus has been investigated most intensively because type II collagen is quantitatively the most important in cartilage, and in one pedigree it has been shown that RFLPs at the COL2A1 locus segregated with the inheritance of osteoarthritis with mild chondrodysplasia (Knowlton et al., 1990). However, in other pedigrees the same locus was excluded from being closely linked to the inherited defect (Knowlton et al., 1990; Sykes et al., 1989), which raises the question of an involvement of more minor cartilage components.
9.4. More new genes in the class II region?

To date the only interval within the class II region which has been intensively studied for the presence of novel genes is that between DNA and DOB, or its equivalent in rodents (this study; Spies et al., 1990; Deverson et al., 1990; Monaco et al., 1990). In the present study, the strategy taken to isolate novel genes was based on identifying CpG islands, as defined by the presence of clusters of sites for rare-cutter enzymes, in the class II region. This approach proved highly successful, with each cluster of rare-cutter sites being associated with at least one gene, and confirms that CpG islands are highly diagnostic for the presence of genes. Cluster 4, between DQB3 and DQB1 (Figure 3.5.), may therefore mark the position of another novel gene. However, there are additional genes in the class II region which are not associated with CpG islands (Spies et al., 1990). Other as yet uncharacterised intervals in the class II region may also contain additional genes even though they do not necessarily contain CpG islands.

It is interesting that previously undiscovered highly diverged class II-related genes, including an α/β gene pair, have been found in the mouse class II region (Monaco et al., 1990; Dr. John Monaco, personal communication). Based on the striking homology between the human and mouse class II regions (Chapter 7) it would be predicted that homologues of these genes should also be found in the human class II region. Indeed, probes from two of the mouse genes have been used in this laboratory to identify a homologous α/β gene pair in the human class II region (Adrian Kelly, personal communication). The product of this gene pair could be the novel class II molecule identified by Carra and Accolla (1987; section 1.6.1.), and may have implications for the understanding of the association of autoimmune diseases with the class II region (section 1.5.).

There is still no confirmation of the identification of the genes which encode the subunits of the LMP complex and which map between Ob and Pb in the mouse class II region (section 1.6.3.), although these may be among the novel genes described by Monaco et al. (1990). If the LMP complex is the same as the high molecular weight proteinase complex, as discussed in section 1.6.3., it may be possible to identify the imp
genes on the basis of recently published protein sequence data from proteinase subunits. N-terminal amino acid sequences have been determined for five human proteinase subunits (Lee et al., 1990a), and these could be used to design oligonucleotide probes which in turn could be used to map the genes encoding the subunits.

It is notable that the mouse and rat class II regions contain two genes, HAM1 and HAM2, about 15kb apart, which are related to the human RING4 gene (Monaco et al., 1990; Deverson et al., 1990). HAM1 and HAM2 share 77% identity at the amino acid level, with HAM1 most closely related to RING4. RING4 probes have been hybridised at reduced stringency to cosmid and YAC clones in the vicinity of the human RING4 gene and weakly cross-hybridising fragments have been detected in some experiments (Dr. John Trowsdale, personal communication). It remains to be determined whether humans, like rodents, have a second, diverged, RING4-like gene.

9.5. Function of the MHC gene cluster

The MHC probably provides an excellent model for a typical region of the mammalian genome, containing a high density of compact genes with a variety of different functions. Constitutively expressed CpG island-associated 'housekeeping' genes are interspersed with tissue-specific genes. Some genes are clearly related to one another by duplication events while others are unique. A number of the genes in the MHC region, such as those for steroid 21-hydroxylase and the α2 chain of type XI collagen, are functionally unrelated to the classical MHC genes. Despite this apparent diversity and complexity it is interesting to note that some of the recently discovered genes in the MHC region may have functions related to those of the classical MHC genes in the mediation of the immune response. It is well established that the classical class I and class II gene products function in shaping the T-cell repertoire and in antigen presentation, while the complement gene products play a key effector role in the humoral immune response. Now there is good evidence that the MHC also encodes the putative peptide transporter which translocates antigens from the cytoplasm into the ER (RING4). The MHC may also encode
subunits of a proteolytic complex which could generate peptides from intracellular antigens (Parham, 1990). The hsp70 gene products may function in the binding and chaperoning of peptides in the cytoplasm and the ER (Flynn et al., 1989; Vanbuskirk et al., 1989). The products of the Tnf genes can modulate the expression of MHC molecules (Collins et al., 1986; Pujol-Borrell et al., 1987). There is also evidence that the class III region encodes a locus which regulates the sensitivity of cells to lysis by natural killer cells (Rembecki et al., 1988).

