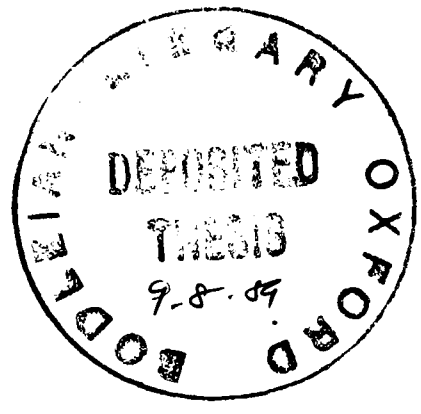
MOLECULAR MAPPING OF THE HLA CLASS III REGION



Carole A. Sargent

St. Peter's College, Oxford

A thesis submitted for the

Degree of Doctor of Philosophy

October 1988

ɜtry Unit

ɔchemistry,

Oxford

CONTACT ADDRESS

Miss. Carole A. Sargent,
Department of Pathology,
University of Cambridge,
Tennis Court Road,
CAMBRIDGE. CB2 1QP

ACKNOWLEDGEMENTS

There are many people who, over the past three years, have been willing to give their time and friendship to advise and encourage me during my research, and to help me enjoy my stay in Oxford. In particular, thanks go to my supervisor, Duncan Campbell, who has provided unfailing support during both the non-productive phases and more rewarding phases of two projects, at undergraduate and graduate levels. In addition, I should like to mention Sandra Smith for her efficient organisation of the laboratory, and all members of the MRC Immunochemistry Unit for their companionship. Unfortunately, there is not space to acknowledge everyone here, but I hope that I have remembered to include all who have aided with the practical aspects of the research in the appropriate places.

I am also grateful to those who have helped in the preparation of this thesis; Carolyn Brooks for her patience in teaching me how to use the word processor, Ken Johnson for the photographic work, and the "volunteers" who have had the task of proof reading, especially Robert Lunn and Wendy Thomson. I hope that my attempts at writing did not prove too much of a strain!

Last, but not least, my thanks are due to the Medical Research Council, which has provided funding for my graduate course in the form of a studentship.

Abbreviations

The nomenclature of complement components is that recommended by the World Health Organization (1968, 1981). Activated components are indicated by bar, e.g. $C\bar{I}$. The one and three letter codes for amino acids are as recommended by the IUPAC - IUB Commission on the Biochemical nomenclature (1969). Restriction enzymes are referred to by the three letter nomenclature of Smith & Nathans (1973). The remaining abbreviations are listed in alphabetical order below, or defined in the text when they are first used.

A_{280}, A_{600}, etc.	- Absorbance at 280, 600 nm etc.
APS	- Ammonium persulphate
BCIG	- Indoyl-β-D-galactoside
BSA	- Bovine serum albumin
C	- Complement
cDNA	- Complementary DNA
DNA	- Deoxyribonucleic acid
DEAE	- Diethylaminoethyl
DTT	- Dithiothreitol
EDTA	- Ethylendiaminetetraacetic acid
HLA	- Human leucocyte antigens
Hepes	- N-2-hydroxyethylpiperazine-N'-2-ethanesulphonic acid
IPTG	- Isopropyl thiogalactoside
MHC	- Major Histocompatibility Complex
mRNA	- Messenger RNA
NaAc	- Sodium acetate
dNTP	- Deoxyribonucleoside triphosphate

	(A, C, G or T specified instead of N where appropriate)
ddNTP	- Dideoxynucleoside triphosphate (N = A, G, C or T)
OD	- Optical density
21OH	- 21-Hydroxylase
PBS	- Phosphate buffered saline
rATP	- Adenosine triphosphate
RNA	- Ribonucleic acid
SDS	- Sodium dodecylsulphate
SLE	- Systemic lupus erythematosus
TEMED	- N, N, N', N' - tetramethylethylene-diamine
Tris	- Tris (hydroxymethyl) aminoethane
U	- Units

ABSTRACT

Molecular Mapping of the HLA Class III Region.

Carole A. Sargent
Submitted for the degree of D. Phil.

St.Peter's College
Michaelmas Term, 1988

(a) The class III region of the human MHC is approximately 1.1 Mbp in size. Almost 50% has been cloned in two overlapping cosmid clusters by chromosome walking from the complement/ 21OH gene cluster and from the TNF α locus. The complement/ 21OH cosmid cluster consists of 45 independent recombinants, spaning 390 kb. The TNFα and TNF β cosmid cluster consists of 16 independent recombinants, spanning 150 kb. Each genomic insert has been characterised by restriction enzyme digestion and Southern blot analysis to produce a detailed restriction map of the cloned region.

(b) Single copy DNA sequences isolated from the cosmid clones have been hybridised to Southern blots of genomic DNA digests separated by pulsed-field gel electrophoresis (PFGE). This analysis has established the orientation and precise position of the complement/ 21OH gene cluster within the class III region relative to the class I and class II genes. Furthermore, probes from the TNF cosmid cluster have been used to show that the TNF genes lie ∿250 kb centromeric to the HLA-B locus. The gene order from the centromere is HLA-DR-21OHB-C2-TNF α-TNF β-HLA-B, with HLA-DR separated from 21OHB by 300-350 kb, and C2 separated from HLA-B by 650 kb.

(c) Genomic probes have been used in conjunction with PFGE to define the positions of HTF islands within the cloned portion of the class III region. Single copy sequences associated with these islands have been hybridised to Southern blots of DNA from a variety of animal species to look for cross-hybridisation indicative of phylogenetic conservation of coding sequences. These probes have been used to screen Northern blots and define island related transcripts. The products of 15 genes have been mapped within the cloned region. Of these, two have been identified as duplicated members of the HSP 70 family. The remaining loci appear to be single copy sequences. One (G11) has been cloned and characterised.

C O N T E N T S

Page No.

CHAPTER I

INTRODUCTION

1.1 GENERAL INTRODUCTION

The major histocompatibility complex (MHC) consists of clusters of polymorphic loci which belong to multigene families involved in immune regulation. In all vertebrate species studied so far, the genes appear to remain linked on one chromosome (Gotze, 1977; Klein and Figueroa, 1986). The MHC has been most extensively investigated in man and mouse (Fig. 1.1). The human MHC, or HLA (for human leukocyte antigen) has been mapped to the short arm of chromosome 6 (Lamm et al., 1974; Franke and Pellegrino, 1977) in the distal portion of the 6p21.3 band (Morton et al., 1984; Lamm and Olaisen, 1985). The murine MHC, termed the H-2 complex, is located on chromosome 17 (Klein, 1975). In both species, the MHC gene products can be divided into three separate classes, depending upon their structures and functions. The class I and class II antigens are heterodimers which form integral membrane glycoproteins (Klein et al., 1983; Hood et al., 1983; Steimetz and Hood, 1983; Kaufman et al., 1984; Strominger, 1987; Duquesnoy and Trucco, 1988). They are important for the recognition of foreign antigens during cell mediated immunity, such as cell lysis by cytotoxic T lymphocytes (CTLs) and the antibody response. In man and mouse, the loci encoding the class I and class II specificities flank the class III region (Hood et al., 1983). The class III genes are more heterogeneous, encoding complement components C2, factor B and C4 (Campbell et al., 1986, 1988; Chaplin, 1985), cytochrome P450 steroid 21-hydroxylase (21OH) (Carroll et al., 1985b; Chaplin, 1985), the cytokines tumour necrosis factors (TNF) α and β (Carroll et al., 1987; Dunham et al., 1987; Inoko and Trowsdale, 1987; Ragoussis et

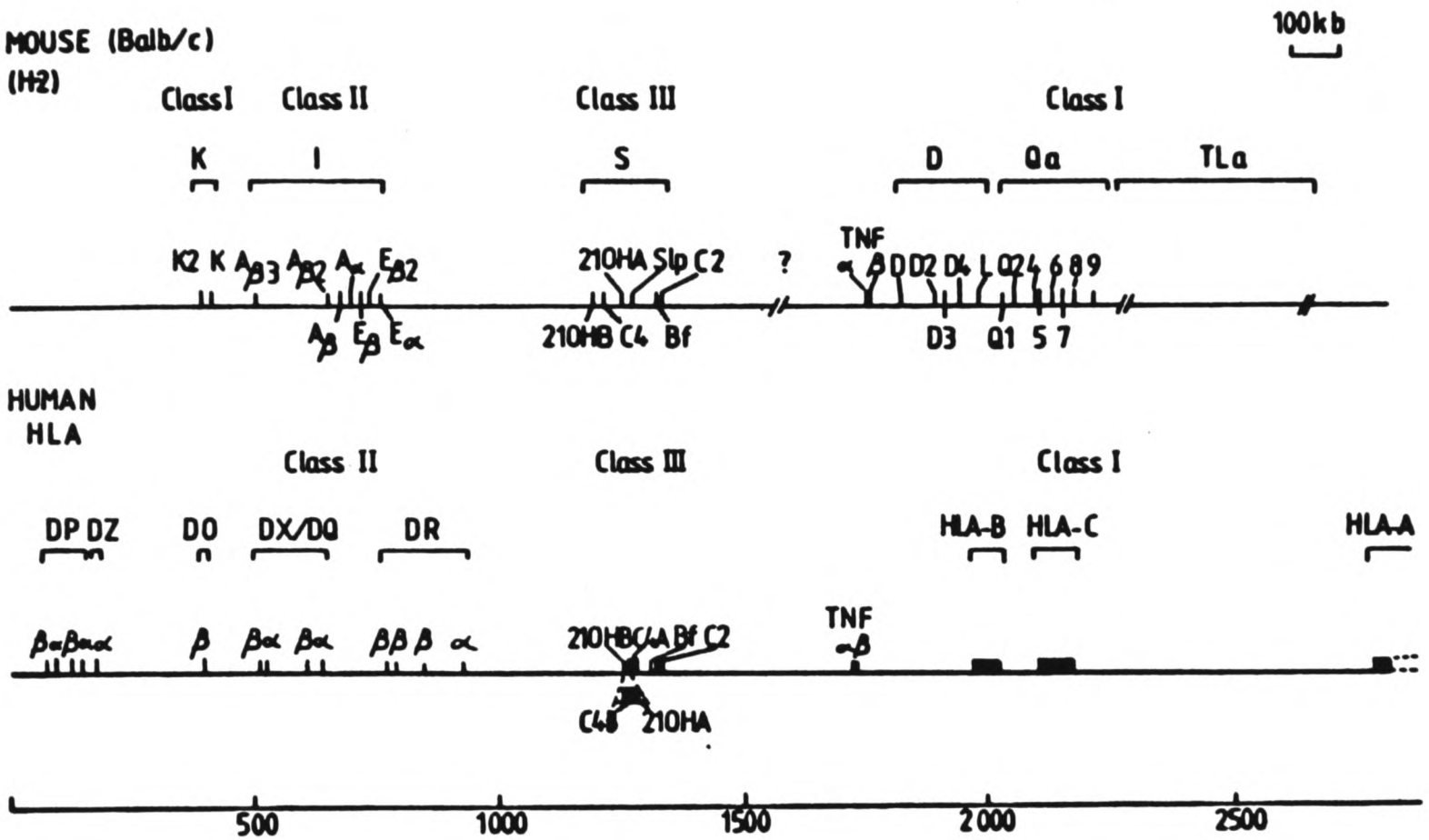


Fig. 1.1

A comparison of the major histocompatibility complexes from man and mouse. The centromere is towards the left. (From Campbell et al., 1988)

al., 1988; Muller et al., 1987b), and two molecules of unknown function, B144 and RD (Tsuge et al., 1987; Levi-Strauss et al., 1988).

The class I antigens were originally identified as the determinants involved in graft rejection (Gorer, 1937), and are the classical transplantation antigens. Each class I molecule consists of a polymorphic heavy (α) chain of relative molecular weight (M_r) 43,000 daltons (43 kD) non-covalently associated with a light chain of M_r 12 kD. The light chain is $\beta_2$microglobulin (β_2m), a non-polymorphic peptide encoded outside the MHC (chromosome 15 in man; chromosome 2 in mouse). Similarly, the class II antigens were first defined as elements governing the level of the immune response to given peptide immunogens in guinea pigs and mice (Katz and Benacerraf, 1976; Benacerraf, 1981). Like class I antigens, they consist of a heavy (α) chain (M_r, 33 kD), which is relatively non-polymorphic, and a light (β) chain (M_r, 28 kD) which is far more variable. Both chains are encoded within the MHC.

The complement components of the class III region are serum glycoproteins. C2 and factor B are serine proteases of unusual structure which perform analogous roles in the classical and alternative pathways of complement activation (Reid and Porter, 1981; Reid, 1986; Campbell et al., 1988). C4 is an early member of the classical pathway, which interacts with C2 to form a C3 convertase (Reid and Porter, 1981; Reid, 1986; Campbell et al., 1988). In contrast, 21OH is a microsomal enzyme, which has an important role in the biosynthesis of mineralocorticoids and glucocorticoids in the adrenal gland (Nerbert and Gonzalez, 1987). TNF α and β are polypeptides, secreted by macrophage and T lymphocytes respectively, which are cytotoxic to tumour cells (Old, 1985; Sugarman et al., 1985; Beutler and Cerami, 1986). In addition, TNF α is a mediator of inflammatory reactions (Beutler and Cerami, 1986; Old, 1987).

Characterisation of the MHC at a molecular level is of interest for several reasons. Firstly, it represents ∿1/750 of the human genome, and the functions of many of its products are well understood. Secondly, there are a number of diseases of both an autoimmune and a non-immune aetiology where a genetic predisposition can be mapped to the HLA region (for reviews, see Batchelor and McMichael, 1987; Tiwari and Terasaki, 1985). In some cases, the disease can be directly associated with the products of the class I, class II or class III regions (Glass et al., 1976; Fielder et al., 1983; Batchelor and McMichael, 1987; Todd et al., 1988a, 1988b). In others, the correlation may be the result of linkage disequilibrium between the HLA marker and another, as yet undefined gene within the MHC (Bodmer and Bodmer, 1978; Batchelor and McMichael, 1987).

The products of the known loci display extensive polymorphism at the protein level (Steinmetz and Hood, 1983; Hood et al., 1983; Klein et al. 1983, Kaufman et al., 1984, Strominger, 1987; Festenstein and Ollier, 1987). Serologically defined determinants have been used to specify the class I loci, A, B and C, at least three class II αβ heterodimers called DR, DP and DQ, and two C4 loci. Each locus encodes a range of alleles of which the number of possible combinations is far greater than that observed in the population as a whole. Linkage disequilibrium is defined as the combination of specific allotypes at linked loci on the same chromosome at a frequency greater than that expected from the individual frequencies. Haplotypes were originally termed to describe the association between the D region and the HLA-B alleles (Bodmer and Bodmer, 1978). Since then, linkage disequilibrium has been shown to include combinations of the complement proteins, leading to the idea of the extended haplotype (Alper et al., 1986). For a given haplotype including DR, C2, Bf, C4 and HLA-B specificities, the possible alleles at other linked loci are also limited.

The distance between HLA-A and HLA-DP in man has been estimated as 3-4 centimorgans (cM) from studies of recombination within families. Of this, ∿1 cM lies between HLA-B and HLA-DR (Olaisen et al., 1987). If the class III region is defined as all the genes between class I and class II, then there may be many which have yet to be characterised. This has been highlighted by the recent mapping of four new genes to the class III region; firstly the confirmation of the location of the TNF loci between HLA-B and the complement cluster (Dunham et al., 1987; Carroll et al., 1987; Inoko and Trowsdale, 1987; Ragoussis et al., 1988), and secondly the identification of the novel genes B144 and RD (Tsuge et al., 1987; Levi-Strauss et al., 1988).

In order to characterise a region of this size in detail, two techniques are especially useful. Firstly, chromosome walking allows the construction of a molecular map of the class III region. The cloned DNA can be analysed for the distribution of unique sequences, followed by the analysis of single copy probes to define which nucleotide sequences are conserved between species. This represents one approach for the identification of potential genes (Monaco et al., 1986). Secondly, pulsed-field gel electrophoresis (PFGE) (Schwartz and Cantor, 1984; Carle and Olson, 1984; Anand, 1986) allows the separation of fragments up to 8,000 kb in size, and can be used to produce long-range restriction maps based on the hybridisation of probes to common fragments on Southern blots (Brown and Bird, 1986; Lindsay and Bird, 1987). Used together, the two techniques are especially powerful for elucidating the overall organisation of a genomic region, as has been shown for the class II genes (Hardy et al., 1986).

The infrequently cutting enzymes used in PFGE are also useful for identifying HTF islands. These are CpG rich, unmethylated stretches of DNA which are often found at the transcriptional start sites of genes

(Bird, 1986; Gardiner-Garden and Frommer, 1987). Comparison with the distribution of the same enzyme sites in cloned DNA can be used to define which are cleaved at the chromosomal level. Once the position of a potential island structure has been found, flanking probes may then be used to search for transcripts by Northern blot hybridisation. By identifying novel gene transcripts, and characterising potential products of these loci, it is hoped that molecular mapping of the class III region will aid in the understanding of the nature of HLA association and disease.

1.2 THE CLASS I REGION

The products of three loci, termed HLA-A, -B and -C have been identified in man on the basis of common epitopes recognised by cross-species antisera. Each locus displays extensive polymorphism at a protein level. Serological testing can be used to group alleles according to public antigenic determinants (shared epitopes which are cross-reactive with a given antiserum) and to split them according to their private antigenic determinants, which define the HLA specificities. A total of 24 HLA-A, 52 HLA-B, and 11 HLA-C specificities have been recognised by the World Health Organisation (Bodmer, et al., 1988).

The class I antigens are expressed on the surface of virtually all nucleated somatic cells. In mice, in addition to the classical transplantation antigens K, D and L, there are a number of class I related proteins encoded by the H-2 associated Qa 2,3 and Tla regions. These are less polymorphic, and exhibit a limited tissue distribution pattern (Flaherty, 1981). Human activated T lymphocytes have also been demonstrated to express unique class I determinants which are not

encoded by the HLA-A, -B or -C loci, and may be equivalent to these murine products (Paul et al., 1987; Tada et al., 1978; van Leeuwen et al., 1980; Pontarotti et al., 1986). In addition, a class I antigen that is preferentially expressed on resting T cells has been described by Koller et al. (1988). The locus encoding this molecule has been mapped to a position between the HLA-C and HLA-A loci, and is designated HLA-E.

The class I heavy chains have a characteristic domain structure (Fig. 1.2) (Pleogh et al., 1981; Duquesnoy and Trucco, 1988). The extracellular portion can be divided into three regions; α_1 and α_2, of 90 amino acids each, and α_3, which is similar to the constant domain of immunoglobulins. The α_2 domain contains an intrachain disulphide bond, and both α_1 and α_2 are glycosylated. Antigenic determinants reside primarily within the α_1 and α_2 domains (Stroynowski et al., 1985; Parham et al., 1988; Srivastava et al., 1987). Here, the polymorphic sites are found to lie within variable and hypervariable regions, the most polymorphic being the segment between residues 62 to 80. The extracellular region is followed by a hydrophobic transmembrane domain, and a short hydrophilic cytoplasmic tail, which contains phosphorylation sites. Alternative splicing of murine class I mRNA species has also been found to produce serum polypeptides which lack the cytoplasmic domain (Handy et al., 1988).

1.2.1 3-D Structure

Computer modelling has been used to predict the secondary structure of the class I heavy chain, and confirms the immunoglobulin-like arrangement of the α_3 domain (Novotry and Auffray, 1984; Vega et al., 1984). The α_2 and α_1 domains appeared to be comprised primarily from β-sheet, but a short region of α-helix (positions 146-150) within the α_2

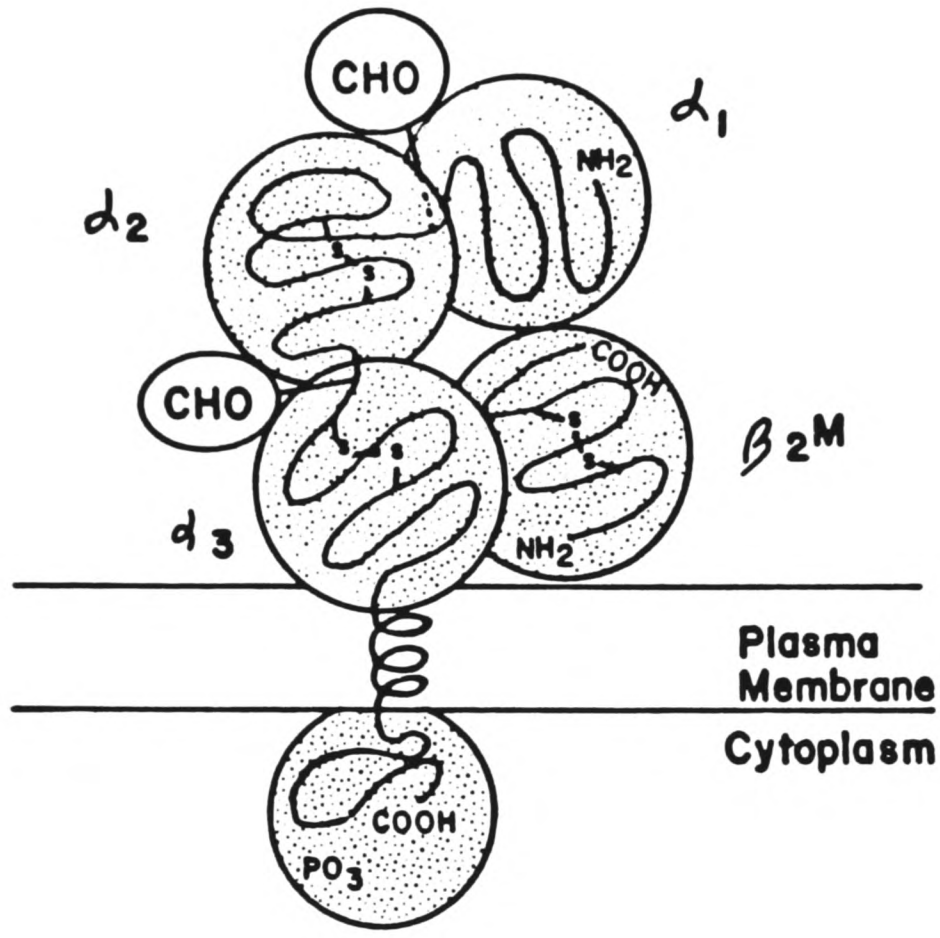


Fig. 1.2

Basic domain strucure of an HLA class I molecule, showing the positions of carbohydrate attachment on the external domains. (From Duquesnoy and Trucco, 1988)

domain may be important for interaction with cytotoxic T cells (CTLs), as substitutions in this region have been shown to decrease the efficiency of cell lysis in CTL mediated pathways (Krangel et al., 1983).

X-ray diffraction studies of a soluble class I molecule (HLA-A2) consisting of the α_1, α_2, α_3 domains and β_2m have been used to analyse the three dimensional structure. It has a twofold rotational axis of symmetry, with the two immunoglobulin-like domains α_3 and β_2m paired together, and the α_1 and α_2 domains paired to form the top of the molecule. The α_1 and α_2 domains have similar folding patterns: the N-terminal regions consist of 3 β-strands forming a single anti-parallel sheet, whereas the C-terminal regions are α-helical. The sheet forms the bottom of a groove at the surface of the molecule, the walls of which consist of the side chains of the residues in the α helices (Bjorkman et al., 1987a).

The cleft is believed to form the antigen binding site of the class I molecule. Studies of the variable residues in class I alleles have defined regions within the α chain domains that could be involved in interactions with peptides and T cell receptors (Bjorkman et al., 1987b; Parham et al.,1988). The α helix of the α_1 domain has a cluster of 11 residues between positions 62 to 80 which show high diversity. The side chains of these amino acids point into the cleft, or up from the floor of the groove, and are probably involved in peptide binding. Variable residues in the α helix of the α_2 domain have side chains which point upwards from the surface of the molecule, and may interact with T cell receptors. A higher level of diversity in the β strand of the α_2 domain than the α_1 domain leads to a hypothesis of peptide binding in which the residues of the β strands of the α_1 domain interact with the conserved

peptide backbone, and the β strands of the α_2 domain interact with the amino acid side chains.

1.2.2 Biosynthesis

The production of the α chain is controlled at both transcriptional and post-transcriptional levels. The biosynthetic pathway includes the association with β_2m, glycosylation and insertion into the endoplasmic reticulum (ER) (Pleogh et al., 1981; Cresswell, 1987; Duquesnoy and Trucco, 1988). Association of the heavy chain with β_2m is essential for expression: in Daudi cells (a Burkitts lymphoma cell line), which are deficient in β_2m, and specially selected human and murine class I mutants, inability to bind β_2m has been demonstrated to correlate with the loss of cell surface expression, although intracellular levels of the free heavy chain remain normal (Zeff et al., 1986; Krangel et al., 1982). Once inserted into the ER, the signal peptide is cleaved from the molecule, and it becomes glycosylated on the a_1 and a_2 domains. The addition of carbohydrate does not appear to be essential, as treatment of cells with tunicamycin (an inhibitor of glycosylation) does not reduce class I expression. A slow or fast processing rate correlates with a low or high level of cell surface molecules. In the mouse, the N-terminal sequences of the class I α chains have been shown to govern the rates at which intracellular processing and transport are achieved (Beck et al., 1986).

The regulation of expression is also governed at the level of transcription by interferons α, β, and γ . An interferon response element (IRE) occurs in association with a functional enhancer in the promoter of the heavy chain gene (Israel et al., 1986). Interferon is produced in vivo in response to viral infection, and the effect on class

I gene transcription may be important in relation to MHC restricted antigen presentation to virus specific cytotoxic T cells.