Another intriguing feature of these genes is that almost all are polymorphic, albeit to varying degrees. The highest level of polymorphism is observed in the products of the classical class I and class II genes (Parham et al., 1988; Marsh and Bodmer, 1989). The C4 gene products are also unusually polymorphic, though less so than those of the class I and class II genes (Carroll and Alper, 1987). Lower levels of polymorphism have been described for the C2 and Bf genes (Campbell, 1987), the rat HAM1 gene (homologous to RING4; Deverson et al., 1990), the TNFB gene (Messer et al., 1991), the hsp70 gene (Caplen et al., 1990), the NK-regulatory gene (Rembecki et al., 1988) and the MHC-encoded Imp genes (Monaco and McDevitt, 1986). It is tempting to speculate that the products of specific alleles at these loci function more efficiently together and that the physical clustering of these genes has been favoured by natural selection because this tends to keep advantageous allelic forms together at meiosis (Parham, 1990; Bodmer et al., 1986). Although in most cases it is not yet known whether the polymorphism observed is of functional significance, evidence to support the hypothesis that certain alleles of MHC gene products may interact favourably has been obtained from studies of antigen presentation in the rat. Naturally occurring allelic forms of the cim locus (which may be the rat peptide transporter) in the class II region affect the specificity and efficiency of antigen presentation by class I molecules (Livingstone et al., 1989; Deverson et al., 1990). The idea that the clustering of these genes may have functional consequences attaches even greater significance to the study of the MHC.
10. References


Bohme J, Andersson M, Andersson G, Moller E, Peterson PA and Rask L (1985). HLA-DR β genes vary in number between different DR
specificities, whereas the number of DQ β genes is constant. J. Immunol. 135: 2149-2155.


Morris NP and Bachinger HP (1987). Type XI collagen is a heterotrimer with the composition (1α, 2α, 3α) retaining non-triple-helical domains. *J. Biol. Chem.* 262: 11345-11350.


Young JAT and Trowsdale J (1990). The HLA-DNA (DZA) gene is correctly expressed as a 1.1kb mature mRNA transcript. *Immunogenetics* 31: 386-388.


Appendices

A. Cosmid clones

A.1. Cosmid vector
The cosmid library was constructed using genomic DNA from the human T-cell line HPB.ALL in the cosmid vector cos202 (Figure A.1.). The genomic DNA was partially cleaved with Sau3A and ligated into the BglII cloning site of the vector.

A.2. Cosmid clones
The orientation of the genomic inserts of the cosmid clones isolated in this study is summarised in Table A.1., which shows which end of the vector (the 2.2kb EcoRI fragment or the 0.2kb EcoRI fragment, Figure A.1.) is adjacent to the proximal (centromeric) end of the genomic insert.