1.2.3 Function

Class I antigens are the restrictive elements for the cytotoxic T lymphocytes (CTLs) of the immune system (Zinkernagel and Doherty, 1974, 1979). In particular, CTLs are responsible for the destruction of virally infected and neoplastic cells. The T lymphocyte responses do not rely upon the expression of intact viral antigens on the cell surface, but on the presentation of peptides in the context of a class I associated complex. In the example of the influenza virus, CTLs are not directed against the haemagglutinin molecule, which normally occurs at the cytoplasmic membrane during infection, but against the nucleoprotein. The response to the infected cell can be analysed using truncated nucleoproteins and peptides to limit the antigenic determinants to a peptide of 14 residues (Townsend et al., 1986).

The T cell receptors from different T lymphocyte subsets are known to be encoded by the same genes (Marrack and Kappler, 1986). The ability to discriminate between MHC antigens correlates with the distribution of the CD4 and CD8 cell surface markers. The class I restricted cells are CD8 positive and CD4 negative. The CD8 molecule has been found to have affinity for intracellular class I heavy chains in activated T lymphocytes, an affinity which can persist through to the level of cell surface expression (Bushkin et al., 1988). There is some evidence that the interaction between the proteins involves the formation of a disulphide bond (Blue et al., 1988), in the vicinity of the junction between the α_1 and α_2 domains (Stroynowski et al., 1985; Flavell et al., 1986).

As class I molecules present peptides from endogenous proteins, a mechanism has been proposed to explain the interaction with antigen. The class I chains may come into close proximity with regions of the Golgi apparatus specialising with improperly folded proteins. Denatured polypeptides, or fragments, from within this compartment escape sorting into secondary lysosomes through binding to the class I antigens, and thus reach the cell surface (Germain, 1986). During T cell ontogeny, the presentation of self antigens in the thymus could be important for the deletion of autoreactive clones, a mechanism which results in immunological tolerance (Robertson, 1988; Kappler et al., 1988; MacDonald et al., 1988). This hypothesis is supported by the discovery that crystallised class I molecules appear to contain a peptide at the antigen binding site (Bjorkman et al., 1987a, 1987b).

1.2.4 Genetic Organisation

The structure of the class I gene reflects the domain organisation of the protein product (Fig. 1.3) (Srivastava et al., 1985; Flavell et al., 1986; Strachan, 1987; Duquesnoy and Trucco, 1988). The first exon encodes the 5' untranslated sequence and the signal peptide of about 24 amino acids in length. Each of the α_1, α_2 and α_3 domains is then encoded by a separate exon of approximately 270 nucleotides. The fifth exon is about 122 nucleotides and encodes the transmembrane segment plus part of the cytoplasmic tail. The remaining three exons contain the rest of the cytoplasmic domain (11 and 15 codons in exons 6 and 7) and the 3' untranslated region (400 nucleotides in exon 8) (Srivastava et al., 1985). In total, the genes are approximately 3 kb in length.

Cosmid clones containing class I gene sequences indicate that there are at least 20 or more members of the class I family, other than the

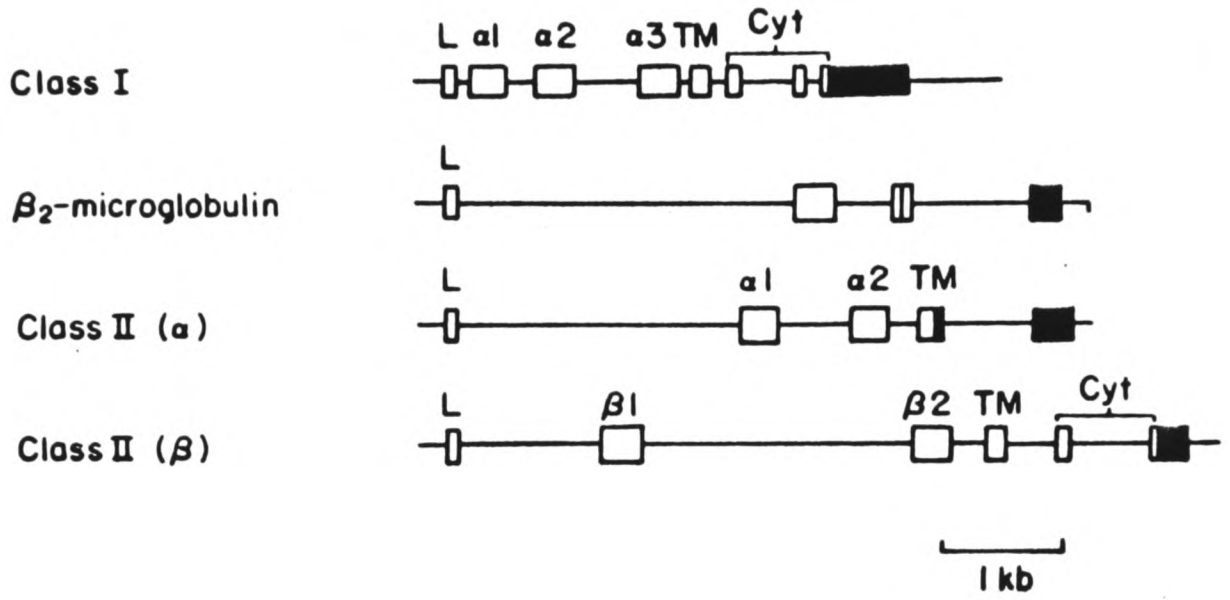


Fig 1.3

The gene stuctures of the class I and class II heavy and light chains. The exons are shown as open boxes, and the 3' untranslated regions are shown as shaded boxes. The exon-intron structure reflects the protein domain organisation, as shown by the designations above the open boxes.

(From Flavell et al., 1986)

classical HLA-A, -B and -C loci. One of these, designated HLA-E, has been mapped between the HLA-C and HLA-A genes (Paul et al., 1987; Carroll et al., 1987). The remainder probably lie telomeric to HLA-A (Orr and DeMars, 1983) and represent both genes and pseudogenes (Malissen et al., 1982; Srivastava et al., 1985, 1987; Strachan, 1987; Ponterotti et al., 1986). These may be analogous to the class I related Qa and Tla regions of the murine H-2. The numbers of the mouse genes vary between inbred strains (Winoto et al., 1983; Weiss et al., 1984; Flavell et al., 1986), and there is some evidence that this may also be true for the class I-like genes in man (Srivastava et al., 1985 and 1987). The functions of these extra genes are unknown, but they may serve as a reservoir of nucleotide sequences for the generation of polymorphism.

The organisation of the HLA class I region has been determined using PFGE (Carroll et al., 1987; Dunham et al., 1987; Chimini et al., 1988) and has been estimated at a minimum of 1,600 kb. The loci are arranged in the order HLA-B, -C, -E, -A, from the centromeric side of the cluster. The distance from HLA-B to HLA-C is about 250 kb, with a minimum distance between HLA-C and HLA-A of 1,000 kb. The HLA-E gene is at least 650 kb telomeric to HLA-C.

1.3 CLASS II REGION

The class II antigens are encoded by a series of linked loci in the D region of the MHC (Trowsdale et al., 1985; Trowsdale 1987). The products of three subregions can be defined serologically, and on the basis of mixed lymphocyte responses (MLR). In the MLR, donor lymphocytes, which are irradiated to prevent proliferation, are mixed with responder lymphocytes and the degree of stimulation depends upon the number of

determinants in common. In total, 26 Dw specificities have been defined. The additional specificities belonging to the individual subregion products are divided into 20 DR, 9 DQ and 6 DP, as recognised by the World Health Organisation (Bodmer et al., 1988). In mice there are fewer class II genes mapping to the I-A and I-E subregions (Klein, 1981; Widera and Flavell, 1985; Flavell et al., 1986).

The class II antigens are selectively expressed only on B cells, antigen presenting cells (such as macrophages and dendritic cells) and activated T cells. DR determinants are usually detected before DQ on T lymphocytes. The β chains of the class II molecules show extensive polymorphism, both at protein and nucleotide levels, but with the exception of DQα, the heavy chains appear to be invariant (Trowsdale et al., 1985; Trowsdale, 1987).

Both the α and β chains have a similar domain structure (Fig. 1.4). The α_1 and β_1 domains occur at the surface of the molecule, and are highly polymorphic. As with the class I heavy chain, the polymorphic regions can be divided into variable and hypervariable segments. The α_2 and β_2 domains are more conserved, and, like the class I α_3 domain and β_2m, are homologous to the Ig constant domain. Each section of the extracytoplasmic portion consists of 85-95 amino acids, and, with the exception of the α_1 domain, contains an intra-chain disulphide bond. The C-terminus of each chain consists of a mostly hydrophobic transmembrane segment of about 25 residues, and a cytoplasmic tail which varies between 10 and 20 amino acids in length. The molecule is glycosylated on the α_1, α_2 and β_1 domains, the differences in the carbohydrate moeities being largely responsible for the observed differences in the molecular weights of the two chains.

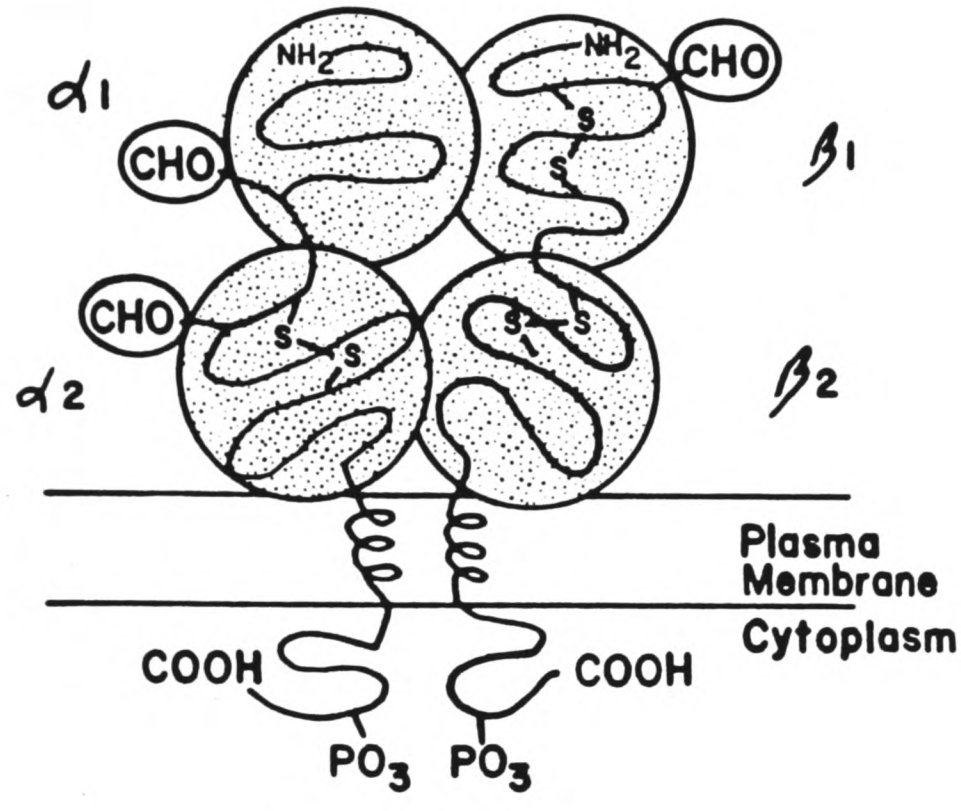


Fig. 1.4

Basic domain structure of an HLA class II molecule, showing the positions of carbohydrate attachment on the external domains. (From Duquesnoy and Trucco, 1988)

1.3.1 3-D Structure

No class II molecule has been crystallised to study in an analogous fashion to the class I antigens, but the overall domain organisation suggests that the two have a similar three dimensional arrangement. Comparison of the predicted class II secondary structure with the known class I crystal structure suggests that the α_1 and β_1 domains can form a binding cleft with N-terminal β-sheet forming the floor and C-terminal α-helices the walls. Homologous α-helices from different molecules show a pattern of conserved residues every 3/ 4 amino acids, which form a face to the helix. In the β-strand, polymorphic residues occur at the bottom of the cleft. Regions between the defined secondary structures may form loops. Like the groove in the class I antigen, the binding cleft of the class II molecule contains some polymorphic sites which would interact with a peptide, and others which could interact with a T cell receptor. The cleft could accomodate an α-helix or extended polypeptide such that a T cell receptor could recognise the MHC determinants and the presented antigen simultaneously (Brown et al., 1988).

1.3.2 Biosynthesis

The signal peptides of the heavy and light chains probably bind to a receptor at the endoplasmic reticulum, and are cleaved off on the luminal side of the membrane (for reviews, see Cresswell, 1987; Cresswell et al., 1987). The immature class II molecule is found in association with a third protein, the invariant chain (I or γ) which is a 216 amino acid polypeptide encoded on human chromosome 5. The complex is transported to the Golgi apparatus, during which time it becomes

glycosylated. Dissociation of the I chain occurs just before insertion into the cell membrane, and is blocked by monensin and chloroquine. The inhibition by these drugs suggests that an intracellular protease in an acidic compartment may be required to catalyse the dissociation process (Blum and Cresswell, 1988).

The level of class II expression is down regulated by prostaglandin E in murine macrophages (Snyder et al., 1982). In addition, factors from activated T lymphocytes and γ interferon can induce class II negative cells to become positive. Although an interferon response element has been located 5' to class II genes, it is not found in conjunction with an enhancer similar to the class I genes, and the mechanism of regulation may be different (Israel et al., 1986).

1.3.3 Function

The MHC class II antigens act as the restriction elements for the helper T cells (T_h) of the immune system, and are therefore involved in the regulation of humoral responses (Shevach and Rosenthal, 1973; Katz and Benacerraf, 1976; Benacerraf, 1981). In rodents, the alleles of the class II genes have been shown to govern the ability of the animal to produce antibody against a given immunogen, and have been termed the immune response (Ir) genes (Benacerraf, 1981).

The T cell receptor of the activated T_h cell recognises only antigen which has been processed by a class II positive accessory cell. As with CTLs, the antigenic determinant is a peptide fragment derived from the original protein immunogen (Watts et al., 1985; DeLisi and Berzofsky, 1985; Ogasawara et al., 1987). The most antigenic peptides are believed to be those capable of forming short stretches of α-helix which are amphipathic in nature (DeLisi and Berzofsky, 1985). This may be

important for stabilisation and interaction with the class II binding site.

T helper cells recognise class II by interaction between the CD4 molecule and the MHC product. The nature of this association is unknown.

A mechanism for the interaction of the class II molecule with the exogenous protein antigens has been proposed by Germain (1986) on the basis of the chloroquine sensitivity of presentation. Antigens are taken up by the processing cell by receptor mediated endocytosis or pinocytosis. They pass to endosomes, where they are denatured or fragmented. Here, the peptides come into contact with class II molecules which are being recycled from the cell surface, and are returned to the cytoplasmic membrane as a class II asssociated complex.

1.3.4 Genetic Organisation

A number of laboratories have been involved in the characterisation of the class II antigens at a molecular level (for reviews, see Trowsdale et al., 1985; Trowsdale, 1987; Strominger, 1987). Analysis of both cDNA and genomic clones has allowed the division of the HLA-D region into subregions DR, DQ, DP, DN/ DO, each containing at least one α and β chain gene. Serologically defined products have only been detected from DR, DQ, and, DP.

Studies of recombination within the HLA-D region and analysis of deletion mutants have defined that DP is centromeric to DR (Shaw et al., 1981; Kavathas et al., 1981). Strong linkage between DQ and DR specificities in given haplotypes implies that these genes are also close together (Trowsdale et al., 1985; Cohen et al., 1985a). The exact organisation cannot, however, be determined from these data, or from cosmid walking, as none of the cloned subregions have been linked. Using

PFGE with chain specific probes, the order has been established by Hardy et al. (1986) as centromere-DP-DNA/ DOB-DQ(2)-DQ(1)-DRB-DRA-telomere, with an overall size of 1,100 kb. This has been confirmed by the work of others (Carroll et al., 1987; Dunham et al., 1987).

The α chains are not very polymorphic, with the possible exception of DQα, as comparison of the gene sequences shows a 50% conservation between the coding sequences, increasing to 60% for the α_2 domain, connecting peptide and transmembrane regions (Trowsdale et al., 1985). Most of the allelic variation which contributes to the serologically defined antigenic determinants resides in the β chain. Both α and β genes share a similar structure which reflects the domain organisation of the protein, although β chains may have extra introns which separate the cytoplasmic domain and the transmembrane region, and make them more like class I α chain genes (Fig. 1.3).

The DP Subregion

Cosmid clones encompassing 100 kb of genomic DNA have been isolated from the DP subregion (Trowsdale et al., 1984; Okada et al., 1985a). These contain two pairs of α and β chain genes, presumably derived from the duplication of a single ancestral unit. An α/β pair is arranged head to head, so that each gene is inverted with respect to its neighbour. The order of the genes within the cosmid clones is DPB2-DPA2-DPB1-DPA1. The DPB1 gene is unusual as it contains a processed pseudogene in the intron 5' to the β_1 coding domain. The pseudogene is derived from the ribosomal large subunit protein L32 (Young and Trowsdale, 1985; Trowsdale et al., 1984). Although the DPA1 and DPB1 genes are expressed (Okada et al., 1985a), sequence analysis suggests that DPA2 and DPB2 are probably pseudogenes (Okada et al., 1985a; Kappes and Strominger, 1986). The functional DPB1 gene is probably the human equivalent of the mouse $A\beta_3$ gene, as determined from comparison of nucleotide sequences (Widera

and Flavell, 1985).

The DO/ DN Subregion

The HLA-DNA gene has been cloned and characterised from cosmid genomic libraries by cross-hybridisation with the DRα chain cDNA (Trowsdale et al., 1983; Spielman et al., 1984). It has been mapped to chromosome six by human-mouse somatic hybrids (Spielman et al., 1984) and by PFGE (Hardy et al., 1986). There is evidence for an expressed transcript (Inoko et al., 1985; Trowsdale and Kelly, 1985), but not for a protein product. The mRNA is 3.5 kb in length, rather than the 1.2 kb predicted from the genomic sequence. The large transcript may arise owing to an abnormal polyadenylation signal, ACTAAA (Trowsdale and Kelly, 1985). No polymorphism has been detected at a genomic level (Inoko et al., 1985; Trowsdale and Kelly, 1985).

HLA-DOB cDNA clones have been isolated from a homozygous cell line. Sequence analysis suggests it is closest to the murine Aβ_2 gene and, by analogy, ought to map between DP and DQ/DR (Tonnelle et al., 1985)

Genomic clones linking DOB to DNA or to other class II subregions have not been isolated. The gene products are unlikely to constitute a heterodimer, as they are not under the same regulatory controls (Tonnelle et al., 1985).

The DQ Subregion

Like the DP subregion, DQ contains two pairs of related α and β genes. In addition, another β chain gene (DV) has recently been mapped to the region between the DQA2 and DQB1 loci (Trowsdale and Campbell 1988). Both cDNA (Schenning et al., 1984) and genomic clones (Spielman et al., 1984; Okada et al., 1985b) have been isolated from the DQ subregion. The clones show that the genes in each pair are inverted with respect to one another, as in the DP subregion. The α genes are $\sim$6 kb, and the β genes 7-10 kb in length, with DQA1 and DQB1 separated by 12 kb, and DQA2 and

DQB2 by 8 kb. The two clusters are not linked, but, as HLA-DQA1 and -DQB1 are in linkage disequilibrium with DR, the suggested gene order is DQB2-DQA2-DQB1-DQA1- DR (Spielman et al., 1984). It is not known whether the DQB2 and DQA2 genes are expressed. Southern blots show that the DQB2 and DQA2 genes are not polymorphic at a genomic level (Okada et al., 1985b), but that both DQB1 and DQA1 encode different allelic forms (Auffray et al., 1983; Okada et al., 1985b; Spielman et al., 1984). The genomic polymorphisms can be equated with serological specificities, and used to define further polymorphisms at the nucleotide level. Sequence comparison infers that DQB1 is analogous to the murine Aβ1 gene (Auffray et al., 1983).

The DR Subregion

The DR subregion contains one DRα chain gene, but a variable number of DRβ chain genes, which depends upon the DR haplotype (Bohme et al., 1985). The DRα chain cDNA and gene sequences were first elucidated by Lee et al. (1982). Only one polymorphism has been found (Trowsdale et al., 1985; Trowsdale, 1987). In contrast, the isolation of DRβ chain cDNA clones (Long et al., 1985) has allowed their classification into 4 major subsets, at least two of which appear to be non-allelic. Characterisation of cosmid clusters (Spies et al., 1985; Meunier et al., 1986) containing genes from the DR subregion of a DR4 haplotype has linked HLA-DRA to HLA-DRBIV (the light chain gene encoding the DRw53 serological specificity). The genes are arranged 3' to 3', and separated by 90 kb. Isolated conserved elements consisting of the DRβ 5' promoter and signal sequences occur six times within the cloned region (Spies et al., 1985) and an individual β_1 exon has been located 15 kb 3' to the DRA gene (Spies et al., 1985; Meunier et al., 1986). The function of these repeat elements is unknown, but there is speculation that they are required for the coordinate expression of the α and β chain genes (Spies

et al., 1985). As the β_1 exon is the most polymorphic of the light chain domains, and encodes the specificities of the DR protein products, the isolated exon may serve as an extra reservoir for DNA exchange (Meunier et al., 1986).

The other two DRB genes are organised 3' to 5', separated by 22 kb. Overall, the genes are probably arranged DRBI-DRBII-DRBIV-DRA. The DRBII locus represents a pseudogene, whereas DRBI encodes the DR4 serological specificity (Spies et al., 1985).

Restriction fragment analysis of homozygous typing cell lines shows that RFLPs with the enzymes Taq I, Pvu II and Bam HI (Bohme et al., 1985; Trowsdale et al., 1985) can be used to classify alleles at the DRB loci. The DR2 to DR7 haplotypes all have three β chain genes, of which DRBI encodes the DR serological specificity. Both DR4 and DR7 have the pseudogene DRBII, and the DRBIV gene. DR1 has two genes (DRBI and DRBII) (Sorrentino et al., 1985), whereas there is evidence for only one gene (DRBI) in the DR8 haplotype. The DRw52 specificity is associated with the DRBIII gene.

The DR subregion of the MHC is equivalent to the I-E subregion of the H-2 complex, as the DRα chain is most homologous to the I-Eα chain.

1.4 THE CLASS III REGION

The loci for the complement components C2, C4 and factor B were linked to the HLA on the basis of inheritance of polymorphic forms of the proteins with HLA haplotypes (Allen, 1974; Fu et al., 1974; O'Neill et al., 1978; Awdeh and Alper, 1980). They could be mapped to a positon between the HLA-D and HLA-B loci from family recombination studies (Weitkamp and Lamm, 1982; Robson and Lamm, 1984). The order of the genes within the class III region was determined from a series of overlapping

cosmid clones (Carroll et al., 1984) as C2-factor B-C4A-C4B, but, the orientation of the cluster and the precise location with respect to the class I and II regions was in dispute. By studying extended haplotypes and recombination frequencies, groups positioned the complement genes closer to HLA-D (Robson and Lamm, 1984; Yunis et al., 1985; Wilton and Charlton, 1986), but could not determine whether C4 was closer to HLA-B than factor B (Olaisen et al., 1983; Wilton and Charlton, 1986) or vice versa (Marshall et al., 1984; Abbal et al., 1987). Using a combination of cosmid cloning and PFGE, the orientation and distances between class III genes and flanking regions of the HLA have now been established (Dunham et al., 1987). C4 is positioned ∿350 kb from HLA-D, and C2, at the opposite end of the complement cluster, is ∿650 kb from HLA-B.

1.4.1 The Complement Pathway

The complement pathway consists of about 13 components, under the control of at least 7 other plasma proteins and 15 membrane bound molecules or receptors (Fig. 1.5) (Reid and Porter, 1981; Reid, 1986; Campbell et al., 1988). There are two pathways of activation, termed the classical pathway and the alternative pathway. The classical pathway is activated by the Fc regions of immunoglobulins IgG and IgM in immune aggregates, whereas the alternative pathway is mainly activated by the polysaccharides of bacterial and yeast cell walls. The pathways merge at the level of membrane attack complex (MAC) formation, with the complement component C3 playing a central role in the coordination of the early and late stages of the cascade.