<table>
<thead>
<tr>
<th>COSMID</th>
<th>FIGURE</th>
<th>PROXIMAL END</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPB.ALL 1</td>
<td>5.2</td>
<td>0.2kb</td>
</tr>
<tr>
<td>HPB.ALL 8</td>
<td>5.2</td>
<td>0.2kb</td>
</tr>
<tr>
<td>HPB.ALL 25</td>
<td>5.2</td>
<td>2.2kb</td>
</tr>
<tr>
<td>HPB.ALL 31</td>
<td>5.2</td>
<td>2.2kb</td>
</tr>
<tr>
<td>HPB.ALL 33</td>
<td>5.2</td>
<td>2.2kb</td>
</tr>
<tr>
<td>HPB.ALL 42</td>
<td>5.2</td>
<td>2.2kb</td>
</tr>
<tr>
<td>HPB.ALL 71</td>
<td>5.6</td>
<td>2.2kb</td>
</tr>
<tr>
<td>HPB.ALL 51</td>
<td>7.4</td>
<td>2.2kb</td>
</tr>
</tbody>
</table>

Table A.1. Orientation of cosmid clones isolated in this study. The second column shows the Figure in which a map of the genomic insert of the cosmid is presented. The third column shows which end of the vector (Figure A.1.) is adjacent to the proximal end of the genomic insert.
Figure A.1.

Restriction map of the cosmid vector cos202.
The CDM8 vector carries:
- supF
- origins of replication for M13, Py, piVX and SV40
- CMV and T7 promoters
- splice site and poly(A) tail

Figure B.1.
Restriction map of the cDNA vector CDM8.
B. cDNA clones

B.1. cDNA vector
All the cDNA clones described in this study were isolated from libraries constructed according to the method of Seed (1987) in the plasmid vector CDM8, or derivatives thereof. The structure of CDM8 is shown in Figure B.1. For cDNA cloning, the 4.8kb vector is cleaved with BstXI to remove the 400bp stuffer fragment and then ligated to cDNA which has been tailed at each end with complementary linkers. The BstXI sites are destroyed in this process, but the cDNA insert can be excised using other enzymes with sites either side of the cloning site (Figure B.1.).

Sequencing of the cDNA clones isolated in this study was initiated using two primers, CDM8-FORWARD (CDM8-F) and CDM8-BACKWARD (CDM8-B) which are complementary to the vector sequence either side of the cloning site as shown in Figure B.1. The sequences of these two primers are:

CDM8-F (ICRF No. 10299) 5'-CCCAAGCTTCTAGAGATCCCT-S'
CDM8-B (ICRF No. 5841) 5'-AGGCGCAGAACTGGTAG-3'

B.2. cDNA clones
The orientation of each cDNA insert within the vector is summarised in Table B.1. which shows which of the two primers, CDM8-F or CDM8-B, gives the sequence at the 5' end of a particular insert.
<table>
<thead>
<tr>
<th>GENE</th>
<th>cDNA</th>
<th>5' PRIMER</th>
</tr>
</thead>
<tbody>
<tr>
<td>RING1</td>
<td>CEM15</td>
<td>ND</td>
</tr>
<tr>
<td>RING2</td>
<td>CEM21</td>
<td>CDM8-B</td>
</tr>
<tr>
<td>RING3</td>
<td>CEM32</td>
<td>CDM8-B</td>
</tr>
<tr>
<td>&quot;</td>
<td>CEM35</td>
<td>CDM8-F</td>
</tr>
<tr>
<td>&quot;</td>
<td>CEM41</td>
<td>CDM8-B</td>
</tr>
<tr>
<td>&quot;</td>
<td>CEM44</td>
<td>CDM8-B</td>
</tr>
<tr>
<td>RING4</td>
<td>2.1</td>
<td>ND</td>
</tr>
<tr>
<td>RING5</td>
<td>γU5</td>
<td>CDM8-F</td>
</tr>
</tbody>
</table>

Table B.1. Orientation of cDNA clones isolated in this study. The third column shows which primer, CDM8-F or CDM8-B, gives the sequence at the 5' end of the cDNA insert. ND, not determined: both CEM15 and 2.1 were subcloned into other vectors before sequencing.

C. Gene symbols and database accession numbers

HGM nomenclature for the new genes RING1-5:

RING1    D6S111E  
RING2    D6S112E  
RING3    D6S113E  
RING4    D6S114E  
RING5    D6S115E  

GenBank accession number for the nucleotide sequence of RING5 (shown in Figure 6.12.): M58660.