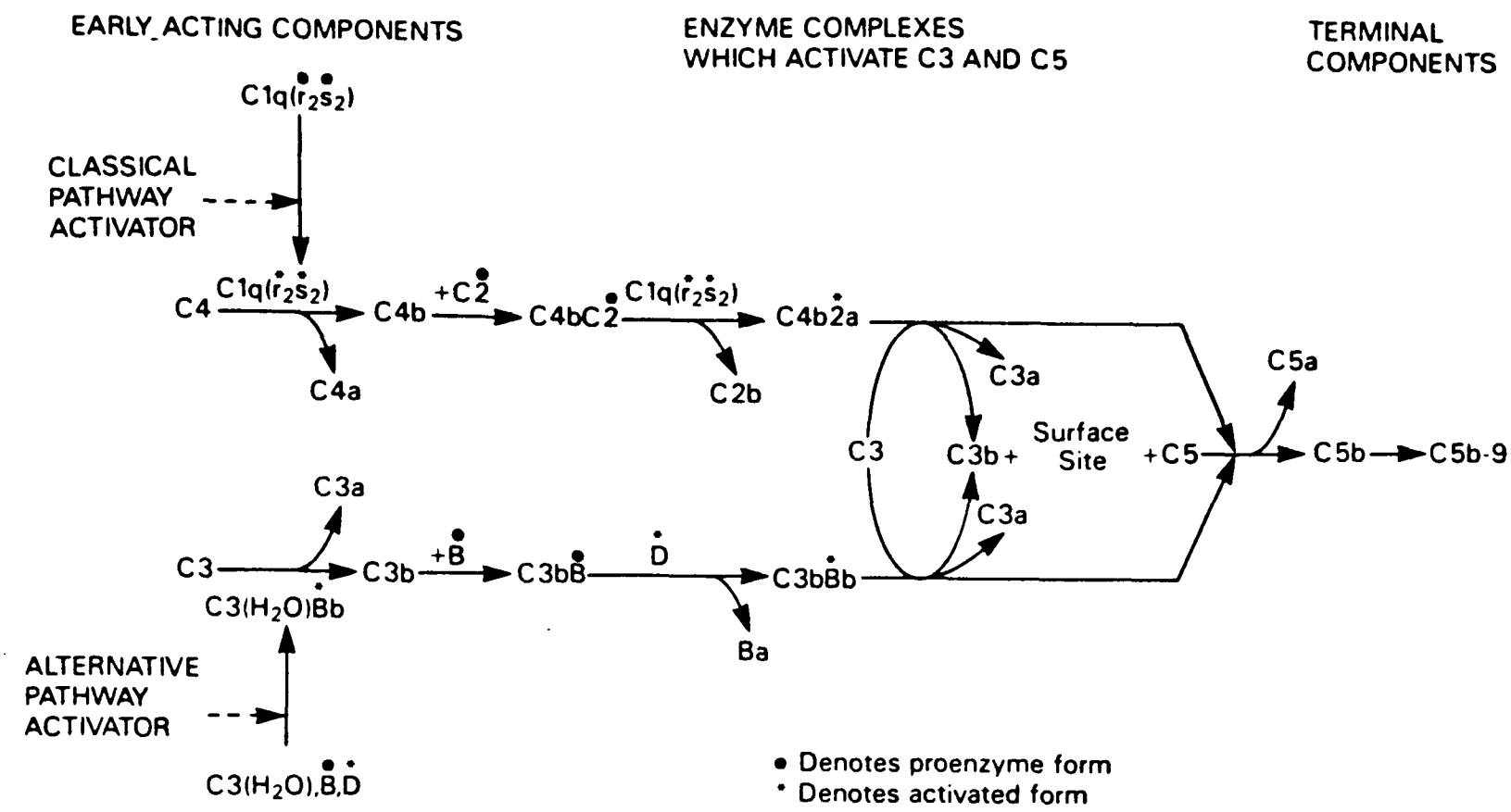


Fig. 1.5

The two pathways of complement activation. The classical pathway is activated by antibody-antigen aggregates. The alternative pathway is activated by antibody-antigen aggregates or bacterial polysaccharides. (From Reid, 1986)

1.4.2 Functions of the Class III Complement Components

C2

C2 is the third serine protease of the classical pathway for the activation of complement (Reid and Porter, 1981; Reid, 1986; Campbell et al., 1988). It is synthesised as a 754 amino acid precursor which is glycosylated and cleaved to yield the 734 residue plasma form of M_r 102 kD (Bentley and Campbell, 1986). Three major protein forms have been identified, of which the most common is C2C (97% of the caucasian population), the others being a more acidic variant (C2A; gene frequency <1%) and a more basic variant (C2B; gene frequency ~2%) (Alper, 1976).

In the activation of the classical pathway, C1 interacts with the Fc portion of IgG or IgM in immune aggregates (Reid and Porter, 1981; Reid, 1986). The $C\bar{1}s$ protein of the $C\bar{1}$ complex activates C4 by cleaving the C4a peptide (M_r, 9 kD) from C4b. C2 then associates with C4b in a Mg^{2+} dependent reaction, and is itself cleaved by $C\bar{1}s$ into a non-catalytic chain of 30 kD (C2b) and a catalytic chain of 70 kD (C2a) which contains the C-terminal serine protease domain. The $\overline{C4b,2a}$ complex acts as the C3 convertase of the classical pathway, splitting the C3 molecule into C3a, a 9 kD anaphylatoxin, and C3b. Binding of the C3b molecule to C4b in the $\overline{C4b,2a}$ complex (Takata et al., 1987) alters its activity to a C5 convertase, thus initiating the formation of the membrane attack complex.

Factor B

Factor B is a serine protease of the alternative pathway of complement with homologous functions to C2 of the classical pathway. It is synthesised as a 764 amino acid polypeptide which is glycosylated and cleaved to give a 739 residue plasma form of M_r 90 kD (Bentley and Campbell, 1986). There are two major variants, F and S, plus two less

common variants F_1 and S_1, and other minor forms (Alper et al., 1972; Mauff et al., 1978).

In the activation of the alternative pathway, factor B associates with the C3b-like molecule $C3(H_2O)$ in a Mg^{2+} dependent manner. In this form it is cleaved by factor D, a 24 kD serine protease, into a non-catalytic chain of 30 kD (Ba), and its C-terminal catalytic chain of 60 kD (Bb). $\overline{Bb,C3b}$, or, $\overline{BbC3(H_2O)}$, acts as the C3 convertase of the alternative pathway. Like $\overline{C4b,2a}$, association of further C3b with $\overline{C3b,Bb}$ leads to the formation of a C5 convertase, initiating the formation of C5b-9.

C: C4

C4 is the second component of the classical pathway of complement activation. More than 35 polymorphic variants, attributable to two isotypic loci have been defined (O'Neill et al., 1978; Olaisen et al., 1979; Mauff et al., 1983). It is synthesised as a single polypeptide chain of M_r 200 kD, which is subsequently glycosylated and cleaved into the disulphide bonded α, β and γ chain plasma form.

The isotypes, C4A and C4B, have different activities in haemolytic assays (C4B>C4A), C4A being more reactive with amino groups than hydroxl groups and C4B being equally reactive with either (Law et al., 1984; Isenman and Young, 1984). The differences in binding to amino or hydroxyl groups may be of biological importance, allowing the removal of a wider range of pathogens with differing surface antigenic structures (Porter, 1983).

C4 is activated by the C1s protein of the C1 complex, which removes the C4a peptide (M_r, 9 kD) from C4b. The C4b molecule binds covalently to cell surfaces via a reactive carboxyl group on a glutamyl residue forming an ester or amide bond (Law et al., 1984; Campbell et al., 1980), or may react with water to form fluid phase C4b. The interaction of C4b with C2 leads to the formation of the C3 convertase of the

classical complement pathway, and, ultimately to cell lysis by the MAC.

In addition to the many allotypic forms of C4, the C4d fragments of the isotypic C4A and C4B proteins, when bound to erythrocytes, carry the Rodgers and Chido blood group determinants, respectively (O'Neill et al., 1978).

1.4.3 Genetics of the Class III Complement Components

The cDNA and genomic structures for C2, factor B and C4 have all been determined (Campbell et al., 1986; Carroll and Alper, 1987; Campbell et al., 1988). The genes have been linked together in a cluster of overlapping clones (Carroll et al., 1984) from a cosmid library prepared from an individual HLA typed as: HLA-A3, 31; -B14, 37; C2C; BfS; C4A 2, 3; C4B 2, 1; HLA-DR1, 2 (Grosveld et al., 1982). Within the 98 kb cloned, single genes for factor B and for C2 were mapped less than 2 kb apart, and subsequently shown to lie within 0.5 kb from sequence analysis (Wu et al., 1987). Two C4 loci were mapped 10 kb apart, the A gene being 30 kb from the 3' end of factor B. All the complement loci were organised in the same transcriptional orientation.

A complete molecular map of the C4 region (Carroll et al., 1985a, 1985b; White et al., 1985) has established that each C4 gene is linked to a gene for cytochrome P450 steroid 21-hydroxylase (21OH). The 21OH loci are 2-3 kb from the 3' ends of the C4 loci, and, again, are in the same transcriptional orientation as the complement genes.

C2 and Factor B

The C2 and factor B genes encode mRNA species of 2.9 and 2.6 kb, respectively (Morley and Campbell, 1984; Campbell et al., 1984; Bentley and Porter, 1984). The cDNAs for both C2 (Bentley and Porter, 1984) and factor B (Woods et al., 1982; Campbell and Porter, 1983) have been isolated from cDNA libraries using oligonucleotide probes based on the

available amino acid sequence. Complete determination of the cDNAs (Morley and Campbell, 1984; Bentley, 1986) reveal that the two proteins share structural homologies which may reflect their functional homologies in the complement pathways.

Both C2 and Factor B belong to a class of proteins which are characterised by a 60 amino acid repeat. The family also includes other complement regulatory factors plus some non-complement proteins such as β_2 glycoprotein 1 and the interleukin-2 receptor (Reid et al., 1986). The N-terminal domains of C2 and factor B contain three of these repeat units, which may have arisen by duplication of a single ancestral segment (Morley and Campbell, 1984). Determination of the gene sequence of factor B confirmed this view, as each repeat is encoded within a distinct exon (Campbell et al., 1984). The C-terminal portion of each protein shares homology with the classical serine proteases, although the catalytic subunits are twice as large as, for example, trypsin (Campbell and Porter, 1983; Morley and Campbell, 1984). The N-terminal portions of the catalytic domains share homology with the von Willebrand factor. Similar structural domains have been found in cartilage matrix protein and the cell surface glycoprotein Mac-1 (Pytela, 1988).

In total the factor B gene is encoded by 18 exons (Campbell and Porter, 1983; Morley and Campbell, 1984). The C2 gene is three times larger (18 kb as against 6 kb), and, although the gene structure has not been published, a similar organisation would be expected. Given the close functional and structural homologies, and the physical proximity of the two loci, C2 and factor B probably arose from a common ancestral gene by duplication.

The genomic sequence of the factor B F allele differs from the S allele in six positions, two of which lie in exons (Campbell et al.,

1985, Campbell et al., 1988). One is a silent mutation (C to T) at position 3 of the codon for Tyr^{199}. The other (G to A) lies at position 2 of codon 7, and results in the substitution of an Arg (S allele) by a Glu (F allele) in the Ba fragment of the protein. This amino acid substitution is sufficient to explain the observed electrophoretic mobilities of the alleles, with F carrying less positive charge.

The C to T substitution introduces an Rsa I polymorphism which is associated with only some of the F alleles (Bentley and Campbell, 1986; Campbell et al., 1988). The second, G to A, mutation introduces an Msp I polymorphism which can distinguish the F and S alleles at the nucleotide level (Bentley and Campbell, 1986; Campbell et al., 1988). Additional polymorphisms within the C2 gene can be used to subdivide C2C allelic variants associated with the F form of factor B (Cross et al., 1985; Bentley et al., 1985; Bentley and Campbell, 1986). A polymorphic Taq I site in the 3' end of C2 results in a 6.6 kb fragment with a frequency of 39%, or a 4.5 kb fragment with a frequency of 61% (Cross et al., 1985; Bentley et al., 1985). These appear to be correlated with variants of F detected by isoelectric focusing (Teng and Tan, 1982; Abbal et al., 1984), F^a (41%) and F^b (59%).

Southern blot analysis of the C2 gene has defined an Sst I polymorphism near the 5' end of the gene (Bentley et al., 1985). Fragments of 2.7, 2.65, 2.6 and 2.4 kb have been detected, and shown to correlate with similar length variations of other restriction fragments (Bam HI and Hind III). This suggests that the observed polymorphisms probably arise from the insertion or deletion of short sequences (Campbell et al., 1988).

C4 and 21OH

The C4A and C4B isotypic variants are encoded by two loci separated by 10 kb (Carroll et al., 1984) The genes differ in size, owing to the

presence or absence of a 6-7 kb intron 2.5 kb away from the 5' end. Long genes, 22 kb, include all C4A and some C4B genes, including most C4B1 alleles linked to C4A3. Other C4B genes are only 16 kb in size (Carroll et al., 1985c; Yu et al., 1986; Palsdottir et al., 1987a, 1987b). The long gene can be distinguished from the short gene on genomic Southern blots by characteristic restriction fragments associated with the 5' end; 4.8 kb Bam HI and 7.5 kb Kpn I fragments in the long gene as opposed to 3.3 kb Bam HI and 8.5 kb Kpn I fragments in the short gene. Additional polymorphisms in Taq I restriction digests have been used as markers for different C4 genes. A 7.0 kb fragment is associated with C4A long genes, a 6.0 kb fragment is associated with C4B long genes (as opposed to a 5.4 kb fragment with C4B short genes) and a 6.4 kb fragment is found in individuals with a C2C, BfS, C4A QO, C4B 1 complotype. The latter probably arises from a recombination event between the two C4 loci which deletes the majority of the C4A gene.

Direct sequence analysis of the C4 cDNAs and genomic clones (Belt et al., 1895; Yu et al., 1986) define only 15 differences between the alleles. Twelve of these are occur on the C-terminal side of the thiolester bond. Four residues, at positions 1101-1106, can be assigned to isotype determination; PCPVLD in C4A and LSPVIH in C4B. These arise owing to 5 nucleotide substitutions within a 17 bp block. The major Rodgers and Chido determinants are defined within the region 1188-1191. The Rg1 epitope, VDLL, and the Ch1 epitope, ADLR, differ by 4 nucleotides within a 16 bp unit. Restriction fragment length polymorphisms with the enzymes Nla IV and Eco 0109 have been used to define the isotypic and Rg1/ Ch1 epitopes at a genomic level (Yu and Campbell, 1987; Yu et al., 1988). The other Rodgers and Chido determinants appear to consist of both continuous and discontinuous epitopes.

These polymorphisms can help to study the nature of C4 null alleles, which occur at a relatively high frequency (5-15% for C4A QO, and 10-20% for C4B QO). Although studies suggest that deletions of one of the two loci may be responsible in 60% of cases (Carroll et al., 1985c; Schneider et al., 1986), RFLP analysis, along with molecular cloning and sequencing suggests that an apparently null allele may arise from the conversion of the gene at the "unexpressed" locus to encode the same isotype or allotype as the expressed locus (Schneider et al., 1986; Yu et al., 1986; Palsdottir et al., 1987b). Therefore, in the haplotype B35, DR1, C4A 3, C4A 2, C4B QO, the adjacent loci encode the same C4 isotype (Palsdottir et al., 1987b; Schneider et al., 1986), but in the haplotype B44, DR6, C4A 3, C4B QO, the analysis indicates that the genotype is C4A 3, C4A 3 (Yu and Campbell, 1987).

Each C4 locus has been shown to be associated with a gene encoding 21-hydroxylase, located 2-3 kb from the 3' end (Carroll et al., 1985a, 1985b; White et al., 1985). The 21OHA and B loci can be identified using the restriction endonuclease Taq I. A 3.7 kb fragment segregates with the 21OHB gene, and a 3.2 kb fragment with the 21OHA gene. This difference has been used with deleted haplotypes to define that the 21OHB gene is functional (White et al., 1985). Individuals with deletions of this locus suffer from the most severe form of congenital adrenal hyperplasia (CAH). Confirmation that 21OHA is a highly homologous pseudogene comes from sequence analysis (Higashi et al., 1986; White et al., 1986; Rodrigues et al., 1987). Comparison of the available nucleotide data infers that there could be some polymorphism at the 21OHB locus in normals, and that in certain cases of CAH the 21OHB gene can have 21OHA-like characteristics (Rodrigues et al., 1987).

1.5 OTHER MHC ASSOCIATED GENES

A number of other genes have been mapped to the MHC on the basis of associations between polymorphic variants or deficiencies with known MHC loci. In mouse, and rat, neuraminidase shows a linkage with class I and class II alleles (Figueroa et al., 1982; Samollow et al., 1987), and in man, there appears to be an association between neuraminidase deficiency and the class III region (Oohira et al., 1985; Harada et al., 1987). The cellular oncogene pim has also been mapped to the appropriate chromosomes in man and mouse, although the exact location with respect to the MHC is unknown (Hilkens et al., 1986; Cuypers et al., 1986; Nagarajan et al., 1986). The haemochromatosis locus is thought to lie close to the HLA-A locus. A region of homology to the ferritin heavy chain probe has been mapped to the appropriate region of human chromosome 6 on the basis of cross-hybridisation (Cragg et al., 1985; McGill et al., 1985). Other genes which could lie within the 6p21-23 region include HSP 70 (Goate et al., 1987; Harrison et al., 1987), microtubule associated protein 2, the prolactin gene, and the α subunit of chorionic gonadotrophin (see Olaisen et al., 1987).

1.6 THE HLA AND DISEASE ASSOCIATIONS

A range of human diseases show linkage to HLA markers, and can be divided into three major groups depending on the aetiology of the disease.The first, which includes rheumatoid arthritis, type I diabetes (IDDM) and coeliac disease is autoimmune in character, whereas the second, including 21-hydroxylase deficiency and idiopathic haemochromatosis is not. The third class of disease encompasses those of

uncertain pathogenesis, such as narcolepsy and psoriasis (Batchelor and McMichael, 1987; Bodmer and Bodmer, 1978).

One of the major problems in correlating disease and HLA associations arises from the phenomenon of linkage disequilibrium. Within a given population, the number of combinations of alleles at the loci is much lower than that expected from the frequencies of the individual allotypes. This means that certain combinations of class I and class II alleles occur as a stable inherited unit, or haplotype. The linkage between DR and HLA-B alleles can be extended to include specific combinations of the complement genes. From over 1,000 possible combinations of factor B, C2, C4A and C4B, only 47 have been found. No cross-overs have been observed in the complement gene cluster, presumably because of the closeness of the genes at a molecular level. Therefore, alleles of HLA-DR, complement and HLA-B can be referred to as an extended haplotype (Alper et al., 1986). Analysis of the other HLA loci indicates that there is only a limited variation at these positions.

Maintenance of linkage disequilibrium may arise from the selective pressure which ensures a specific combination of alleles essential for a varied immune response repertoire. Interestingly, one of the extended haplotypes which is often found in autoimmune diseases appears to show male segregation distortion, similar to the t-haplotype phenomenon in mice. Thus, GLO2 (but not GLO1)-DR3-C2C-BfS-C4A QO-C4B 1-B8 is preferentially inherited in IDDM and gluten sensitive families, propagating a deleterious haplotype as a result of a chromosomal selective advantage (Awdeh et al., 1983).

To determine how significant the association with the HLA is, all the markers ought to be considered. Once a specific allele or haplotype has been identified as being prevalent in the patient population, but not

amongst normal individuals, it is important to consider the ethnic origins of the study group. If a given allele always shows a high concordance, irrespective of the racial background, there is a reasonable chance that this allele is the primary causative agent. If different alleles at the same locus show a high correlation, then the true disease associated gene may be in linkage disequilibrium with this locus.

In some cases a direct role for the class I and class II antigens in the pathogenesis of an autoimmune disease can be defined (Todd et al., 1988a, 1988b; Batchelor and McMichael, 1987). Such diseases arise as the result of the normal function of the immune system when the balance between the recognition of foreign antigens and tolerance towards self is broken down. In order for this to occur, a triggering antigen must be different enough to break the self-tolerance, but similar enough to induce a cross-reactivity (Oldstone, 1987). In the case of Reiter's syndrome and ankylosing spondilitis, association with B27 persists regardless of the race of the patient. Both diseases can occur after infection with a number of bacteria such as Shigella, Yersinia, Salmonella and Klebsiella. Sequence analysis of one of the immunodominant antigens in Klebsiella pneumoniae has revealed a sequence homology between a peptide of 6 amino acids and part of the hypervariable domain of HLA-B27. Thus, presentation of the peptide during infection could result in the proliferation of cytotoxic T lymphocytes directed against the B27 antigen (Oldstone, 1987)

Haplotype matched patients and controls can also be used to define genetic polymorphisms that split serological determinants into subgroups that correlate with the normal or diseased population. In the case of the B8-DR3-DQw2 associated diseases, RFLP analysis has identified a DQB1 associated 3.7 kb Bam HI fragment absent in IDDM patients, but common in

haplotype matched controls (Owerbach et al., 1986). In addition, a DQB1 15 kb Hind III polymorphism was found to occur in myasthenia gravis, but not in other autoimmune disease or normal individuals (Bell et al., 1986).

In IDDM, human, mouse and rat models have all been shown to require particular recessive MHC encoded genes to develop susceptibility to the disease (Todd et al., 1988a, 1988b; Prochazka et al., 1987; Eisenbarth, 1986). Furthermore, autoimmunity is T cell mediated, as illustrated by the transfer of T lymphocytes from diabetic to normal animals. In man, DR3/4 heterozygotes have a higher relative risk of contracting IDDM than DR3 or DR4 homozygotes, whereas DR2 individuals appear to be protected. Analysis of Swedish patients indicates that DQβ chain associated polymorphisms in DR2 diabetics differ from normals, although in DR3 patients they do not (Bohme et al., 1986). Sequencing DQB genes from humans, and the Aβ genes from non-obese diabetic (NOD) mice, has shown that no IDDM sequence is unique to patients, although the occurrence in normals may be rare. In both man and mouse, normal β chains contained Asp^{57}, but patient chains contained Ser (also Ala or Val in man) at this position (Todd et al., 1988a; Acha-Orbea and McDevitt, 1987). By comparison with the class I 3-D model (Bjorkman et al., 1987a, 1987b), residue 57 of the class II molecule would occur at the site involved in antigen binding, and may alter either the ability to bind the antigen or T cell recognition of the complex. The DQ molecule is unusual in that it has a polymorphic heavy chain, as well as a polymorphic light chain (Trowsdale et al., 1985). The increased incidence of DR3/4 heterozygotes in the diabetic population may reflect a complementation which produces a DQ heterodimer encoded by genes on different chromosomes (Bohme et al., 1986; Todd et al., 1988a). A model could then be proposed in which the DQ molecule in DR2 resistant individuals induces tolerance to the

triggering factor by deletion of the T cell subset which would normally recognise the antigen. In diabetic individuals, the T cell subset involved in the immune response is primarily directed against an allotype of DQ, and is not deleted during T cell development. When it encounters the relevant environmental trigger (such as a cross-reactive viral peptide), or self-antigen in the context of the DQ molecule, it is activated. The autoreactive clone would then be responsible for the induction of the antibody response against the β -ells of the pancreas which is observed in IDDM patients. In addition, this hypothesis explains the lack of 100% concordance between diabetes and monozygotic twins, or inbred mouse strains. Although the somatic genetic background is identical, the production of T cell receptors depends on random rearrangements of DNA in T lymphocytes (Eisenbarth, 1986; Todd et al., 1988a).

Complement deficiencies may also lead to associations between the HLA region and diseases. Both C2 and C4 null alleles occur at a higher frequency in systemic lupus erythrematosis (SLE) patients than in normals. DR2 is commonly associated with a C2 null extended haplotype (Alper et al., 1986) and DR3 with a C4A null haplotype. The latter includes a deletion of the C4A and 21OHA genes (Carroll et al., 1985b, 1985c). In addition, drugs which cause drug-induced SLE appear to inhibit the activity of C4, which may result in inefficient removal of immune complexes (Sim et al., 1984). Tissue damage in SLE results from a failure to limit the size of immune complexes at the site of infection, and to contain the response to invading organisms by attracting phagocytes. As C4A binds to, and solubilises immune complexes more effectively than C4B (Law et al., 1984) this could explain a direct involvement of the C4A locus.

C4 deficiencies can also result in defective antibody mediated complement activation, causing a reduced host response to invading organisms. This may account for the development of subacute sclerosing panencephalitis in C4 deficient patients which arises from the persistent subclinical infection by the measles virus.

Other HLA associated diseases that are not related to class I, class II or complement genes include 21OH deficiency, neuraminidase deficiency, haemochromatosis, tuberculosis and leprosy. 21OH deficiency is correlated with Bw47. However, analysis of patients from the same ethnic background revealed common extended haplotypes (Fleischnick et al., 1983), including some with rare C4 variants. Mapping of two 21OH genes 3' to the two C4 loci (White et al., 1985; Carroll et al., 1985a) allowed further characterisation of the haplotypes on the basis of Taq I restriction fragment differences which distinguished between the 21OHA and 21OHB genes (White et al., 1985). It was concluded that the 21OHB gene was important for steroid biosynthesis, as this gene was deleted in the most severe form of the disease. Different extended haplotypes appear to be increased in other forms of the disease, reflecting its heterogeneous nature (Alper et al., 1986).

Recent studies have indicated that neuraminidase deficiency in man correlates with 21OH deficiency in some families (Oohira et al., 1985). Genes for neuraminidase have been tentively mapped to the MHC in both mouse and rat (Samollow et al., 1987; Figueroa et al., 1982), and it is possible that, by analogy, the neuraminidase deficiency is in linkage disequilibrium with 21OH genes owing to the presence of a neuraminidase locus within the class III region.

Both tuberculosis and leprosy are diseases caused by mycobacteria, M. tuberculosis and M. leprae, respectively. Population studies have revealed an HLA association with DR2 (pulmonary TB), DR3 (tubercloid

leprosy) and DQw1 (lepromatous leprosy) (Lamb and Rees, 1988). Analysis of T cell clones indicates that at least 20% of CD4+ lymphocytes in mice inoculated with M. tuberculosis are directed against a member of the HSP 70 family of stress proteins (Young et al., 1988). This may lead to problems in distinguishing between self and non-self analogues, owing to the high degree of homology between species (Lindquist, 1986). One of the major human HSP 70 genes has been mapped to the region of chromosome 6 which includes the MHC (Goate et al., 1987; Harrison et al., 1987) and could be in linkage disequilibrium with the observed associations in the HLA-D region.

1.7 TECHNIQUES

At the start of this project, in 1985, only 120 kb of genomic DNA from the class III region of the MHC had been cloned. This represents, at most, about 12% of the class III region. In order to produce a detailed map of the whole of this region, two techniques were particularly suitable: chromosome walking and pulsed-field gel electrophoresis.

1.7.1 Chromosome Walking

Chromosome walking involves the isolation of a series of overlapping genomic clones from a given chromosomal location. Genomic libraries prepared in cosmid vectors are generally used in preference to bacteriophage λ, owing to the twofold increase in the size of the insert DNA which can be achieved with the former. The walking process can be divided into three stages. Firstly, the isolation of the initial cosmid clones. Secondly, the characterisation of the genomic inserts by restriction endonuclease digestion and Southern blotting. Finally, the

identification of unique DNA sequences, and the preparation of new genomic probes for rescreening the cosmid libraries.

For ease of library construction and molecular mapping of the genomic clones, the DNA used in this study was derived from a homozygous haplotype containing only single copies of the C4 and 21OH genes. This reduced the probability of walking along different sister chromatids.

1.7.2 Pulsed-Field Gel-Electrophoresis (PFGE)

Using standard gel electrophoresis systems, DNA fragments larger than 20 kb in size are difficult to resolve accurately. Pulsed-field gel-electrophoresis (PFGE) is a method which allows the separation of fragments upto 5-10x 10^3 kb in size (Schwartz and Cantor, 1984; Carle and Olson, 1984; Arnand, 1986), which has been used to resolve intact yeast, trypanosome and S.pombe chromosomes. For mammalian systems, the smallest chromosomes are still too large, and PFGE must be combined with a method of generating fragments in the appropriate size range. As the CpG dinucleotide frequency of the vertebrate genome is lower than predicted from the base composition, and most CpGs are methylated, enzymes which recognise only CpG rich unmethylated sequences produce restriction elements suitable for use with PFGE (Lindsay and Bird, 1987; Brown and Bird, 1986; Anand, 1986). In addition, there are some rare cutting endonucleases with 8 bp recognition sequences, such as Sfi I and Not I.

Southern blots of PFGE gels can be made using standard techniques, and probed with unique DNA sequences to produce long-range restriction maps (Brown and Bird, 1986).

Two methods of PFGE were used in this laboratory. Firstly, the orthogonal field system (OFAGE), as described by Brown and Bird (1986)

and secondly, a cross field gel electrophoresis ("Waltzer") gel system, as described by Southern et al. (1987). In the OFAGE system, non-uniform fields are generated using long cathodes and short anodes (Fig. 1.6A) arranged at an angle of approximately 90^{o}. The direction of the field with respect to the migration route of the DNA molecules is altered by switching between the pairs of electrodes. During re-orientation, small molecules rotate more quickly and display a higher mobility than large molecules. By altering the switching times, DNA fragments within given size ranges can be separated. The OFAGE system has the disadvantage that the outer tracks become curved, making it difficult to estimate the sizes of restriction elements. With the "Waltzer" apparatus, much straighter tracks can be obtained, increasing the accuracy of the long-range maps generated from the data. This system uses a uniform field. Again, the DNA molecules are subjected to alternating perpendicular fields, but the switching time controls the movement of the rotating platform, on which the gel rests (Fig. 1.6B).

1.8 AIMS OF THE PROJECT

The aims of the project were as follows:

(a) to define the orientation of the known class III genes within the human MHC,

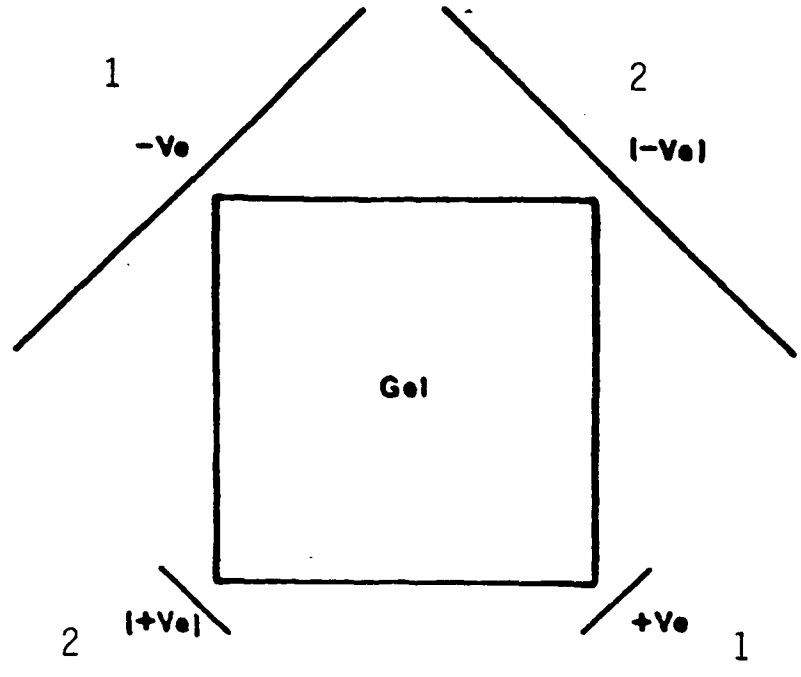
(b) to produce a detailed restriction map of the class III region by characterisation of overlapping genomic cosmid clones isolated by chromosome walking,

(c) to search for novel products of the class III region.

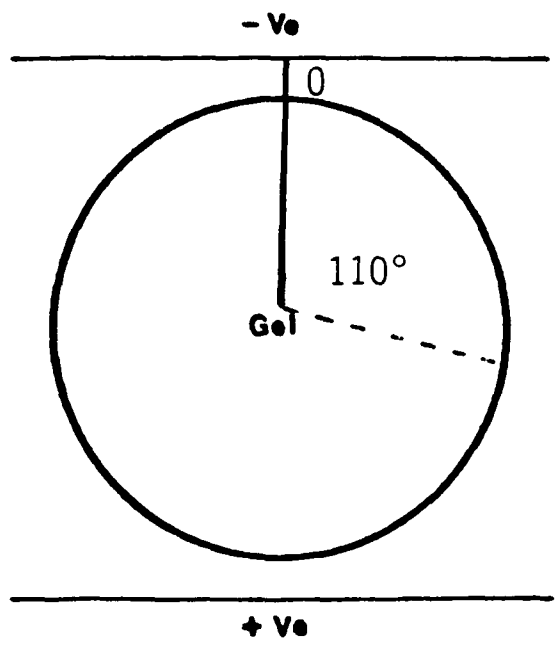
With the latter, it was hoped that the analysis of novel loci would aid in elucidating the relationship between disease association with the MHC, especially by trying to define a possible molecular basis.

Fig. 1.6

The basic organisation of the OFAGE (A) and "Waltzer" gel systems.In OFAGE, the direction of the field is altered by switching between the pairs of electrodes numbered 1 and 2. With the "Waltzer" gel apparatus, the gel is rotated through 110° relative to the horizontal line marked 0, then returned to the original position. DNA molecules of different size ranges can be separated by altering the switching time, or the intervals between gel rotations.



A The OFAGE gel system



B The "Waltzer" gel system

CHAPTER II

MATERIALS AND METHODS

2.1.1 Enzymes

All restriction enzymes were purchased from Amersham, BRL, New England Biolabs, or Boehringer and used according to the supplier's recommendations.

Lysozyme and RNase were obtained from Sigma; T4-DNA ligase and the Klenow fragment of *E.coli* DNA polymerase I from Amersham; T4 polynucleotide kinase and calf intestinal phosphatase from Boehringer; and proteinase K from BDH.

Enzymes and reagents for radiolabelling double stranded DNA probes were obtained from Amersham ("Multiprime labelling kit" and "Nick translation kit")

2.1.2 Chemicals

All chemicals (analar quality) were obtained from BDH, with the following exceptions:

Agarose (Seakem and Seaplaque)	ICN Biomedicals
Bacto-agar	Difco
Bactotryptone	Difco
Bind silane	LKB
Caesium chloride	Sigma
Dextran sulphate	Pharmacia
DEAE-Cellulose (DE52)	Whatman
Ficoll	Pharmacia
β-mercaptoethanol	Fluka

Polyvinylpyrrolidone	Sigma
Sephadex G-100	Pharmacia
Tris	Boehringer
Yeast extract	Oxoid

2.1.3 Radioactive Nucleotides

Radioactive nucleotides were bought from Amersham International. [α-^{32}P]-dATP and [α-^{32}P]- dCTP were supplied at 10 mCi/ml, specific activity 3000 Ci/mmol.

2.1.4 Bacterial Strains and Vectors

All bacterial strains were derivatives of Escherichia coli K12.

JM 103	Messing et al., 1981
BHB 2688	Hohn and Murray, 1977
BHB 2690	"
MC 1061	Casadaban and Cohen, 1980
1046	Cami and Kourilsky, 1978
NM 554	gift of N. Murray

vectors:

pAT 153/Pvu II/8	Anson et al., 1984
pDVcos	gift of G. Brownlee
pUC18	Vieira and Messing, 1982
M13mp8	Messing and Vieira, 1982

2.1.5 Media

2x TY — 15 g bactotryptone, 10 g yeast extract, 5 g NaCl, to 1l in water. Autoclave before use.

L broth — 10 g bactotryptone, 5 g yeast extract, 10 g NaCl, 2.46 g $MgSO_4$ to 1l in water. Autoclave before use.

HMF 1 — 10 g bactotryptone, 5 g yeast extract, 10 g NaCl. 2.46 g $MgSO_4$, 50ml glycerol, 1.79 g KH_2PO_4, 6.27 g K_2HPO_4, 5.88 g tri-sodium citrate, 15 g bacto-agar to 1 l in water. Autoclave before use.

HMF 2 — 10 g bactotryptone, 5 g yeast extract, 10 g NaCl. 2.46 g $MgSO_4$, 250ml glycerol, 1.79 g KH_2PO_4, 6.27 g K_2HPO_4, 5.88 g tri-sodium citrate, 15 g bacto-agar to 1 l in water. Autoclave before use.

Agar plates were prepared using 15 g bacto-agar/l broth. Ampicillin was prepared as a stock solution at 20 mg/ml in 1 M sodium bicarbonate. Liquid cultures routinely contained ampicillin at 50 µg/ml. All agar plates for cosmid colonies contained 50 µg/ml, and plates for plasmid recombinants 100 µg/ml.

2.1.6 Standard Solutions

40% acrylamide stock solution - 38 g acrylamide, 2 g N,N'-methylenebis--acrylamide to 100 ml in water. Deionised 30 min with 5 g "Amberlite" MB-3 monobed

resin, filtered and stored at 4°C.

4% native gel mix - 5 ml 40% acrylamide, 5 ml 10x TBE, to 50 ml in water. Polymerise with 400 µl 10% (w/v) APS and 40 µl TEMED.

0.5x TBE gel mix - 15 ml 40% acrylamide, 5 ml 10x TBE, 46 g urea to 100 ml in water. Polymerise 40 ml with 200 µl 10% APS, 80 µl TEMED.

1.0x TBE gel mix - 15 ml 40% acrylamide, 10 ml 10x TBE, 46 g urea to 100 ml in water. Polymerise 40 ml as above.

2.5x TBE gel mix - 15 ml 40% acrylamide, 25 ml 10x TBE, 46 g urea, 5 g sucrose, trace of bromophenol blue to 100 ml in water. Polymerise as above.

Buffer Q1 - 30 µl 1 M Tris-HCl, pH 7.5, 3 ml 100 mM spermidine, pH 7, 90 µl 1 M $MgCl_2$, 300 µl 250 mM ATP, pH 7, 10 µl β-mercaptoethanol to 5 ml in water.

Buffer A - 20 µl 1 M Tris-HCl, pH 8, 5 µl 1 M $MgCl_2$, 50 µl 1% β-mercaptoethanol, 10 µl 1 M EDTA, pH 7.5 to 1 ml in water.

50x Denhardt's solution - 1% ficoll, 1% polyvinylpyrrolidone, 1% BSA (w/v, in water).

5x glycerol loading dyes - 10 mg bromophenol blue, 10 mg xylene cyanol, 3 ml glycerol, 100 µl 0.5 M EDTA pH 8, to 10 ml in water.

Hogness Buffer - 4% (v/v) glycerol, 3.6 mM Na_2HPO_4, 1.3 mM NaH_2PO_4, 2 mM tri-sodium citrate, 1 mM $MgSO_4$ pH 7.5.

10x MOPS - 41.83 g 3-(N-morpholino)-propanesulphonic acid, 4.10 g NaAc, 3.72 g EDTA to 1l in water. Adjust to pH 7.5 with 5 M NaOH and autoclave before use.

20x SSC - 702 g NaCl, 309 g tri-sodium citrate, to 4 l in water.

10x TBE - 109 g Tris base, 55.6 g boric acid, 9.3 g EDTA to 1l in water. Autoclave before use.

Phenol for extraction of DNA from aqueous solutioins was prepared by dissolving in 1 M Tris-HCl solution, pH 8.0, then sequentially equilibrating in 0.1 M Tris-HCl, pH 8.0, and 10 mM Tris-HCl, pH 8.0. Aliquots were stored under 10 mM Tris-HCl, pH 8.0, at -20^{o}C.

2.2 BACTERIAL CULTURES

Overnight cultures of bacteria were set up from single colonies picked into 10 ml of the appropriate liquid culture medium, supplemented with antibiotic where necessary. The cultures were grown up at 37^{o}C with aeration for a minimum of 12 hours.

Stocks were prepared from 1 ml of overnight culture, as follows: the cells were pelleted in a 1.5 ml Eppendorf tube by centrifugation (1 min, 12,000 g) then resuspended in liquid culture medium containing 1x Hogness buffer. Bacterial stocks were maintained at -70^{o}C for long term storage.

2.3 PREPARATION OF NUCLEIC ACIDS

2.3.1 Recovery of Nucleic Acids From Aqueous Solution

(a) organic extraction: an equal volume of phenol equilibrated in 10 mM Tris-HCl, pH 8.0, was added to the aqueous solution. The mixture was vortexed (5 seconds), centrifuged for 3 min at 12,000 g, and the aqueous layer transferred to a fresh tube. The aqueous solution was sequentially extracted with an equal volume of phenol/chloroform/isopropanol (25:24:1, v/v), and chloroform/ isopropanol (24:1, v/v) (when required), prior to ethanol precipitation.

(b) ethanol precipitation: one-tenth volume of NaAc, pH 7.0, and 2.5 volumes of cold absolute alcohol were added to one volume of aqueous solution. The nucleic acids were precipitated at -70^{o}C for 30 min, or overnight at -20^{o}C, prior to recovery by centrifugation.

2.3.2 Preparation of High Molecular Weight Genomic DNA from Whole Blood/ Cell Pellets (Bell et al., 1981)

DNA was prepared from 10 ml whole blood, or from tissue culture cell pellets. Whole blood was centrifuged at 2,200 r.p.m. for 10 min (MSE-6L) to separate plasma from the buffy coat, which was removed to a fresh tube for DNA extraction. Tissue culture cell pellets contained $\sim 10^7$ cells.

Forty ml of lysis buffer (10 mM Tris-HCl, pH 7.4/ 5 mM $MgCl_2$/ 1% Triton X-100/ 0.32 M sucrose) was added to the cells, and the suspension incubated at 0^oC for 15 min. Following centrifugation at 2,000 r.p.m. and 4^oC for 10 min (MSE-6L) the supernatant was removed, and the pellet resuspended in a small volume of lysis buffer. The volume was increased to 40 ml and the cells incubated on ice for 15 min, then centrifuged as before. The pellet was resuspended in 5 ml SET (150 mM NaCl/ 5 mM EDTA/ 50 mM Tris-HCl, pH 8.0). The volume was increased to 10 ml SET containing 0.5% SDS and 200 µg/ml proteinase K (BDH, fungal) (final concentrations) prior to incubation at 37^oC for 12 h.

Protein was extracted from the solution as follows. One volume of phenol was added, mixed with the aqueous phase, and the mixture was allowed to stand at room temperature for 5 min. The organic layer was separated from the DNA solution by centrifugation at 2,000 r.p.m. and 4^oC for 15 min (MSE-6L), then removed with a syringe. The DNA was similarly extracted once more with phenol, twice with phenol/ chloroform/ isopropanol (25:24:1, v/v), and once with chloroform/ isopropanol (24:1, v/v). The aqueous layer was removed to a fresh tube. The DNA was precipitated with three volumes of absolute ethanol (pre-chilled to -20^oC) by gentle agitation of the sample tube. The precipitate was removed with a sealed capillary tube, washed in 70%

alcohol (v/v, in water), air-dried briefly and resuspended in 10 mM Tris-HCl/ 0.1 mM EDTA, pH 7.4 at 4°C. Once in solution the DNA was dialysed against 10 mM Tris-HCl/ 0.1 mM EDTA, pH 7.4, for 24 h at 4°C.

The DNA concentration was estimated at A_{260}, assuming that one absorbance unit is equivalent to 50 µg/ml DNA.

2.3.3 Large Scale Preparation of Plasmid DNA (Radloff et al., 1967; Maniatis et al., 1982)

Large scale preparations of plasmid DNA were performed by ethidium bromide/ caesium chloride density gradient centrifugation. A starter culture was prepared from a single bacterial colony picked into 20 ml L-broth/ ampicillin and grown at 37°C overnight.

Two 2 l flasks, each containing 500 ml L-broth/ ampicillin, were inoculated with 5 ml of starter culture and grown at 37°C, with aeration, until A_{600} reached 0.85-0.9. Cultures were maintained at 37°C for a further 12-14 h in the presence of 200 µg/ml chloramphenicol, to allow amplification of the plasmid DNA.

The cultures were centrifuged at 5,000 r.p.m. for 10 min (Sorvall RC-5B) and the pellets recombined in 20 ml (one volume) of lysis solution (50 mM glucose/ 10 mM EDTA/ 25 mM Tris-HCl, pH 8.0) containing 200 µg/ml lysozyme. After incubation on ice for 30 min, two volumes of 0.2 M NaOH/ 1% SDS were added, and the sample kept on ice for 5 min. This was followed by addition of 1.5 volumes of 3 M NaAc, pH 4.8, and left on ice for 30 min. The supernatant was cleared by centrifugation at 8,000 r.p.m. and 10°C for 20 min (Sorvall RC-5B) and transferred to a silanised flask. The DNA was precipitated by adding twice the total volume of cold ethanol, and leaving at -20°C for 2 h.

The DNA was recovered by centrifugation in silanised Corex tubes at

13,000 r.p.m. and 4°C for 10 min (Sorvall RC-5B), dried *in vacuo*, and the pellets recombined in a total volume of 6 ml 50 mM Tris-HCl, 150 mM NaCl, 5 mM EDTA, pH 8.0 containing 200 µl 0.5% ethidium bromide. Eight grammes of caesium chloride (Sigma, optical grade) were added, and the solution allowed to return to room temperature. The density was checked by weighing 1 ml, to ensure it lay within 1.61-1.70 g/ml.

The solution was centrifuged at 55,000 r.p.m. and 18°C for 16 h (Ti70.i rotor, Beckman L-6 ultracentrifuge) in a sealed polyallomer tube overlayered with paraffin, and relaxed to 45,000 r.p.m. for one hour prior to stopping without a brake. The DNA band was removed from the tube using a syringe. The ethidium bromide was extracted five times by mixing with an equal volume of butan-1-ol. The final sample was diluted with 3 volumes of de-ionised water, and the DNA precipitated with twice the total volume of ethanol at -20°C for four hours. The precipitate was recovered by centrifugation at 10,000 r.p.m. and 4°C for 30 min (Sorvall RC-5B) in silanised Corex tubes and resuspended in 400 µl 0.3 M NaAc, pH 6.0. The solution was transferred to a 1.5 ml Eppendorf, ethanol precipitated, and the DNA resuspended in 10 mM Tris-HCl/ 1 mM EDTA, pH 7.4. Concentrated solutions of DNA prepared by this method were held at 4°C for long term storage.

2.3.4 Small Scale Preparations of Plasmid/ Cosmid DNA (Birnboim and Doly, 1979)

Ten ml of overnight culture was started from a single bacterial colony. One ml was retained for making Hogness buffer stocks (2.2). The remaining culture was centrifuged at 2,000 r.p.m. and 4°C for 30 min (MSE-6L) and the pellet resuspended in 200 µl (1 volume) of 50 mM glucose/ 10 mM EDTA/ 25 mM Tris-HCl, pH 8.0 before transfer to a 1.5 ml

Eppendorf. The cells were lysed at room temperature for 10 min. The solution was then treated with 2 volumes 0.2 M NaOH/ 1% SDS for 5 min on ice, and 1.5 volumes 3 M NaAc, pH 4.8 for 10 min on ice. The supernatant was cleared of cellular debris and genomic DNA by two centrifugation steps, each of 5 min. Plasmid/ cosmid DNA was precipitated at -70^{o}C for 10 min by adding an equivalent volume of isopropanol, and centrifuging. The pellet was resuspended in 200 µl 0.3 M NaAc, pH 7.0, and extracted once with phenol. DNA was recovered from the aqueous phase by precipitation at -70^{o}C for 10 min with an equal volume of isopropanol. After centrifugation, the pellet was washed with 70% ethanol (v/v, in water), dried *in vacuo*, and resuspended in 50 µl 10 mM Tris-HCl/ 0.1 M EDTA, pH 7.4. RNA was removed by incubation with 1 µl heat inactivated RNase (stock solution 1 mg/ml) at 37^{o}C for 30 min prior to storage at -20^{o}C.

2.3.5 *Preparation of RNA* (Chirgwin *et al.*, 1979; Maniatis *et al.*, 1982)

All glassware was autoclaved for at least 30 min, or baked at 100^{o}C overnight, before use. Samples containing RNA were handled with gloves to avoid contamination with RNases from the skin.

Pellets of 2-4x 10^{7} cells from tissue culture (prepared by B. Moffat) were washed 2-3 times with 20 ml PBS (phosphate buffered saline, Flow Laboratories). The final pellet was resuspended in 3 ml freshly prepared 4 M guanidinium isothiocyanate/ 14% (v/v) β-mercaptoethanol. At this stage the cell suspension could be held at room temperature prior to caesium chloride ultracentrifugation. An equal volume of 5.7 M caesium chloride/ 25 mM tri-sodium citrate/ 100 mM EDTA was added to the suspension and mixed by inversion. A 3 ml cushion was placed in the bottom of an ultraclear centrifugation tube. The 6 ml of cell suspension

was layered on top, and the tube filled with liquid paraffin before heat sealing. The sample was centrifuged at 20°C and 35,000 r.p.m. for 16 h in a Beckman L8 ultracentrifuge (no brake). The supernatant was removed carefully from the RNA pellet. The pellet was washed with 70% ethanol (v/v, in water) and dissolved in 400 µl TES (25 mM Tris-HCl/ 1 mM EDTA/ 0.2% SDS, pH 7.5. To estimate the RNA concentration, 1 µl of solution in 1 ml of water was monitored at A_{260}. One absorbance unit was taken to be equivalent to 40 µg of RNA.

The RNA was stored as an ethanol precipitate at -20°C.

2.4 RESTRICTION DIGESTS

Restriction enzyme digests were carried out according to the supplier's recommended conditions. All buffers containing >50 mM NaCl also contained 5 mM spermidine. Cosmid or plasmid DNA was digested to completion in 1-2 h using a fivefold excess of enzyme. Analytical digests were performed using 50 ng- 300 ng of DNA/ 10 µl reaction volume. Preparative digests were scaled up to contain 5-10 µg of material.

Genomic digests were carried out using 20-30 units of enzyme/ 5 µg of DNA for a minimum of 5 h.

2.5 FRACTIONATION OF NUCLEIC ACIDS AND RECOVERY OF DNA FOLLOWING GEL ELECTROPHORESIS

2.5.1 Separation of DNA by Agarose Gel Electrophoresis (Southern, 1975)

Restriction enzyme digested genomic DNA samples were fractionated on submerged horizontal slab gels 24 cm x 20 cm x 0.6 cm prepared using 300

ml 0.7% agarose (Seakem, HGT) (w/v, in 1x TBE running buffer). Samples were loaded in 1x glycerol dyes and electrophoresed at 45 mA for 14 h in 1x TBE. For visualisation of DNA, gels were stained in 0.001% ethidium bromide for 1 h and destained for 30 min in water.

Digests of plasmid/ cosmid DNA were fractionated on 0.8% agarose/ 0.01% ethidium bromide w/v, in 1x TBE horizontal slab gels. For reaction volumes >50 µl, the conditions were as above. For samples of 10 µl or less, minigels (11 cm x 10 cm x 0.4 cm) were prepared and run at 50 mA for 2-3 h.

Ethidium bromide stained gels were visualised at 300 nm on a Fotodyne transilluminator. Photographs were taken using a Polaroid MP4 camera fitted with a red filter and Polaroid 667 film.

2.5.2 Separation of DNA by Polyacrylamide Gel Electrophoresis

For analytical and preparative purposes end-filled radiolabelled DNA was fractionated on 4% polyacrylamide gels. Gels were prepared between glass plates, one of which was treated with 2 % dichlorodimethyl silane (v/v, in 1,1,1-trichloroethane) for easy removal. Gels were run at 25 mA for 2-3 h in 1x TBE running buffer. After electrophoresis the gels were wrapped in cling film and the DNA bands observed by autoradiography at room temperature (2.7.6).

2.5.3 Fractionation of RNA (Lehrach et al, 1977; Fourney et al., 1988)

Fifteen µg of RNA was recovered from ethanol by centrifugation and resuspended to a total volume of 5 µl in 25 mM Tris-HCl/ 10 mM EDTA, pH 7.4. The sample was denatured by heating to 65^{o}C for 15 min in 25 µl electrophoresis buffer (1 mM EDTA/ 0.4% bromophenol blue/ 0.4% xylene

cyanol/ 50% glycerol w/v, in water). Ethidium bromide was added to a final concentration of 0.03% (w/v). The denaturing gel was prepared by cooling 300 ml 1% agarose (Seakem, HGT) (w/v, in 1 x MOPS running buffer) to 50^{o}C then adding deionised formaldehyde to a final concentration of 1.9% (v/v). The gel was set in a 24 cm x 20 cm tray for 1 h before use. Samples were run at 40 mA for approximately 18 h in 1x MOPS running buffer.

Following electrophoresis, RNA was visualised on a transilluminator, as described above (2.5.1).

2.5.4 Recovery of DNA from Acrylamide Gel Slices.

The acrylamide slice was removed to a silanised Eppendorf tube and the DNA eluted in 400 µl 0.2 M ammonium acetate solution at 37^{o}C for 14 h. The tube was centrifuged and the aqueous solution transferred to a fresh Eppendorf. The gel slice was washed in a further 200 µl 0.2 M ammonium acetate solution, spun, and the washings combined with the original eluate. DNA was precipitated with 1.5 volumes of ethanol at -70^{o}C for 30 min. After centrifugation, the pellet was resuspended in 300 µl 0.3 M NaAc, pH 7.0, and reprecipitated in the presence of 2.5 volumes of ethanol. The DNA was again recovered by centrifugation, resuspended in 200 µl 0.3 M NaAc. pH 7.0, and reprecipitated. The final pellet was washed with 200 µl 70% ethanol, dried in vacuo, and resuspended in 10 mM Tris-HCl/ 0.1 mM EDTA, pH 7.4.

2.5.6 Recovery of DNA from Low Melting Temperature Agarose

Preparative digests for making probes >0.5 kb were fractionated on 1% LGT agarose (Seaplaque) (w/v, in 1x TBE running buffer) horizontal slab

gels (11 cm x 10 cm x 0.4 cm). Samples were loaded in 2x glycerol loading dyes and run at 5 mA and room temperature for 14 h, or, 30 mA and 4°C for 3-4 h. Agarose slices were removed, weighed, and heated to 70°C in 3x v/w of deionised water. The gel mixture was vortexed, left to cool to room temperature, and extracted as described in 2.3.1 (a). The final aqueous solution was precipitated with ethanol, recovered by centrifugation, washed and dried in vacuo before resuspending in 10 mM Tris-HCl/ 0.1 mM EDTA, pH 7.4.

2.6 TRANSFER OF NUCLEIC ACIDS: BLOTTING PROTOCOLS

2.6.1 Transfer of DNA from Agarose Gels (Southern, 1975)

Standard agarose gels were treated as follows:

(a) twice for 15 min at room temperature in 0.25 M HCl to depurinate the DNA,

(b) once for 30 min at room temperature in 0.5 M NaOH/ 1 M NaCl,

(c) once for 30 min at room temperature in 0.5 M Tris-HCl/ 3 M NaCl.

DNA was transferred onto nitrocellulose membranes (Hybond-C), or nylon backed nitrocellulose (Hybond-Extra), by capillary action, using sponges to enhance the process. The membranes were pre-wetted in 2x SSC before use. Gels were blotted in 20x SSC for 24 h (genomic digests) or 30 min to overnight (cloned DNA digests).

Minigels could be similarly used for Southern blot analysis, but were treated for a maximum of 15 min at each of steps (a), (b), and (c).

DNA was fixed to the nitrocellulose filter by baking at 80°C for 2-2.5 h.

2.6.2 Transfer of RNA from Agarose Gels (Fourney et al., 1988; Thomas, 1980)

RNA gels were treated for 20 min in 1x SSC/ 0.05 M NaOH followed by two 20 min washes in 10x SSC, to increase the transfer efficiency of large RNA species. The RNA was blotted onto nitrocellulose pre-wetted in 10x SSC using 10x SSC transfer buffer and standard capillary blotting procedures.

Ethidium bromide stained RNA could be visualised on the membrane, after transfer, by using a transilluminator. Blots were baked at $80^{o}C$ for 2-2.5 h to fix the RNA to the filter.

2.7 HYBRIDISATION TECHNIQUES

2.7.1 Preparation and Radiolabelling of Double-Stranded DNA Probes

Probes larger than 0.5 kb were prepared from the appropriate restriction enzyme digest of cosmid or subclone DNA by recovery from LGT agarose (2.5.6). Probes smaller than 0.5 kb were prepared by recovery of the correct fragment from 4% acrylamide gels (2.5.4). Concentrations of the samples were estimated from the intensities of the bands observed on ethidium bromide stained agarose gels.

All walking probes were labelled to specific activity > 10^{8}-10^{9} c.p.m./ µg by the random hexanucleotide priming method of Feinberg and Vogelstein (1984). Approximately 20-50 ng was radiolabelled using 20 µCi [α-^{32}P]-dCTP for 3-5 h at room temperature with the Amersham "Multi-priming Kit".

Total genomic DNA was labelled to a specific activity of about 2x 10^{7}

c.p.m./ µg with 10 µCi [α-^{32}P]-CTP and 10 µCi [α-^{32}P]-dATP using the nick translation method of Rigby et al. (1977).

All probes were prepared free from unincorporated nucleotides by elution from a 2 ml Sephadex G-100 column with 200 mM NaCl/ 10 mM Tris-HCl/ 1 mM EDTA, pH 7.4.

2.7.2 Preparation and Radiolabelling Oligonucleotides

Oligonucleotides were synthesised on an Applied Biosystems 381A DNA synthesiser by Dr. T. Day.

Crude ammonia solutions of oligonucleotides were purified by ethanol precipitation at -70^{o}C for 10 min. Following centrifugation the pellet was washed in 80% ethanol (v/v, in water), dried in vacuo, and resuspended in 10 mM Tris-HCl/ 1 mM EDTA, pH 7.4. The sample was mixed with an equal volume of formamide loading buffer (10 mM EDTA/ 0.1% xylene cyanol/ 0.1% bromophenol blue w/v,in formamide), boiled for two minutes and loaded onto a 20% acrylamide/ 4.2 M urea/ 1x TBE denaturing gel prepared between silinated glass plates. The gel was run for 7 h at 500 V in 1x TBE buffer, then removed from the plate and placed between two layers of Saranwrap. The major oligonucleotide band was cut from the gel after visualisation at 254 nm against a TLC plate. The DNA was eluted in 10 ml 0.5 M ammonium acetate/ 10 mM magnesium acetate/ 1 mM EDTA at 37^{o}C for 12 h. The oligonucleotide was recovered from a G_{18} Sep-Pack cartridge (Waters Associates) primed by pre-wetting with 10 ml HPLC grade methanol followed by 10 ml sterile water. The solution containing the oligonucleotide was passed through the column twice, and the eluate retained. The cartridge was washed with 10 ml sterile water, then the oligonucleotide was eluted into a 1.5 ml Eppendorf using 1.5 ml 60% HPLC grade methanol (v/v, in sterile water). The oligonucleotide was

recovered by drying under vacuum, and the DNA resuspended in 10 mM Tris-HCl/ 0.1 mM EDTA, pH 8.0.

To estimate the concentration, 5 µl DNA solution was diluted in 1 ml water and monitored at A_{260}. One absorbance unit was assumed to be equivalent to 37 µg of single stranded DNA.

Twenty-five ng of oligonucleotide was radiolabelled to a specific activity > 10^8 c.p.m./µg. The sample was incubated at 37°C for 30 min in the presence of 50 µCi [α-^{32}P]-dATP and 5 units of T4-DNA kinase.

The probes were separated from free radioactivity by elution from a Sephadex GS-50 superfine column using 200 mM NaCl/ 10 mM Tris-HCl/ 1 mM EDTA, pH 7.4

2.7.3 Removal of Repetitive Sequences from Genomic Probes (Sealey et al., 1985)

Genomic probes containing repetitive sequences (as identified from genomic Southern blots: such probes gave a background smear owing to hybridisation to repetitive elements, but also hybridised to specific fragments of the correct size which could be identified from the molecular map) were boiled for 10 min in 5x SSC with sufficient human genomic DNA to give a final concentration of 3-10 µg/ml of hybridisation fluid. Reassociation of the repeat elements was allowed to occur by cooling to 65°C, and holding at this temperature for 15 min before adding to the hybridisation buffer.

2.7.4 Hybridisation Conditions (Jeffreys and Flavell, 1978; Bernards and Flavell, 1980)

All filters were washed in 0.1 M NaCl/ 1 mM EDTA/ 50 mM Tris-HCl/ 0.1%

SDS, pH 8.0 at 42^{o}C for one hour prior to pre-hybridisation. For hybridisation with walking probes, filters were pre-hybridised in 50% formamide/ 5x Denhardts/ 1 M NaCl/ 10% dextran sulphate/ 50 mM Tris-HCl, pH 7.4/ SDS containing 200 µg/ml salmon sperm DNA and 20 µg/ml sonicated E.coli/ pATX DNA for a minimum of four hours at 42^{o}C.

For hybridisation, probes were boiled for 10 min and snap-cooled on ice before adding to the buffer. Hybridisation was continued for 16-36 h at 42^{o}C.

Filters were washed in 2x SSC at room temperature for 30 min to 1 h, followed by two high stringency washes at 65^{o}C, each for 30 min using 0.1-0.2x SSC/ SDS and a brief wash in 0.1-0.2x SSC.

For nitrocellulose membranes (Hybond-C and Hybond-Extra) 0.1% SDS was used throughout. For pulsed-field blots which had been taken onto nylon (GenescreenPlus, New England Nuclear) all steps included SDS to a final concentration of 1%.

For hybridisation to oligonucleotide probes, filters were pre-hybridised and hybridised in 0.9 M NaCl/ 90 mM Tris-HCl, pH 7.4/ 6mM EDTA/ 10 % dextran sulphate/ 5x Denhardts/ 0.1% SDS. Blots were washed in 6x SSC/ 0.1% SDS for 1 h. All steps, including washing, were carried out at the appropriate temperature, as calculated from the base composition of the probe.

2.7.5 Removal of Probe From Membranes

Blots which were used for successive rounds of screening with different probes were stripped between hybridisations by washing in 2 mM Tris-HCl, pH 7.4/ 1 mM EDTA/ 0.1% SDS at 80^{o}C for upto 2 h. The efficiency of the removal of bound radioactivity was checked by auto-radiography overnight.

2.7.6 Autoradiography

Unless otherwise stated, autoradiography was carried out at -70^{o}C (Laskey and Mills, 1977) between two tungstate intensifying screens (Cronex ultra-lightning plus, Dupont) for 5 h to 14 days. Filters were autoradiographed using X-OMAT "S" X-ray film (Kodak), which was sensitised by pre-flashing before use. Films were developed in a Kodak MEI X-Omat automatic processor.

2.8 SUBCLONING

Many walking probes were subcloned by religation of restriction enzyme digested cosmid DNA to form "minicosmids". Other genomic fragments were ligated into Bam HI cut and phosphatased pATX or Hinc II cut and phosphatased pUC18. All DNA subclones were transfected into MC 1061 competent cells, and retained as Hogness stocks at -70^{o}C.

2.8.1 Preparation of Plasmid Vector DNA

Ten µg of plasmid DNA was cut to completion using 30 units of enzyme at 37^{o}C for 2 h. The DNA was then treated with 5 units of calf intestinal phosphatase at 37^{o}C for 30 min, and the reaction stopped by heating to 70^{o}C for 10 min. The vector was tested for;

(a) background levels of uncut plasmid,

(b) religation to itself following dephosphorylation,

(c) ability to accept insert sequences.

Genomic fragments prepared by digestion with Bam HI and/ or Bgl II were ligated into Bam HI cut pATX. All other restriction digests were

blunt-ended with dNTPs and DNA polymerase Klenow fragment (2.8.2) and ligated into Hinc II cut pUC18.

2.8.2 End Filling (Wu and Taylor, 1971)

Following digestion, restriction fragments were made blunt-ended by 3' end repair using 250 µM dNTPs and 2 U of E.coli DNA polymerase Klenow fragment. Reactions were incubated at room temperature for 30 min, and stopped by heating to 70°C for 10 min.

If digests were to be separated on 4% native gels, 10% of the DNA was end-labelled using 5 µCi [α-^{32}P]-dATP or [α-^{32}P]-dCTP in the incubation mixture.

2.8.3 Ligations

All ligations were performed using 10 ng vector and 5-10 ng insert DNA in 50 mM Tris-HCl/ 10 mM DTT/ 10 mM $MgCl_2$/ 1 mM ATP with BSA to 100 ng/µl.

Incubations were carried out at room temperature for 4-5 h using 1 U T4-DNA ligase in a reaction volume of 10 µl.

2.8.4 Preparation of Competent Cells and Transformation Procedure

A starter culture was prepared by picking a single colony of MC 1061 into 10 ml L-broth and growing at 37°C overnight. Two hundred ml L-broth was inoculated with 2 ml starter culture and grown until A_{600} reached 0.2. The cells were chilled on ice for 10 min, transferred to pre-chilled 50 ml Falcon tubes and harvested by centrifugation at 2,600 r.p.m. and 4°C for 10 min (MSE-6L). The pellets were resuspended in 100 ml 50 mM $CaCl_2$/ 15% glycerol/ 10 mM Pipes, pH 6.6, and left on ice for

20 min. The cells were centrifuged as before, resuspended in 16 ml calcium chloride buffer, and aliquoted into pre-chilled Eppendorf tubes. After 40 min on ice, the cells were snap-frozen on dry ice/ IMS, then stored at -70^{o}C.

Competent cells were thawed on ice, then 100 µl added to each of the ligation tubes. The mixture was incubated on ice for 10 min. Samples were heat shocked at 37^{o}C for 5 min, diluted fourfold with L-broth, and maintained at 37^{o}C for 30-60 min. Aliquots of 200 µl were plated onto L-agar plates containing ampicillin to 100 µg/ml, and incubated at 37^{o}C overnight.

Efficient transformation figures were in the region of 10^{7} colonies per µg uncut plasmid DNA.

2.9 DNA SEQUENCING

2.9.1 Subcloning

Double-stranded DNA fragments were prepared by restriction digestion of cDNA or genomic clones, blunt-ended (2.8.2) and recovered from 4% native acrylamide gels (2.5.2/2.5.4). The fragments were then ligated into 20 ng Sma I cut and phosphatased M13mp8 (Amersham International) using the standard protocol (2.8.3).

2.9.2 Preparation of competent cells

Competent cells were prepared by inoculating 300 µl of an overnight culture of JM 103 into 30 ml sterile 2x TY broth and growing at 37^{o}C until A_{600} reached 0.4-0.6 (2-3 h). All subsequent steps were carried out at 4^{o}C.

The culture was centrifuged at 1,600 r.p.m. for 5 min (MSE-6L) and the pellet resuspended in 10 ml 50 mM $CaCl_2$/ 10 mM Tris-HCl, pH 7.5. The cells were held on ice for 30 min, centrifuged as before, and resuspended in 2-5 ml of $CaCl_2$ solution.

The competent cells were stored at 4°C, and used within 24 h.

2.9.3 Transfection Procedure

Two hundred µl of competent cells was added to the 10 µl ligation mixture, and incubated on ice for 45 min. The cells were heat-shocked at 45°C for 5 min, then rapidly transferred to sterile tubes containing 3 ml top agar, 25 µl BCIG (25 mg/ml in dimethylformamide) 25 µl of IPTG (25 mg/ml in water) and 500 µl of overnight JM 103. The mixture was poured onto a 2x TY agar plate and incubated at 37°C overnight. Vector containing insert formed white plaques in the presence of the chromogenic substrate (BCIG).

2.9.4 Phage Growth and DNA Purification

Each positive plaque was grown with vigorous shaking at 37°C for 5 h in 1.5 ml of a one hundredth dilution of overnight JM 103. The culture was centrifuged twice and the supernatant transferred to a fresh Eppendorf at each step. Approximately 200 µl was retained as a stock after the second spin. The remaining supernatant was treated with 200 µl of PEG (20% polyethylene glycol in 2.5 M NaCl) and left on ice for 20 min. After centrifugation, the phage pellet was resuspended in 100 µl 10 mM Tris-HCl/ 0.1 mM EDTA, pH 8.0. The aqueous solution was extracted with an equal volume of phenol/ chloroform/ isopropanol (25:24:1, v/v), and precipitated with 0.1 volume 3 M NaAc and 1 volume of isopropanol at

-70°C for 30 min. The DNA was washed with 70% ethanol (v/v in water), dried *in vacuo* and resuspended in 50 µl 10 mM Tris-HCl/ 0.1 mM EDTA, pH 8.0.

2.9.5 Characterisation of M13mp8 Clones

DNA samples from clones containing the same restriction fragment insert were incubated together under annealing conditions to determine their relative orientations. Hybridisations were carried out using 2 µl tester DNA, 2 µl sample DNA, 4 µl buffer (600 mM NaCl/ 70 mM $MgCl_2$/ 60 mM Tris-HCl, pH 7.4.) and 1 µl of glycerol dye (50% glycerol/ 1% SDS/ bromophenol blue) at 65°C for 1 h. After chilling on ice, samples were characterised on an agarose minigel. If the tester and sample DNA molecules contained inserts of the opposite orientation two prominent bands were observed; if they were in the same orientation, only one band was seen.

2.9.6 DNA Sequencing (Sanger *et al.*, 1977, 1980; Biggin *et al.*, 1980)

DNA samples were sequenced by extension from the M13 universal primer (5'-GTAAAACGACGGCCAGT-3', Amersham International) (Duckworth *et al.*, 1981) in the presence of dideoxynucleoside triphosphate chain-terminating inhibitors, as described by Sanger *et al.* (1977).

For each set of sequencing reactions, 8 µl of sample DNA was annealed to 1.5 µl M13 primer (2.5 ng/µl) in the presence of 1.5 µl 100 mM Tris-HCl/ 50 mM $MgCl_2$, pH 8.4 at 65°C for 1 h. Two µl of template DNA was used per reaction (T,C,G,or,A) with 2 µl of the appropriate ddNTP mix and 2 µl of Klenow mix (2.5 µCi [α-^{32}P]-dATP/ 1µl 0.1 M DTT/ 0.5µl 2.5 mM dATP/ 4 U Klenow/ 6µl 10 mM Tris-HCl, pH 8.0, per clone). The

reaction was incubated at room temperature for 15 min, then 2 µl of 0.5 mM dNTP was added for the chase reaction of 15 min at room temperature. Finally, the DNA duplexes were denatured in the presence of 4 µl formamide loading dye (10 mM EDTA/ 0.1% xylene cyanol/ 0.1% bromophenol blue w/v, in formamide) by boiling for 5 min.

Samples were run on 6% acrylamide/ 8 M urea gels. Short runs (30 mA, 1 h 35 min) were performed on gradient buffer gels prepared essentially as described by Biggin et al. (1980) using 2.5x and 0.5x TBE gel mixes. 0.5x TBE running buffer was used in the cathodal reservoir, and 1.0x TBE in the anodal reservoir. Long gel runs (30 mA, 4-5 h) were carried out using 1.0x TBE gel mixes and running buffers.

Pretreatment of back plates with 2.5 ml ethanol/ 70µl 10% acetic acid (v/v, in water)/ 7 µl Bind Silane (LKB) allowed gels to be fixed in 10% acetic acid/ 10% methanol (v/v, in water) for 15 min at room temperature, washed (15 min, in water), and dried directly onto the glass at 110^{o}C. Autoradiography was carried out at room temperature in the absence of an intensifying screen for 12 h to 3 days.

ddNTP mixes were prepared from stock solutions of 0.5 mM dNTPs and 10 mM ddNTPs as follows:

A mix: 500µl each of dTTP, dCTP, dGTP, and 10 mM Tris/ 0.1 mM EDTA, pH 8.0, with 3µl ddATP.

C mix: 500µl each of dTTP, dGTP, 25µl dCTP, 8µl ddCTP, and 1ml 10 mM Tris/ 0.1 mM EDTA, pH 8.0.

G mix: 500µl each of dTTP, dCTP, 25µl dGTP, 16µl ddGTP, and 1ml 10 mM Tris/ 0.1 mM EDTA, pH 8.0.

T mix: 500µl each of dCTP, dGTP, 25µl dTTP, 50µl ddTTP, and 1ml 10 mM Tris/ 0.1 mM EDTA, pH 8.0.

2.9.7 Analysis of Sequence Data (Staden, 1982a, 1982b, 1984)

Sequence data were analysed using a VAX II/780 computer. DNA nucleotide sequences were read from gels and entered by hand. The data were analysed using DBSEARCH and ANALYSEQ programs.

2.10 PULSED FIELD GEL ELECTROPHORESIS

2.10.1 Preparation of Chromosomal DNA in Agarose Blocks (van Ommen and Verkerk, 1986; Dunham et al., 1987)

(a) purification of agarose with DEAE-cellulose: (Jackson and Cook, 1985) DE-52 was pre-swollen in sterile water, then equilibrated with 5x PBS by washing four times and finally mixing in an equal volume of 5x PBS containing 0.1 mM EDTA.
To purify, 2% agarose (w/v, Sigma type VII in sterile water) was melted by boiling and held at 50^{o}C. Ten ml of DE-52 suspension was spun down to give 5 ml packed DE-52 in a 50 ml Falcon tube. To this was added 45 ml agarose. After mixing,the agarose was held at for 30 min, spun down, and removed to a fresh tube containing 5 ml packed DE-52. This process was repeated a further three times. The agarose was stored at 4^{o}C.

(b) Tissue culture cells were harvested and washed twice with 20 ml PBS. The cells were resuspended to a final concentration of 2x 10^{7} cells/ ml using a 1:1 mixture of 1x PBS and 2% DE-52 purified agarose. Blocks were made by setting the agarose on ice in a perspex former with cell dimensions of 9 mm x 7 mm x 2 mm.
The blocks were digested in 10 ml 10 mM Tris-HCl/ 0.5 M EDTA/ 1% sarkosyl, pH 9.5 (NDS solution) containing 1 mg/ml proteinase K

(Boehringer) at 50°C overnight. They were further incubated in fresh NDS/ proteinase K for 24-48 h. After digestion, the blocks were washed twice, for 30 min each time, at 50°C in NDS buffer, and stored in NDS at 4°C.

2.10.2 Restriction Enzyme Digests

All restriction enzyme digests were carried out according to the supplier's recommendations. Agarose blocks containing approximately 5 µg DNA were washed three times in 10 ml 10 mM Tris-HCl/ 0.1 mM EDTA/ 0.1 mM phenylmethylsulphonyl fluoride, pH 8.0 at 4°C for 30 min, and once in restriction enzyme buffer without DTT. Each block was transferred to an Eppendorf containing 100 µl of 1x restriction buffer with 5 mM DTT and bsa to 10 µg/ml (for buffers with > 50 mM NaCl, digests also included 5 mM spermidine) and incubated with 20 U enzyme at 37°C for 2 h. For double digests, the blocks were rewashed with Tris/ EDTA solution and restriction enzyme buffer prior to the addition of the second enzyme.

2.10.3 Fractionation of DNA on Agarose Gels (Carle and Olson, 1984; Southern et al., 1987)

Agarose blocks were cut in half after digestion with restriction enzymes and loaded directly into the sample wells. Molecular weight markers (gift from I. Dunham) consisted of intact yeast chromosomes (Saccharomyces cerevisiae, strain X2-180-1B) and concatamers of bacteriophage λ c1857 S7.

Both orthogonal field alternation gel electrophoresis (OFAGE) (Carle and Olson, 1984) and "Waltzer" (Southern et al., 1987) were used. For OFAGE the samples were loaded in 0.5x TAE (20 mM Tris acetate/ 1 mM

EDTA, pH 8.5) / 1x glycerol dyes into 14 mm x 4 mm x 1.5 mm wells on a 20 cm x 20 cm x 0.5 cm 1.5% agarose (Sigma type I) gel and electrophoresed in 0.5x TAE running buffer at 12^{o}C and 330V for 22 h.

For the "Waltzer" system, the blocks were loaded in 0.5x TAE/ 1x glycerol dyes into 7 mm x 4 mm x 2 mm wells on a 22 cm diameter x 0.5 cm 1.5% agarose (Sigma, type I) gel. The DNA was electrophoresed in 0.5x TAE running buffer at 18^{o}C and 150 V for 30-36 h.

Gels were stained in 0.5x TAE containing 0.01% ethidium bromide for 30 min, and destained for up to 1 h, prior to visualisation on a transilluminator, and photography, as previously described (2.5.1).

2.10.4 Transfer of DNA by Southern Blotting

DNA was transferred as described in section 2.6.1 onto nylon membranes (GenescreenPlus, New England Nuclear). The membranes were dampened on water and then soaked in 10x SSC for 30 min before use. The filter was placed with side "B" in contact with the gel surface, as recommended by the manufacturer. After blotting, DNA was fixed to the nylon as follows:

(a) the DNA was denatured in 0.4 M NaOH for 30 s,

(b) the membrane was neutralised in 0.2 M Tris, pH 7.4/ 2x SSC for 60 s,

(c) the DNA was crosslinked to the nylon by ultra violet irradiation at 254 nm for 4 min (CAMAG Universal lamp, fixed at 15 cm from the membrane surface).

2.11 PREPARATION OF COSMID LIBRARIES (Grosveld et al., 1981; Steinmetz et al., 1986)

2.11.1 Preparation of Vector Arms

Approximately 100 ug of cosmid pDVcos DNA was digested to completion with 150 U Pvu II at 37°C for 2 h. The DNA was then phenol extracted and ethanol precipitated according to standard procedures.

The DNA was resuspended in 100 µl of phosphatase buffer (50 mM Tris-HCl, pH 9.0/ 1 mM $MgCl_2$/ 0.1 mM $ZnCl_2$/ 1 mM spermidine) and treated twice with 2 U calf intestinal phosphatase at 37°C for 30 min. The reaction was stopped by adding EDTA to a final concentration of 15 mM and incubating at 70°C for 10 min.

Protein was removed from the aqueous phase by organic extraction , and pDVcos recovered by ethanol precipitation. The DNA was resuspended in 1x Bam HI buffer. Arms were prepared by digestion with 100 U Bam HI at 37°C for 2 h. The DNA was extracted and precipitated as before. It was resuspended in 10 mM Tris-HCl/ 0.1 mM EDTA, pH 8.0, to a final concentration of 0.5 µg/µl prior to long term storage at -20°C.

2.11.2 Preparation of Insert DNA

High molecular weight genomic DNA was prepared as described in 2.3.1. Digests, followed by phosphatase treatment, were performed in a low salt Mbo I buffer (Ish-Horowitz and Burke, 1981) using 60 µg DNA in a reaction volume of 300 µl. Five digests, with a concentration range of 40, 20, 10, 5, and 2.75 U Mbo I, were incubated at 37°C for 1 h, then stopped by heating to 70°C for 15 min. One and a half units of calf intestinal phosphatase were added to each tube, and dephosphorylation

carried out for 15 min at 37°C. The reaction was stopped by addition of EDTA to a final concentration of 15 mM, and heating to 70°C for 15 min. Digests were checked by running samples of 30 µl on a 0.4% agarose gel (w/v, in 1x TBE) against a λ Xho I (14 and 35 kb) standard.

The digested and phosphatased DNA was recovered by phenol extraction and ethanol precipitation. The precipitate was recombined in a total volume of 300 µl prior to sucrose density gradient centrifugation.

2.11.3 Size Fractionation of Insert DNA

The digested and phosphatased DNA was size-fractionated on a stepped sucrose density gradient. Each step consisted of 6 ml solution, with a range of 45-20% sucrose (w/v, in 1 M NaCl/ 20 mM Tris-HCl/ 10 mM EDTA, pH 7.5), in 5% increments. The DNA was layered onto the surface of the gradient and covered with a film of paraffin. Centrifugation was performed at 26,000 r.p.m. and 20°C for 20 h using an SW 27 rotor in a Beckman L5-65 ultracentrifuge (no brake).

Samples of 0.7 ml were collected from the bottom of the gradient at a flow rate of 36 ml/ h. From each fraction, 15 µl were run on a 0.4% agarose gel (w/v, in 1x TBE) at 75 mA for 12 h against bacteriophage λ (49 kb) and λ Xho I (35 and 14 kb) markers prepared in 30% sucrose. Samples containing DNA in the size range 35-50 kb were precipitated in ethanol at -20°C for 4 h to overnight. The DNA was recovered by centrifugation and resuspended in 100 µl 10 mM Tris-HCl/ 0.1 mM EDTA, pH 8.0.

The concentrations of the DNA samples were estimated by spotting 1:1 serial dilutions (in water) onto ethidium bromide/ agarose plates (1 µg ethidium bromide/ ml 0.8% agarose), and comparing with bacteriophage λ standards.

2.11.3 Preparation of Packaging Extracts (Grosveld et al., 1981)

Two complementary, temperature sensitive mutants were used in the preparation of packaging extracts. The infected E.coli strains BHB 2688 and BHB 2690 were streaked out onto 2x TY plates and tested at permissible (32°C) and non-permissible (42°C) temperatures, prior to use.

(a) freeze-thaw lysate: a starter culture of 3 ml was set up from a single colony of BHB 2688 in L-broth, and grown at 32°C overnight. Two flasks with 500 ml L-broth were each inoculated with 1 ml starter culture and grown at 32°C with aeration until A_{600} reached 0.35. The cultures were transferred to 45°C for 5 min to induce the λ phage, then returned to 37°C and grown with vigorous agitation (200 r.p.m.) for about 1 h. A small sample was removed and tested for lysis using a few drops of chloroform.

The culture was cooled in iced water, then transferred to centrifuge bottles and the cells harvested at 9,000 r.p.m. and 4°C for 10 min (Sorvall RC-5B). The supernatant was removed and the pellets drained (on ice). The cells were resuspended in a total volume of 2.2 ml 10% sucrose (w/v, in 50 mM Tris-HCl, pH 7.4). Two hundred µl of freshly prepared lysozyme (2 mg/ml in 250 mM Tris, pH 7.5) was added, and mixed by gentle inversion. The sample was snap-frozen in liquid nitrogen, then thawed at room temperature for 20 min, followed by 40 min on ice. One hundred µl of buffer Q1 was added, and the solutions mixed as before. The phage components were recovered by centrifugation at 35,000 r.p.m. and 2°C for 45 min (Beckman L5-65 ultracentrifuge, type 65 rotor). The supernatant was aliquoted into precooled Eppendorf tubes and snap-frozen in liquid nitrogen before long term storage at -70°C.

(b) sonicated extract: a single 500 ml culture was prepared as for the freeze-thaw lysate, using E.coli BHB 2690. The phage were induced, and the bacteria returned to 37°C. The culture was cooled, and centrifuged as above, but the pellet was resuspended in 3.6 ml of buffer A. The solution was sonicated until it was no longer viscous (6x 5 s bursts), keeping the sample below 6°C.

The extract was centrifuged at 6,000 r.p.m. and 4°C for 20 min (Sorvall RC-5B), and the supernatant aliquoted into precooled Eppendorf tubes. The samples were snap-frozen in liquid nitrogen and stored at -70°C.

2.11.4 Ligations and Packaging (Scalenghe et al., 1981; Grosveld et al., 1982)

Ligations were carried out using the standard protocol (2.8.3) with 0.5 µg of vector and 0.5 µg of insert DNA per 10 µl reaction. The ligation mixtures were incubated at 16°C overnight in preference to 4 h at room temperature.

Packaging reactions were set up as follows:

Ligation mixture	10.0
Buffer Q1	7.5
Buffer A	25.0
Sonicated extract	17.5
Freeze-thaw lysate	67.5
Total	127.5 µl

The packaging mixture was incubated at 25°C for 3 h. At the end of the incubation period, SM phage buffer (0.1% gelatin w/v, in 50 mM Tris-HCl,

pH 7.5/ 10 mM NaCl/ 1 mM $MgSO_4$) was added to give a total volume of 1.25 ml. The packaged DNA was stored at 4°C before transduction into E.coli NM 554.

2.11.5 Testing RecA⁻ Phenotype

The host cell line NM 554 was tested for reversion to a recombination positive phenotype by exposure to ultra violet irradiation. Two identical aliquots of 200µl were spread onto 2x TY agar plates. One was allowed to grow at 37°C, and the other was irradiated at 254 nm (CAMAG Universal u.v. lamp, fixed at 15 cm above the plate). Zones of 0, 5, 10, 15 and 20 s exposure to u.v. were produced by protecting the agar culture plate with a sheet of glass, which was moved every 5 s. E.coli Q 358 was taken as a RecA⁻ control.

No colonies were found in the regions which were exposed to the u.v. light, confirming the RecA⁻ nature of NM 554.

2.11.6 Transduction (Grosveld et al., 1981)

Single colonies of NM 554 were picked into L-broth supplemented with 0.4% maltose and grown at 37°C overnight. One ml of culture was spun down and the cells resuspended in 10 mM $MgCl_2$/ 10 mM Tris-HCl, pH 8.0/ 0.4% maltose. To one volume of packaged DNA an equal volume of cells was added, and the mixture was held at room temperature for 20 min, to allow adsorption of the phage particles. The samples were diluted with four times the total volume of L-broth, and were incubated at 37°C for 1 h and 30 min to allow expression of the ampicillin resistance gene.

An aliquot of 2 µl was retained to estimate the number of recombinants by plating onto L-agar containing 50 µg/ ml ampicillin. The remaining

cells were brought to a total volume of 16 ml, by harvesting and resuspending in fresh L-broth if necessary, and spread equally over 22 cm x 22 cm nitrocellulose filters on L-agar/ ampicillin to a density of 1.25×10^5 per filter.

2.11.7 Plating the Cosmid Library (Hanahan and Meselson, 1983; Grosveld et al., 1981)

The recombinants were plated over 22 cm x 22 cm nitrocellulose filters (Schleicher and Schuell) on L-agar containing 50 µg/ ml ampicillin in 23 cm x 23 cm square culture plates (Nunc). The approximate density was estimated at 125,000 colonies per filter. The plates were incubated at 37°C for 12-14 h, so that the colonies were visible, but no larger than 0.5 mm in diameter.

Impressions were taken onto fresh nitrocellulose pre-wetted on L-broth/ ampicillin plates. The copy was placed on top of the master and covered with a layer of 3MM Whatman paper, then a replica of the colonies was made by pressing the sandwich between two glass plates. The top plate was removed, and the filters keyed with indelible ink, before they were peeled apart and placed onto fresh plates. The master was returned to L-broth/ ampicillin and the copy to HMF 1/ ampicillin. Three further copies of each master were taken in the same way, but these were recovered on L-broth/ ampicillin. The copies were incubated at 37°C until the colonies were 0.5 mm in size. The first replica was transferred onto HMF 2/ ampicillin, recovered at room temperature for 2 h, then frozen at -70°C. The second replica was transferred to HMF 1/ ampicillin and retained at 37°C for a further hour. This was used to prepare a sandwich by following the replication procedure, using filters pre-wetted on HMF 1. The two filters were not peeled apart, but sealed

into plastic bags, and frozen at -70°C. The third and fourth impressions were used to make filters for screening. After recovery at 37°C, the colonies were lysed *in situ* by placing the nitrocellulose on 3MM Whatman papers soaked in the following solutions, and incubating for 3 min at each step:

(a) 10% SDS,

(b) 0.5 M NaOH,

(c) 1 M Tris-HCl, pH 7.4.

The filters were washed in 2x SSC for 15 min at room temperature, and the DNA fixed by baking at 80°C for 2 h.

The master filter was allowed to recover for 2 h between each round of replication. It was stored at 4°C to provide a copy from which the first positive colonies were taken after screening. Positives obtained from subsequent screens were picked from the plate stored at -70°C.

New filters were prepared from the frozen sandwich as follows. The sandwich was thawed at room temperature for 30 min. The filters were checked to ensure all keying marks were correct. The two nitrocellulose sheets were separated, and each was placed onto a 23x 23 cm Nunc plate containing L-agar/ ampicillin. The original was recovered at 37°C for three hours. At this point it could be transferred to HMF1/ ampicillin and used to prepare a fresh sandwich, or it could be washed 3-4 times using 2.5 ml 1x Hogness/ L-broth to prepare amplified library stocks. These were stored at -70°C. The second filter was recovered for 12 h at 37°C, then stored at 4°C overnight. Replicas were made from the second filter using the same protocol as above. If the original plate was used to prepare amplified stocks, two impressions were taken, and one returned as a fresh sandwich. If the original plate was used to prepare sandwich filters, only one impression was taken. Lysis of filters, and *in situ* fixing of DNA was as described above.

2.11.8 Screening

Cosmid library filters were hybridised according to the standard procedure at 42°C for 48 h. A minimum of 1x 10^5 c.p.m./ ml of freshly labelled probe DNA was used. They were washed at 65°C, as described in 2.7.4. Autoradiography was between two intensifying screens for between 4 h and 5 days.

2.11.9 Removal of Positive Recombinants from Frozen Master Plates (Steinmetz et al., 1986)

Autoradiographs were matched against the filters, and the keying marks transferred to the X-ray film. Duplicates were compared to establish whether potential positives were detected on both filters.

To remove positive recombinants the frozen master plates were thawed for 45 min at room temperature. Filters were air dried on 3MM Whatman paper before transferring onto fresh L-agar/ ampicillin plates. They were recovered for 2 h at room temperature. Master filters were aligned with autoradiographs on a light box, using a layer of ethanol-washed cling film to separate them. An area of about 2-3 mm in diamater was scraped from the positive area of the master filter using a sterile tooth-pick, or the cut end of an adapted Gilson pipette tip of similar cross-section. The colonies were put into 500 µl L-broth and recovered for 2-4 h at room temperature. The master filters were recovered on HMF 2/ ampicillin for 2 h at room temperature prior to refreezing at -70°C.

2.11.10 Rescreening Positive Recombinants

The stock prepared by scraping the positive area from the master

filter was diluted between 10^2 and 10^6 times in L-broth. From each dilution 200 µl was plated onto nitrocellulose on 9 cm L-broth/ ampicillin plates and incubated at 37°C overnight. The remaining stock was stored at -70°C in 1x Hogness/ L-broth.

Two replicas of the rescreen plates were taken using the same protocol as for the original library preparation (2.11.7). These were recovered on L-broth/ ampicillin plates at 37°C, lysed, and baked to provide filters for screening. The master rescreen plate was stored at 4°C.

Rescreen filters were hybridised, washed and autoradiographed using standard procedures. The rescreening was repeated until single positive colonies were available.

2.11.11 Preparation of Cosmid DNA

Single positive colonies were grown in 10 ml L- broth/ ampicillin, and used to prepare cosmid DNA according to the standard "miniprep" protocol (2.3.3).

2.11.12 Characterisation of DNA inserts

Inserts were digested with Bam HI, Bgl II, Cla I, Eco RV, Hind III, Kpn I (or, its isoschizimer, Asp 718), Sal I and Xho I in single and double digest combinations. Fragments were separated on 0.8 % agarose gels and blotted onto nitrocellulose for analysis by hybridisation with various walking probes, vector DNA, and genomic DNA (see chapter III).

CHAPTER III

PREPARATION AND ANALYSIS OF THE COSMID LIBRARIES: CHROMOSOME WALKING IN THE MHC.

3.1 INTRODUCTION

Chromosome walking involves the systematic isolation of DNA from a specific genomic location in a series of overlapping recombinant clones. Each step relies upon the preparation of unique sequence probes from the outer limits of the cloned DNA. Once isolated, a new probe is used to screen a library of genomic fragments to find the next set of overlapping recombinants. The inserts are characterised by restriction enzyme digestion and Southern blotting to produce a molecular map of the region. The three stages (probe isolation, screening and analysis of the DNA) are repeated in a sequential fashion to extend the original data. Chromosome walking is facilitated by using a vector system which can introduce large fragments of foreign DNA into the chosen host. Cosmids are hybrid DNA molecules derived from plasmids based on the ColE1 replicon carrying the cohesive end site (*cos*) of bacteriophage λ (Collins and Hohn, 1978). Like plasmids, they retain the ability to replicate autonomously within *E.coli*, and carry selectable drug resistance markers. The presence of the *cos* site permits *in vitro* encapsulation using a bacteriophage λ system, but the smaller size of the vector (in the range of 4.5 to 6 kb) allows 40 to 45 kb of foreign DNA to be inserted. The DNA can then be recovered using standard plasmid preparative methods (Birnboim and Doly, 1979).

Cosmid libraries have been used successfully for the characterisation of both human and murine MHC genes (Carroll *et al.*, 1984, 1985a; Strachan, 1987; Malissen *et al.*, 1982; Trowsdale, *et al.*, 1985; Spies *et al.*, 1985; Strominger, 1987; Muller *et al.*, 1987b). In addition,

overlapping cosmid clusters containing the class I, II and III region loci from the H-2 complex in mouse have been linked together by pulsed-field gel electrophoresis (PFGE) to construct a molecular map spanning ∿2,300 kb (Muller et al., 1987a). In man, a similar approach has been used to determine the organisation of the class II subregions (Hardy et al., 1986), and to produce an overall physical map of the HLA region (Dunham et al., 1987; Carroll et al., 1987). Within the class III region, the complement genes C2, factor B (Bf), C4A and C4B, and the associated cytochrome P-450 steroid 21-hydroxylase (21OH) loci 21OHA and 21OHB have been mapped in overlapping cosmids (Carroll et al., 1984, 1985a) spanning 120 kb. The genes show a head-to-tail organisation, with less than 0.5 kb between C2 and factor B (Wu et al., 1987), and a gap of 30 kb between the factor B and C4A genes. The C4A and C4B genes are separated by 10 kb, and each has an associated 21OH locus within 3 kb of its 3' end (Carroll et al., 1985a). Attempts to determine the orientation of the complement genes with respect to other MHC markers by analysis of recombination events has proved contradictory, with some groups favouring C2 at the centromeric side of the cluster (Bodmer and Bodmer, 1984; Wilton and Charlton, 1986) and others at the telomeric end (Abbal et al., 1987).

In order to prepare a detailed molecular map of the class III region in man, cosmid libraries were constructed by Mbo I partial digestion of DNA isolated from an EBV transformed lymphoblastoid cell line. The donor individual was the offspring of a consanguineous (first cousin) marriage with a homozygous HLA type (A2, B7, DR2, C2 C, Bf S, C4A 3, C4B QO). A deletion encompassing the 21OHA gene and the C4B gene is responsible for the C4B QO phenotype, and has been described previously (Carroll et al., 1985b; Yu and Campbell, 1987). An initial cosmid cluster containing the complement genes was used to confirm their arrangement in this haplotype

with respect to the published data (Carroll et al., 1984, 1985a). New probes isolated from the two extremities were subsequently used for walking, and to establish the precise orientation and location of the class III genes within the MHC by PFGE (Dunham et al., 1987). A second walk was initiated from the tumour necrosis factor (TNF) α and β genes, following their linkage to the region between complement and HLA-B (Dunham et al., 1987; Carroll et al., 1987; Ragoussis et al., 1988), to isolate a total of 541 kb of cloned DNA.

3.2 RESULTS

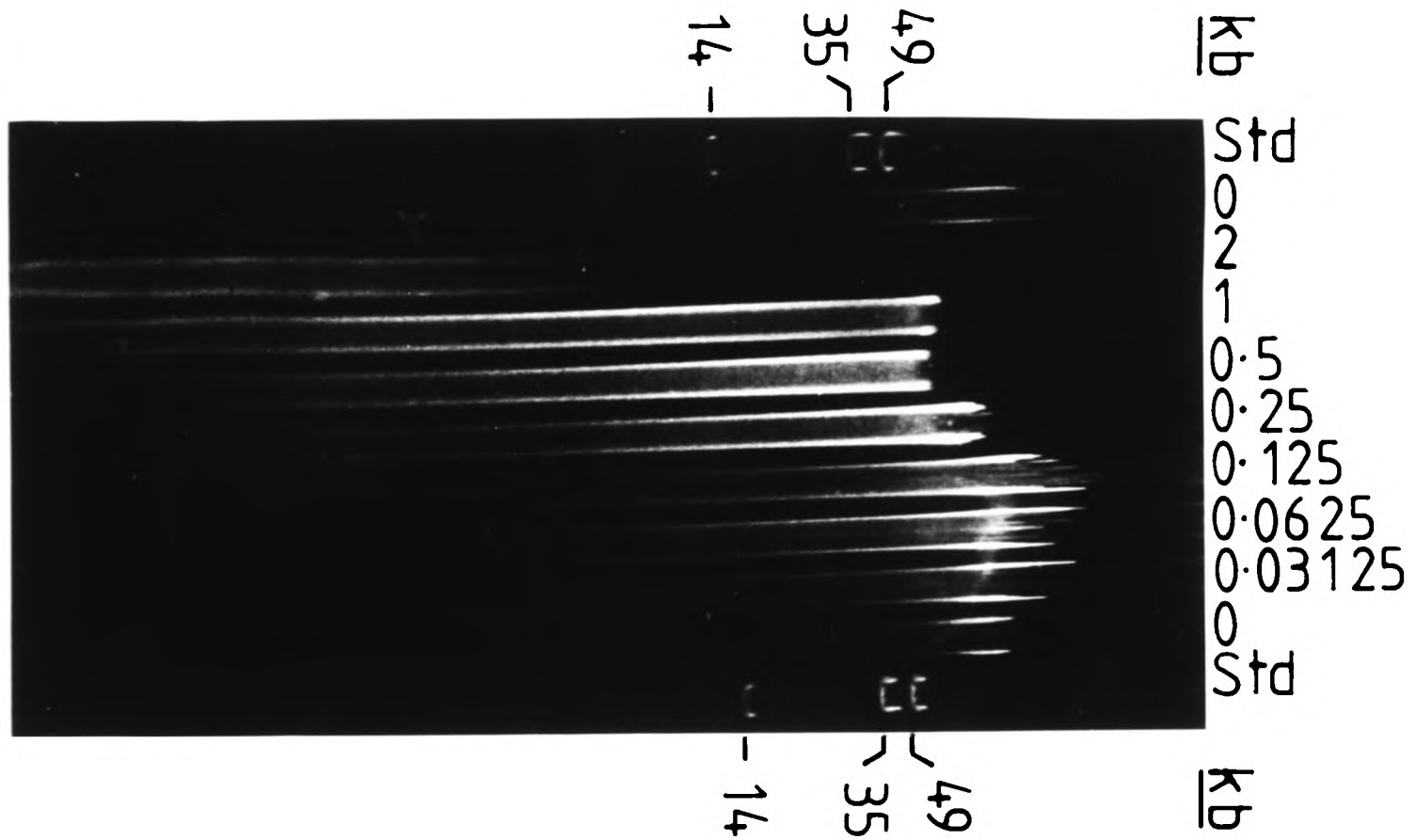
3.2.1 Preparation of the Cosmid Library

3.2.1.1 Preparation of the Insert DNA

High molecular weight DNA was restricted with Mbo I and phosphatased in a series of trial experiments. Samples of 3 µg of genomic DNA were incubated with 2.0-0.03125 U Mbo I in twofold serial dilutions, as described in 2.11.2. The samples were fractionated on a 0.4% agarose gel using uncut λ (49 kb) and Xho I cut λ (35 and 14 kb) as standards (Fig. 3.1). To ensure that regions of DNA which were more, or less, accessible to enzyme were represented in the final 35-50 kb fraction used to construct the library, five enzyme concentrations were chosen. These included partial digests which were over- and under-digested with respect to the optimal conditions, in which the bulk of the cleaved DNA lay within the correct size range. The appropriate range of digests was scaled up exactly by 20 times, using 40, 20, 10, 5 and 2.75 U Mbo I for each 60 µg of genomic DNA. The digests were performed in low salt Mbo I buffer, followed by treatment with phosphatase to reduce the risk of

Fig. 3.1

Mbo I partial digests of genomic DNA. The uncut DNA (0) is compared against 3 µg samples cleaved with two-fold dilutions of enzyme, and fractionated on a 0.7% agarose gel. Samples 1 to 0.0625 were scaled up twenty times for preparation of the cosmid library.


kb
49
35
14
Std
0
2
1
0·5
0·25
0·125
0·0625
0·03125
0
Std
14
35
49
kb

ligation between non-contiguous DNA fragments.

3.2.1.2 Fractionation of insert DNA

The conditions for the size fractionation of the Mbo I partial digests were predetermined with trial runs using bacteriophage λ DNA. The λ genome was digested with Xho I or Hind III prior to end filling with [α-^{32}P]-dATP (2.8.2) and desalting on a Sephadex G-100 column. Duplicate sucrose density gradients were set up using 9 ml steps of 10-40% with 10% increments, and 6 ml steps of 20-45% in 5% increments. One of each pair of gradients was loaded with Xho I cut λ, and the other with a 1:1 mixture of the Xho I and Hind III digests. The samples were centrifuged and 0.7 ml fractions collected as described in 2.11.3. The total radioactivity in each fraction was monitored and plotted against fraction number. Peak separation was more efficient with the 20-45% gradient (Fig. 3.2A). From fractions 20 to 37, 50 µl were run on a 0.6% agarose gel which was dried onto Whatman 3MM paper using a vacuum drier (Aquavac, Uniscience), and autoradiographed (Fig. 3.2B). This confirmed the separation of the major size markers from 35-1.98 kb by centrifugation.

Electrophoresed samples taken from the fractions recovered after the centrifugation of Mbo I partial digests of genomic DNA are shown in Fig. 3.3. These can be compared against the samples of unfractionated partial digests in tracks 16-19. The standards, a mixture of uncut λ, λ Xho I and λ Hind III with fragment sizes of 49, 35, 24, 14, 9.5, 6.6, 4.3, 2.26 and 1.98 kb, emphasise that each fraction contains a distinct band of DNA which lies within a narrow size range. Fractions containing DNA fragments of 35-50 kb were ethanol precipitated, and the pellets resuspended in 25 µl 10 mM Tris-HCl/ 0.1 mM EDTA, pH 7.4. Each 0.7ml

3·2B

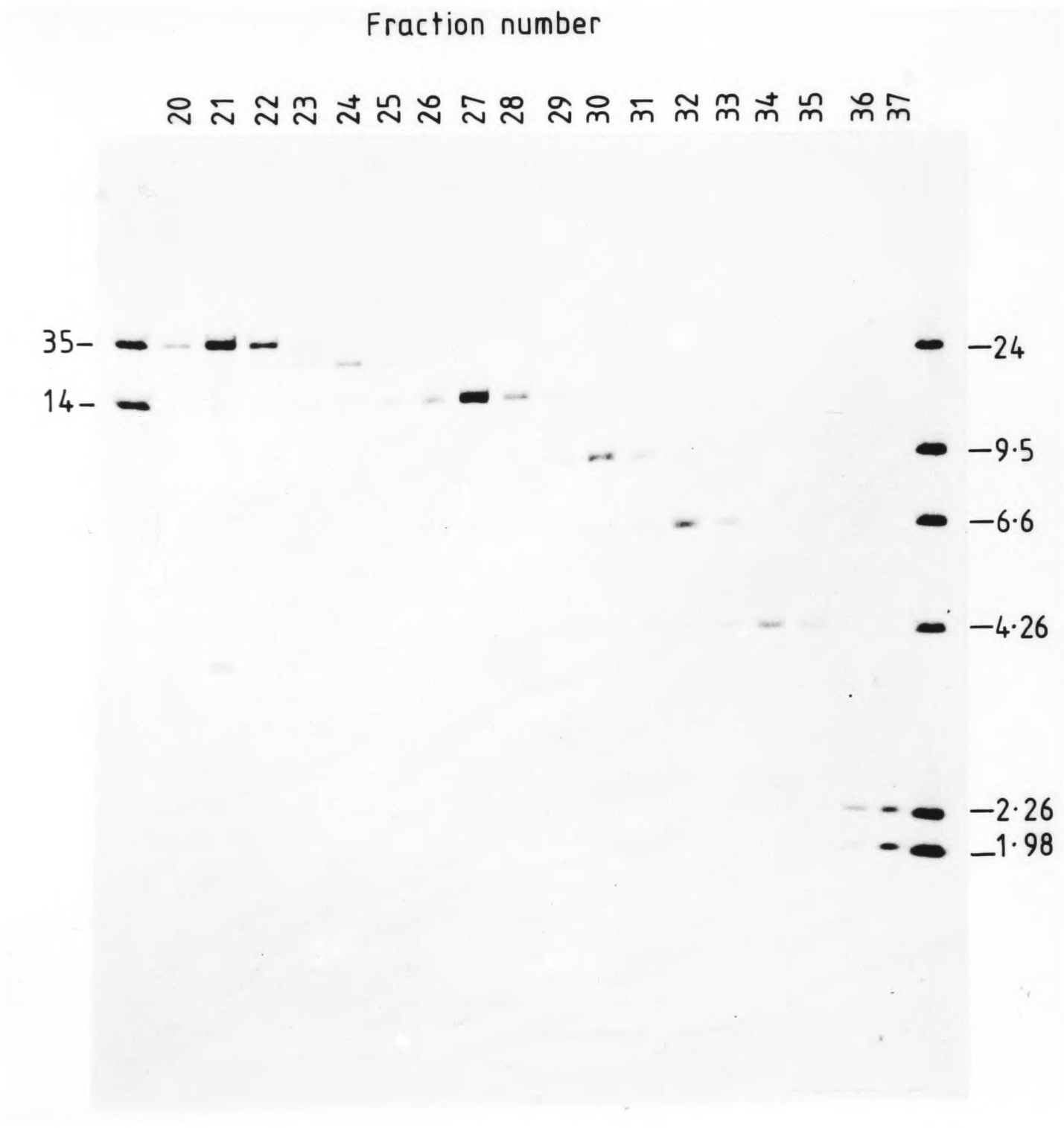
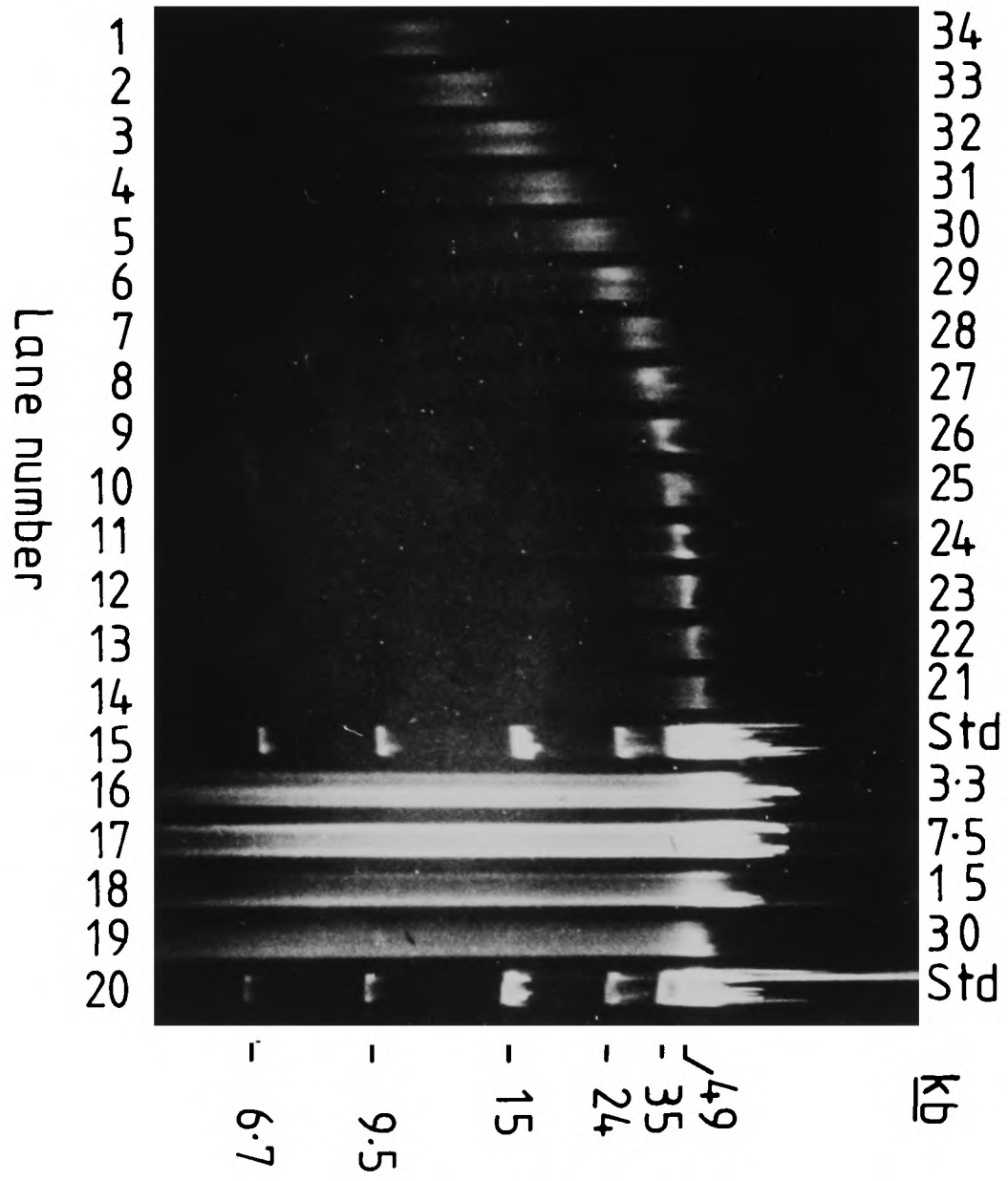

Fraction number
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
35–
14–
—24
—9·5
—6·6
—4·26
—2·26
—1·98

Fig. 3.3

Samples from sucrose density gradient centrifugation of Mbo I digested genomic DNA (lanes 1 to 14) compared with unfractionated partial digests (lanes 16 to 19). Standards consisted of mixtures of uncut λ, λ Xho I and λ Hind III. Fractions containing DNA in the 35 to 50 kb size range were ethanol precipitated and used to prepare the cosmid library.


Lane number
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Fraction number
34
33
32
31
30
29
28
27
26
25
24
23
22
21
Std
3·3
7·5
15
30
Std
kb
49
35
24
15
9·5
6·7

fraction yielded about 5 µg DNA. Only the most efficiently ligated and packaged fraction, as determined from a series of trial experiments (3.2.1.4), was used to prepare a library.

3.2.1.3 Preparation of the Vector Arms

pDVcos is a small cosmid vector (4.5 kb) containing a duplicated *Xho* I/ *Pvu* II fragment which encompasses the *cos* site necessary for DNA encapsulation with bacteriophage λ *in vitro* packaging extracts. This duplication permits the simple preparation of vector arms in a two step process. Firstly, cleavage between the *cos* sites at *Pvu* II followed by phosphatase treatment prevents religation of vector fragments to form oligomers. Secondly, cleavage at *Bam* HI produces two vector arms with ends compatible with the *Mbo* I used to prepare the insert DNA. Each arm has its own *cos* site. In addition, one contains the ampicillin resistance marker (amp^R) and the origin of replication from ColE1 (ori) (Fig. 3.4).

Since the *cos* site at each end of the recombinant must be in the same orientation, and separated by 38-53 kb, to allow encapsulation into the λ phage heads, only the correct combination of arms with insert DNA can be packaged. Dimers formed by religation at the *Bam* HI site are excluded on the basis of size, and oligomers of vector should be eliminated by phosphatasing the *Pvu* II site. Furthermore, as the digests are performed sequentially on the same DNA sample, the arms are present in the ligation mixture at the correct ratio of 1:1.

3.2.1.4 Ligations and Packaging

Trial ligations and packaging experiments were performed to establish

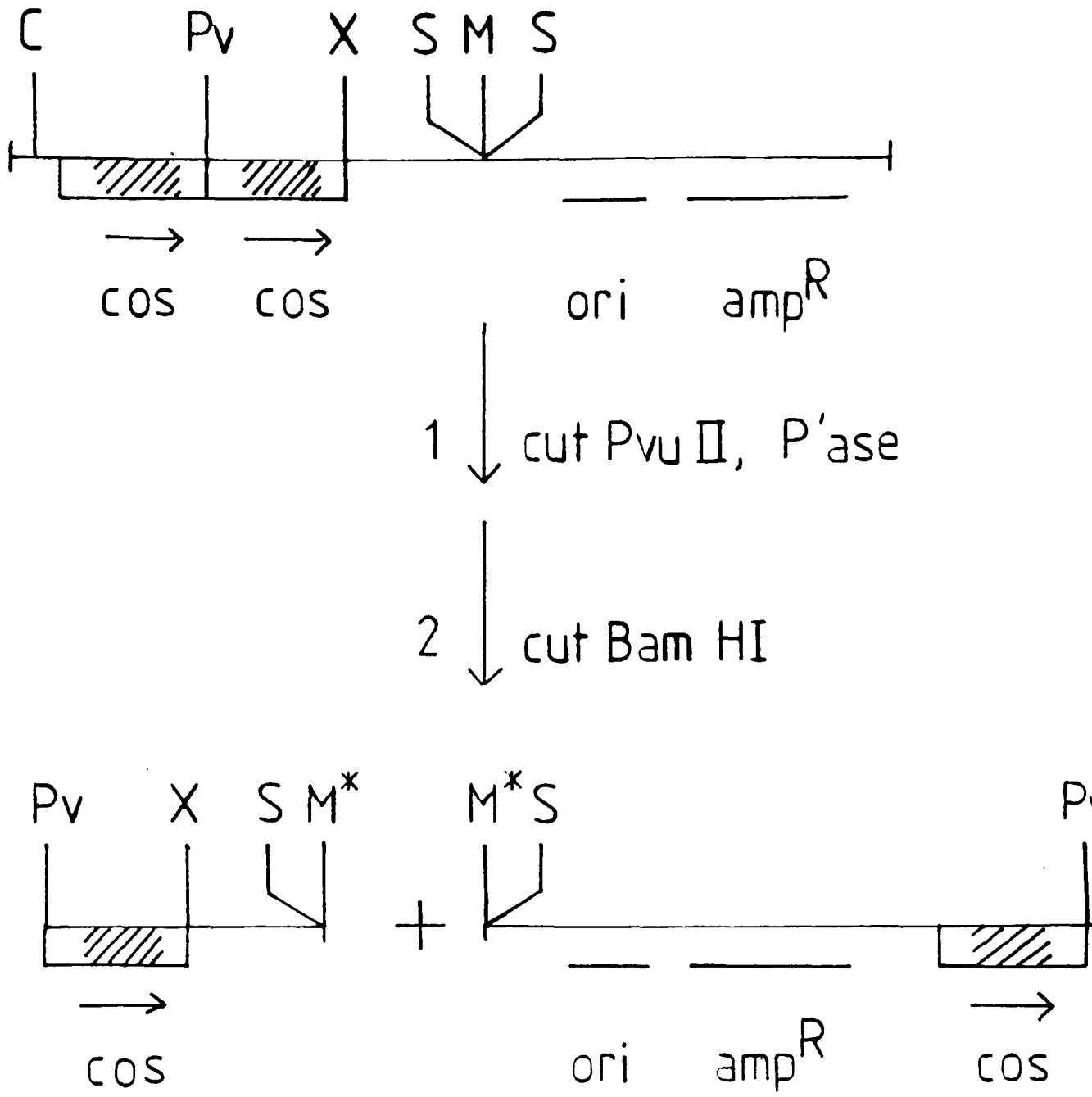


Fig. 3.4

A schematic representation of the preparation of pDVcos vector arms. The orientations of the cos sequences are indicated by the horizontal arrows. The sites for the enzymes Cla I (C), Pvu II (Pv), Xho I (X), Sal I (S) amd Bam HI (M) are shown. Following step 2, the Bam HI cloning site is marked by a *.

which genomic DNA fractions from the 35-50 kb range were suitable for library construction. The optimal conditions were found to be 1:1 (w/w) vector DNA to genomic DNA during the ligation, with a packaging time of 3 h.

All trial ligations were carried out in a 10 µl reaction volume using 250 ng of vector arms. After 2 µl were taken for packaging, a tenth of the encapsulated DNA was introduced into E. coli by transduction. The cells were plated onto L-agar containing 50 µg/ml ampicillin, and incubated overnight. The numbers of colonies per plate were scaled up to calculate the total number of recombinants /µg of insert DNA, as shown in table 3.1. The most colonies were obtained with fraction 25 (Fig. 3.3), which was ligated and packaged almost twice as efficiently as the preceeding fraction (24) containing slightly larger DNA fragments, and tenfold better than the following fraction (26) containing slightly smaller DNA fragments (Table 3.1).

The titrations used NM 554 and 1046 as host bacterial strains for transduction with the encapsulated DNA. Both exhibited a low rate of back mutation to the RecA phenotype following ultra-violet testing (2.11.5). However, transduction efficiencies with NM 554 were twofold higher than with 1046, leading to its selection as the host strain for cosmid library preparation.

3.2.1.5 Transduction and Plating

A maximum of 2×10^5 recombinants per µg of genomic insert DNA was attained with the optimal conditions for ligation, packaging and transduction (Table 3.1). Two libraries were prepared from Mbo I partial digests, each using 5 µg of size-fractionated genomic DNA. This gave a potential 2×10^6 recombinants. In total 1.5×10^6 recombinants were

Table 3.1

Packaging, ligation and transduction efficiencies of DNA samples.

F	ng	no. R	
		1046	NM554
23	50	3.0×10^3	1.2×10^4
	250	3.5×10^4	9.6×10^4
24	50	5.9×10^4	1.3×10^5
	250	6.8×10^4	1.4×10^5
25	50	3.5×10^4	1.2×10^5
	250	1.0×10^5	2.6×10^5
26	50	8.4×10^3	1.0×10^4
	250	1.2×10^4	2.8×10^4

F= fraction number
ng= amount of DNA (in ng) per ligation mixture
no. R= the number of recombinants per ug of insert DNA.

plated, at a density of about 125 000 colonies per 22x 22 cm nitrocellulose filter. The remaining 5x 10^5 clones were lost after an attempt to store packaged cosmids at 4^oC in phage dilution buffer was found to result in a tenfold decline in their transduction efficiency.

3.2.1.6 Screening of Library Filters

Both Hybond-C (nitrocellulose) and Hybond-N (nylon) were used to prepare replica filters and replacement filters for screening. Hybond-N, although more resistant to damage caused by repeated handling, was found to give a higher background radioactivity owing to non-specific probe hybridisation. Furthermore, the intensity of the positive signal decreased more dramatically than when nitrocellulose was used for successive screenings. Typically, nitrocellulose membranes were autoradiographed at -70^oC for 3 h to 2 days, whereas nylon membranes required 12 h to 5 days.

3.2.1.7 Removal of Positive Recombinants

Unless otherwise stated, each round of screening during the chromosome walk involved four pairs of duplicate filters, or 5x 10^5 recombinants. Positives were removed from the -70^oC master plates as described in 2.11.9. The frozen masters were subjected to at least 6-7 cycles of thawing without observing a substantial decline in the viability of the recovered stocks.

3.2.2 Characterisation of the Complement and 21-Hydroxylase Gene Cluster

The initial cosmid clones were isolated using a 21OH genomic probe

(Rodrigues et al., 1987), the factor B cDNA (Morley and Campbell, 1984) and a 1.6 kb Hind III genomic fragment (probe K) located 6.5 kb 5' to the C2 gene (R.D.Campbell, unpublished). All three positives which were isolated with the factor B cDNA and the six which hybridised with probe K were taken for restiction enzyme analysis (Fig. 3.5).

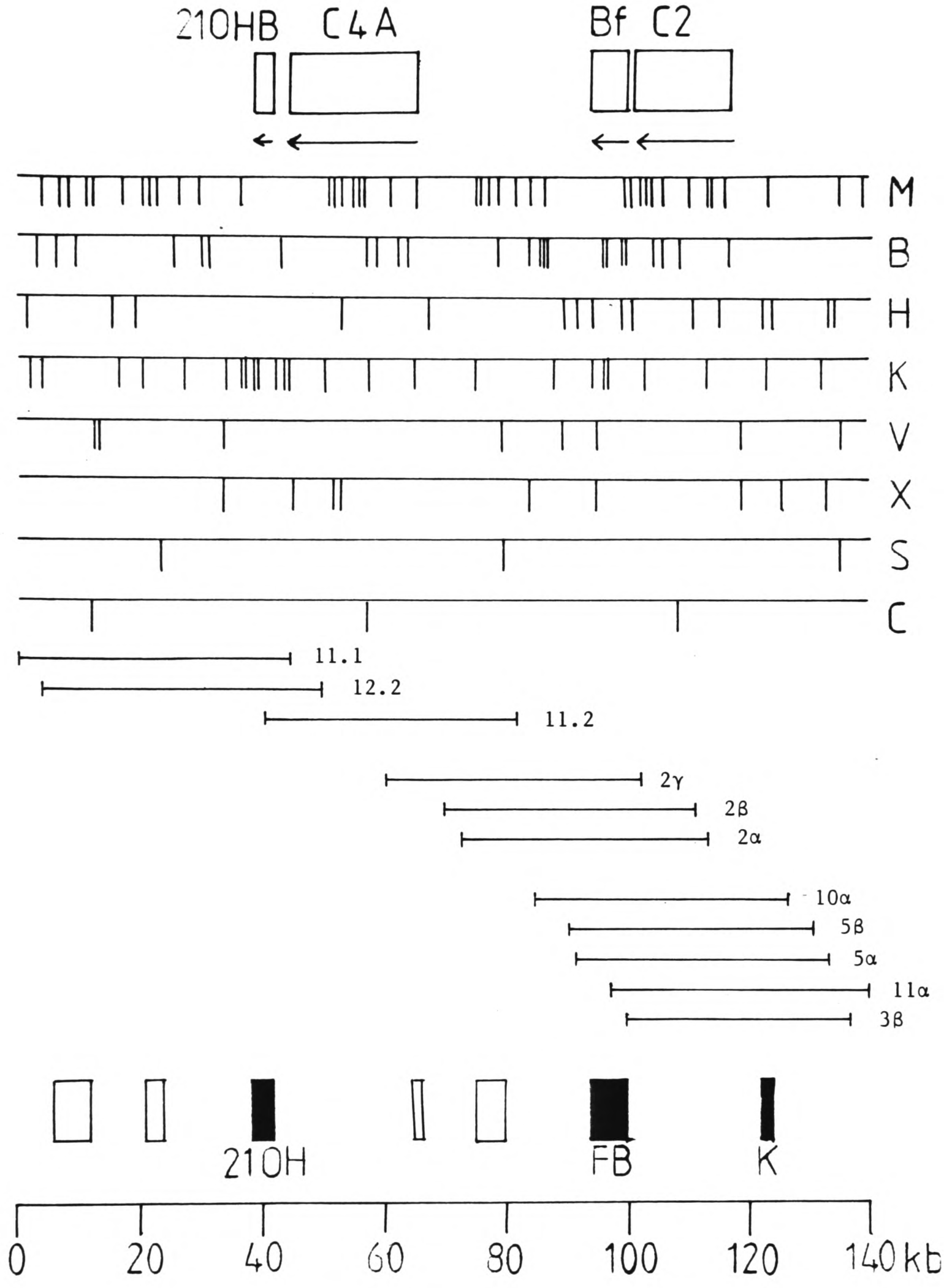
Over 60 positives were detected with the 21OH probe. To eliminate multiple clones covering the same region of DNA, 2µl from each of the rescreen stocks were dotted onto nitrocellulose grids, incubated at 37^{o}C, lysed, and hybridised with a C4 probe (Belt et al., 1984). Only two cosmids were found to extend away from the 21OHB locus; these, along with one other cosmid known to contain the C4A gene, were selected to complete the characterisation of this region.

The cosmid inserts were analysed by restriction enzyme digestion and Southern blotting. Each clone was cut with Bam HI, Bgl II, Cla I, Eco RV, Hind III, Kpn I (or its isoschizimer, Asp 718), Sal I and Xho I in single and double digest combinations (Fig.3.5). The enzyme Sal I was used to cleave the pDVcos out of the DNA as a 4.5 kb fragment. The sizes of all restriction fragments were estimated to within 5%. Bands smaller than 0.3 kb were, however, below the detection limits of the ethidium bromide stained agarose gels used to separate the restriction digest products.

Following fractionation of the restriction digests on 0.8% w/v agarose gels, three Southern blots were prepared. One was hybridised with the identifying probe, and one with pDVcos DNA, to confirm the restriction map generated from the digest data. The third Southern blot was hybridised with total human genomic DNA to identify bands with little or no repetitive DNA elements (see 3.2.2.1).

Fig 3.5

Restriction map of the complement/ 21OH gene cluster. Cosmid genomic inserts were mapped using Bam HI (M), Bgl II (B), Hind III (H), Kpn I (or its isoschizimer Asp 718) (K), Eco RV (V), Xho I (X), Sal I (S) and Cla I (C). The 5' to 3' orientations of the known genes are indicated by horizontal arrows. The limits of the cosmid inserts are shown by the horizontal bars (⟝⟞). Regions of unique DNA sequence are defined below the cosmid inserts by the open boxes, and the probes used to isolate the cosmids, by the shaded boxes.

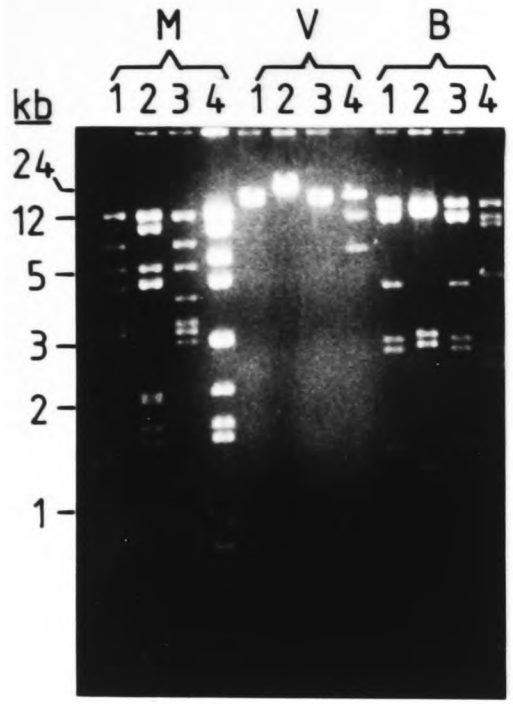

210HB
C4A
Bf
C2
M
B
H
K
V
X
S
C
11.1
12.2
11.2
2γ
2β
2α
10α
5β
5α
11α
3β
210H
FB
K
0
20
40
60
80
100
120
140 kb

Digests of the cosmids isolated with 21OH are shown in Fig. 3.6A, alongside autoradiographs of probed Southern blots. In each case, the fragments which hybridised to 21OH (Fig. 3.6B) and pDVcos (Fig. 3.6C) could be related to the restriction maps of the individual cosmids, and shown to be consistent with the data. The autoradiograph of the blot probed with radiolabelled genomic DNA (Fig. 3.6D), showed that fragments of ∿0.9, 1.0 and 3 kb in the Bam HI digests and ∿1.6 and 3 kb in the Bgl II digests of cos12.2 and cos11.1 were potentially non-repetitive. Comparison of the positions of these fragments with the map of cos11.1 (Fig. 3.6E), indicated that the Bam HI and Bgl II fragments at 3 kb were the furthest away from the 21OH probe, and suitable for isolating a new walking probe.

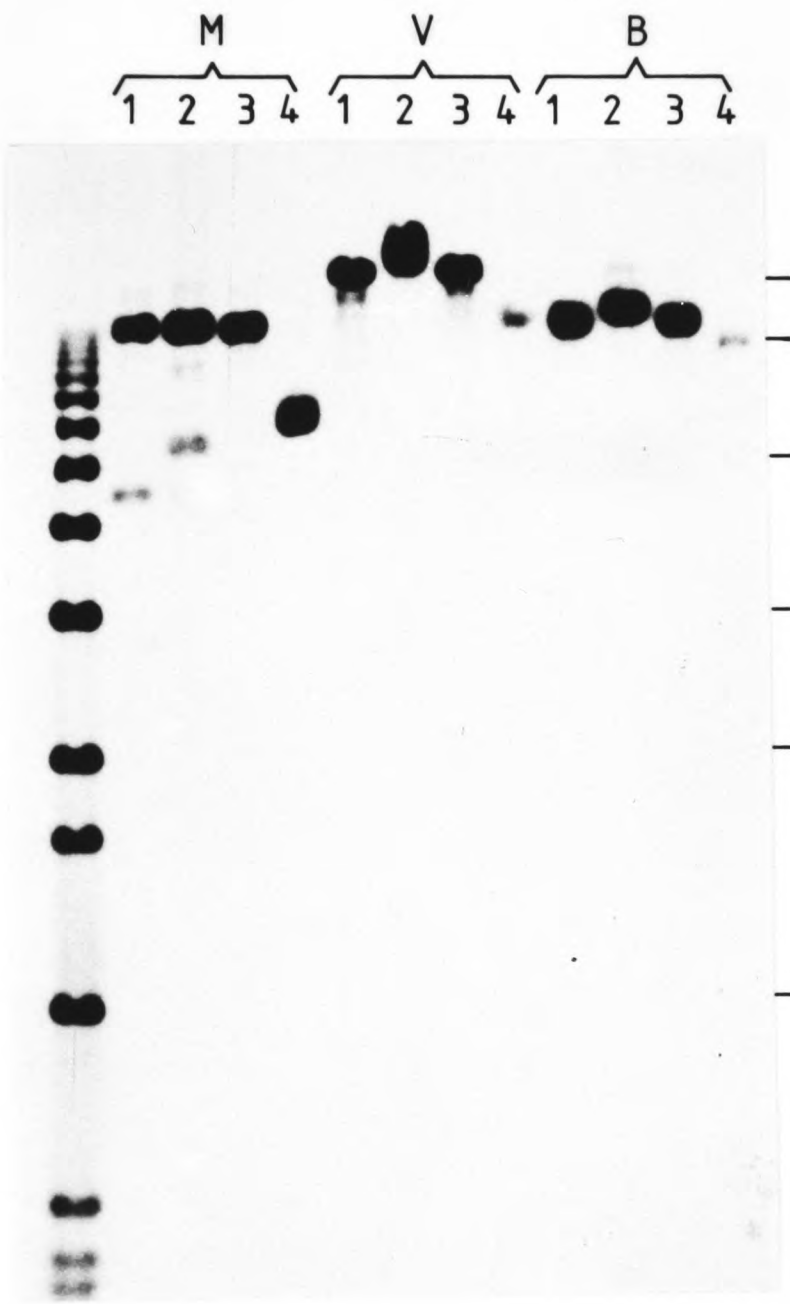
In total, the cluster of 12 cosmids was found to cover 140 kb (Fig. 3.5). The clones isolated with the 21OH probe mapped 37.5 kb beyond the 3' end of the 21OHB gene, encompassed 21OHB and the whole of the C4A gene. The three positives detected by the factor B cDNA spanned the 30 kb gap between C4A and the factor B gene, and 75% of the C2 gene. All of the cosmids obtained with probe K contained a complete C2 gene, and with the exception of cos3β, at least part of the factor B gene (Fig. 3.7B). One of these, cos10γ, was non-contiguous in the region 5' to C2. Although the Bam HI digest appeared very similar to cos11α and defined endpoints within the region covered by the cosmid cluster, the Xho I/ Sal I digest showed that not all the fragments generated were in common with other cosmids containing the C2 gene (Fig. 3.7A). This insert was probably generated by the ligation of a large restriction fragment to a smaller genomic element which had contaminated the 35-50 kb DNA fraction isolated from the sucrose density gradient. The maximum extension at the C2 end of the cluster was 22 kb.

Fig 3.6

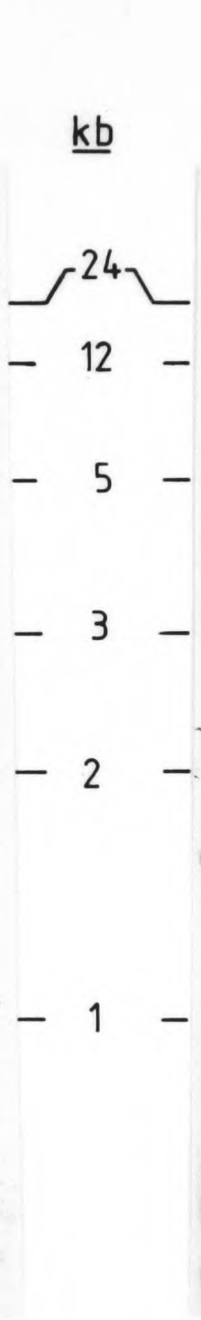
Characterisation of the cosmids isolated with the 21OH probe. Cosmids cos12.2 (1), cos11.2 (2), cos11.1 (3) and cos2 γ (4) (a factor B positive cosmid) were cleaved with Bam HI (M), Eco RV (V) or Bgl II (B) and fractionated on a 0.8% agarose gel (Fig 3.6A). Southern blots were hybridised with the 21OH probe (Fig. 3.6B), pDVcos (Fig. 3.6C) or genomic DNA (Fig. 3.6D). The positive result of 4 with the 21OH probe probably arises from cross-hybridisation of the probe to vector sequences (compare with Fig. 3.6C). For cos11.1, the sizes of the hybridising fragments (kb) are indicated above the restriction map (pDVcos) or below (21OH). Potential unique sequences, which fail to hybridise with the genomic DNA, were mapped in cos11.1 (Fig. 3.6E). A new walking probe, probe A, isolated from the 3 kb Bgl II fragment was tested on a genomic Southern of Bam HI digested DNA from two individuals (Fig. 3.6F), before being used to rescreen the cosmid libraries.


M
V
B
1 2 3 4 1 2 3 4 1 2 3 4
kb
24
12
5
3
2
1

A


M
V
B
1 2 3 4 1 2 3 4 1 2 3 4

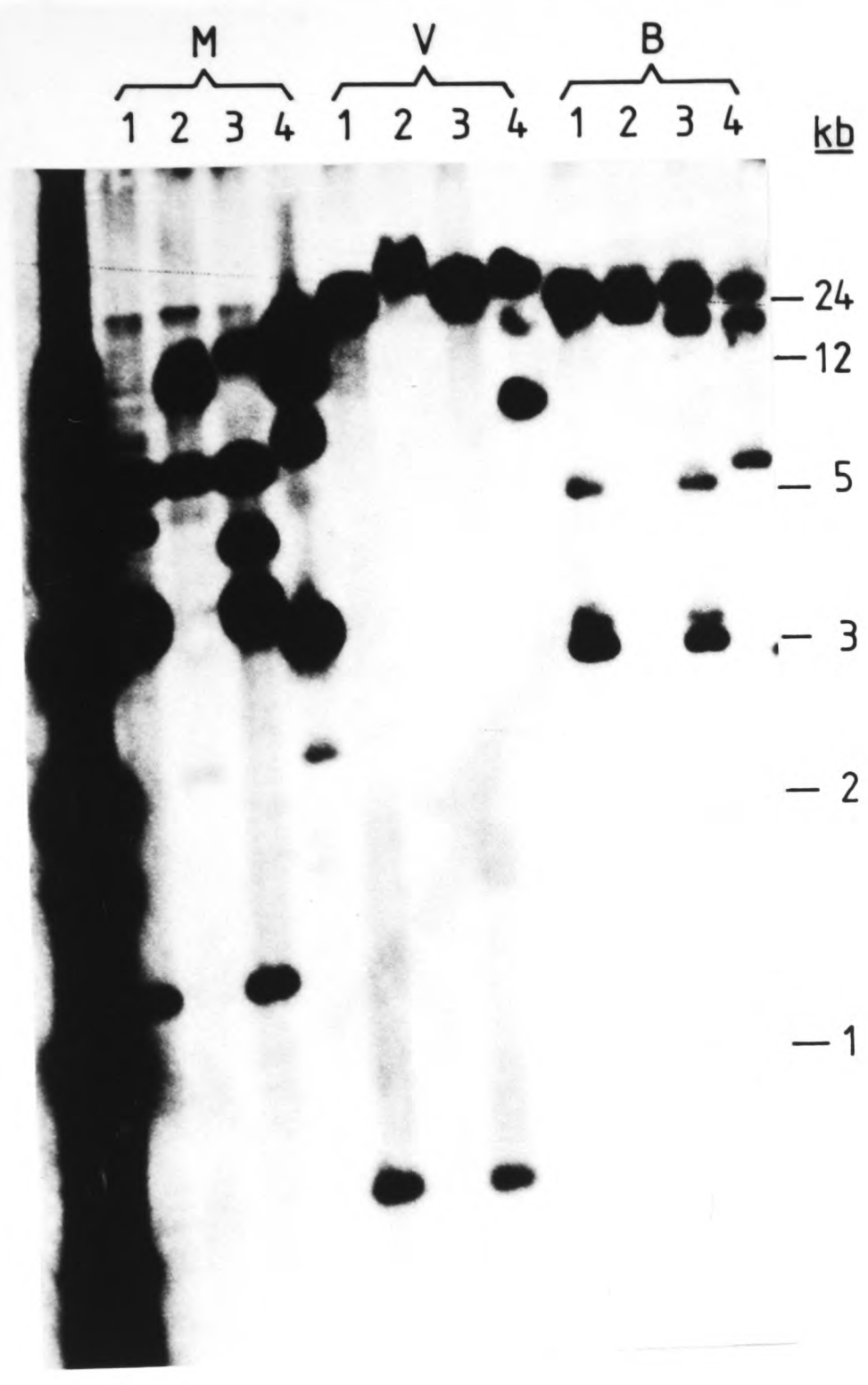
B


kb
24
12
5
3
2
1

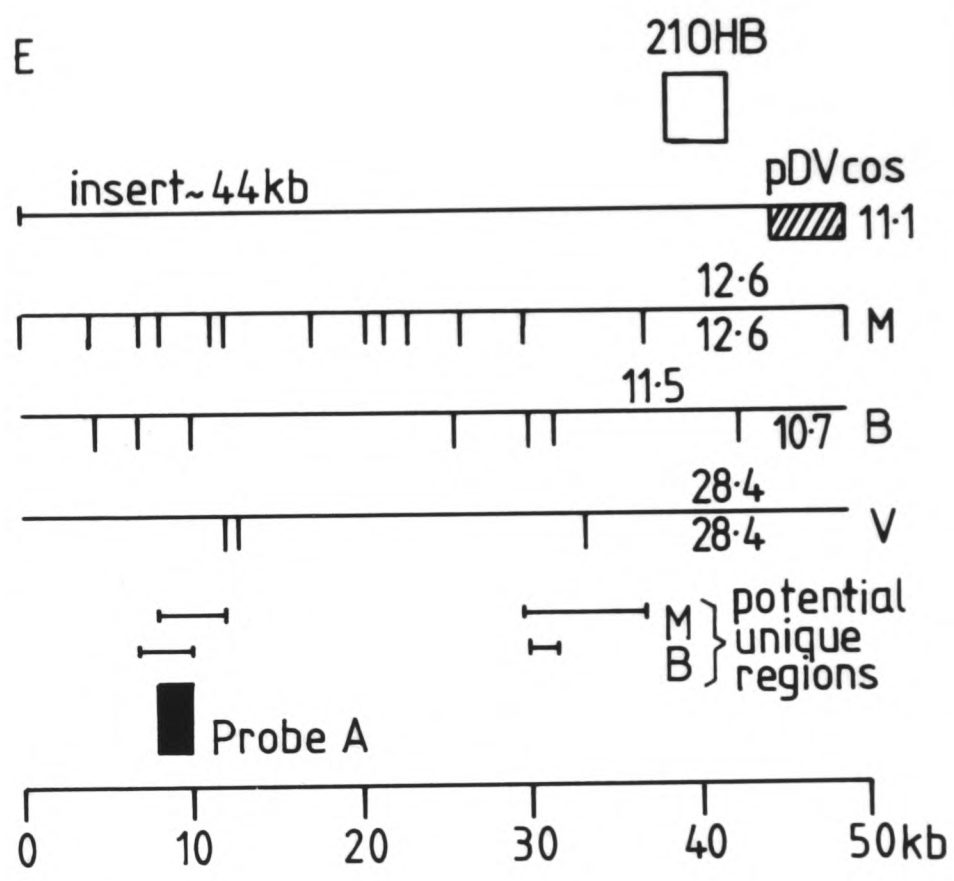

M
V
B
1 2 3 4 1 2 3 4 1 2 3 4

C

PTO


M
V
B
1 2 3 4 1 2 3 4 1 2 3 4
kb
24
12
5
3
2
1

D


E
21OHB
insert~44kb
pDVcos
11·1
12·6
12·6
M
11·5
10·7
B
28·4
28·4
V
M
B
potential
unique
regions
Probe A
0
10
20
30
40
50kb

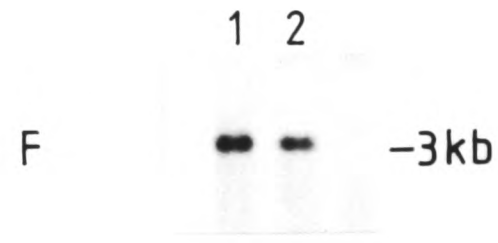
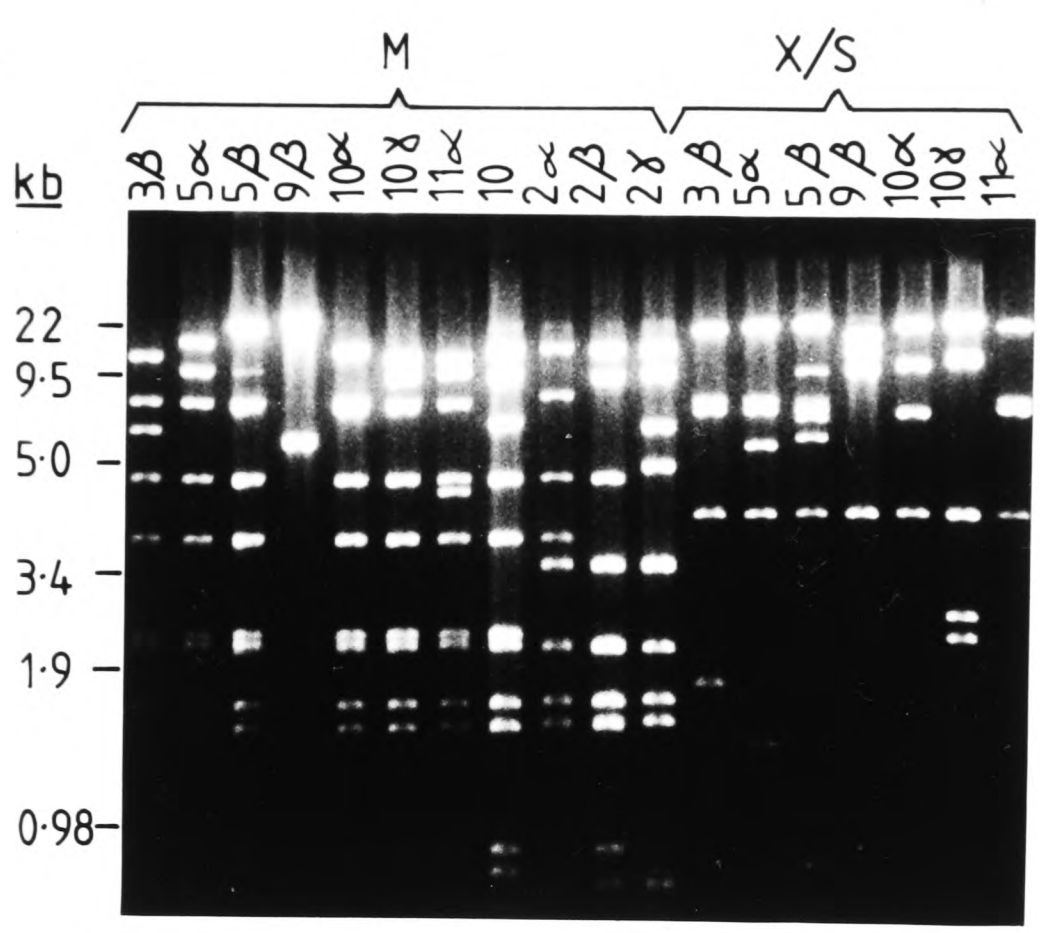

1 2
F
~3kb

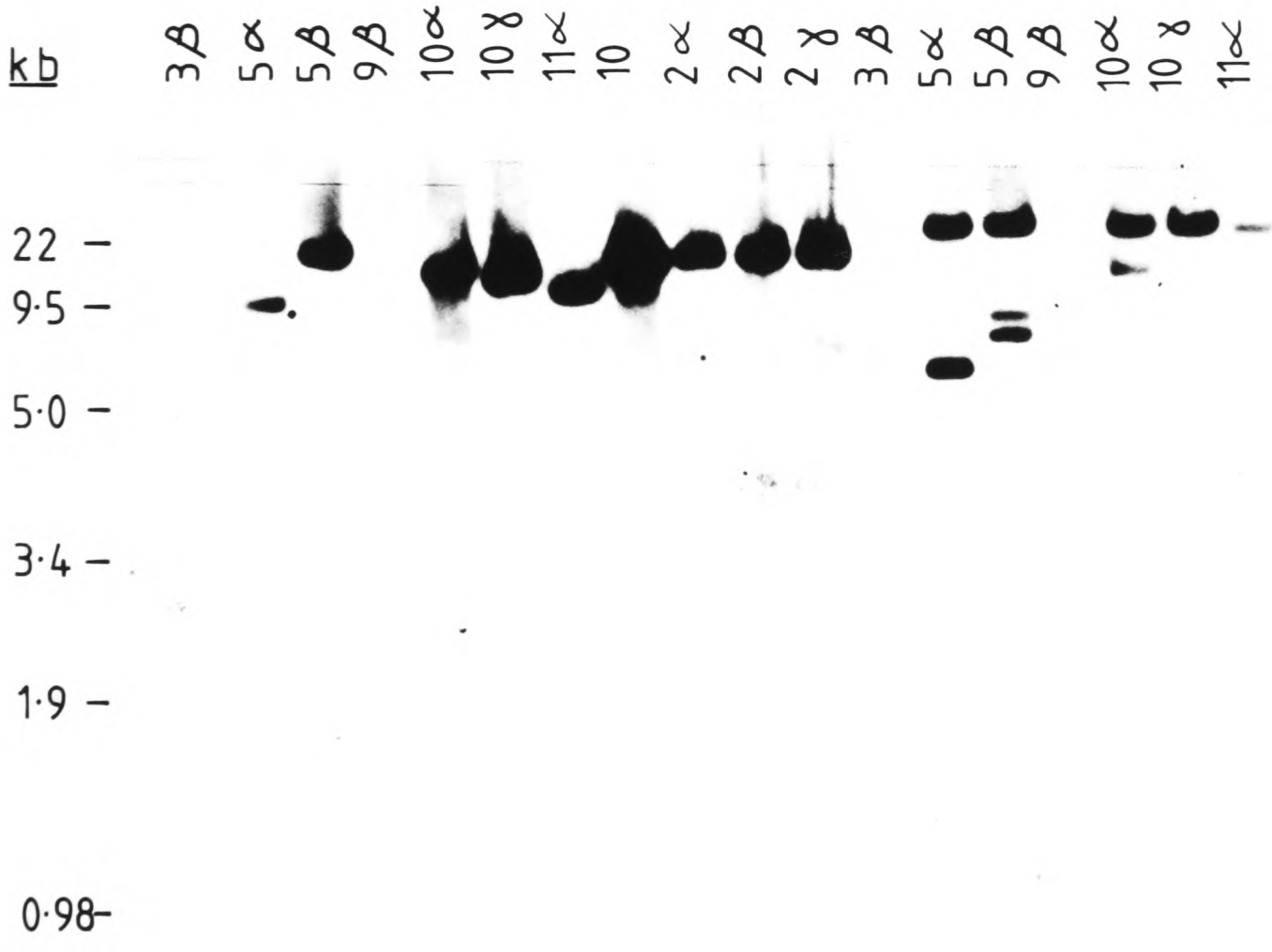
Fig. 3.7

Positives isolated with the factor B probe and probe K. DNA samples were digested with Bam HI (M) or Xho I/ Sal I (X/S), and fractionated on a 0.8% agarose gel (Fig. 3.7A). A Southern blot was hybridised with the factor B probe (Fig. 3.7B). From the results, cos9β is probably an incorrect pick, as the restriction digest patterns are inconsistent with the other clones, and the DNA does not hybridise with the Factor B probe. Cos3β, which was isolated with probe K, does not contain any factor B sequence. Cos10 is an internal standard containing the factor B and C2 genes (Carroll et al., 1984).

A


M
X/S
kb
3β
5α
5β
9β
10α
10γ
11α
10
2α
2β
2γ
3β
5α
5β
9β
10α
10γ
11α
22
9·5
5·0
3·4
1·9
0·98

B


kb
3β
5α
5β
9β
10α
10γ
11α
10
2α
2β
2γ
3β
5α
5β
9β
10α
10γ
11α
22
9·5
5·0
3·4
1·9
0·98

The overall organisation of the genes in the complement/ 21OH gene cluster was consistent with that published by Carroll et al. (1984) allowing for the deletion that encompasses the 21OHA and C4B loci in this cell line. Analysis of Taq I genomic digests has shown that the 21OH gene present in this haplotype is a B gene, and that the C4 gene is characteristically C4A like in size. In addition, it has been shown to encode the C4A isotypic determinants (Yu and Campbell, 1987). All of the restriction fragments mapped in the cloned DNA are in agreement with these findings, and support the hypothesis that the deletion which removes the intervening loci must have occurred either between the very 3' ends of the C4A and C4B genes, or, in the regions between the 3' ends of the C4 and the 5' ends of the 21OH loci.

3.2.2.1 Isolation of Walking Probes

Restriction fragments from the cosmid inserts were blotted onto nitrocellulose and probed with total human genomic DNA, radiolabelled by nick translation. After washing and autoradiography, X-ray films were compared with photographs of the ethidium bromide stained gels to identify which bands appeared to be devoid of repetitive elements. Potentially unique sequences were isolated from LGT agarose (2.5.5), radiolabelled by random hexanucleotide priming (2.7.1) and hybridised with Southern blots of Bam HI and Bgl II digested genomic DNA. Genomic samples were taken from the cell line used to prepare the library, and from a control individual, to ensure that no rearrangements of chromosome 6 had been introduced as a result of the EBV transformation procedure. No restriction fragment polymorphisms were obtained with any of the walking probes isolated from the cloned DNA.

Fragments which were confirmed to be non-repetitive from the results of hybridisation with genomic DNA were linked to pre-existing markers of the class III region by common bands on Southern blots from PFGE. Once linkage had been established, these unique regions were used as probes for chromosome walking.

3.2.3 Cosmid Chromosome Walking

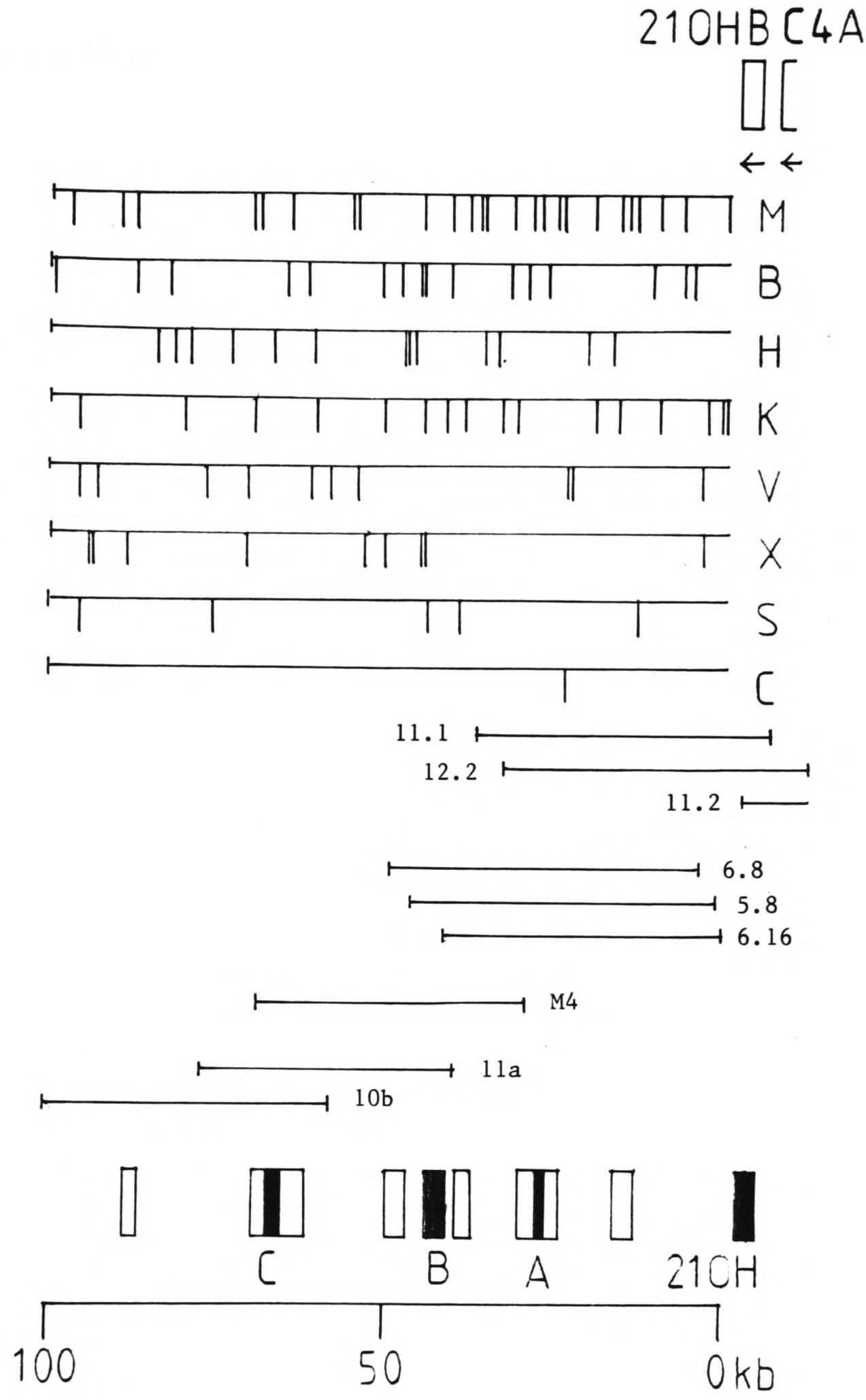
(a) From the 21OHB locus

In order to walk from the 21OHB gene, a new probe had to be isolated from the furthest extending cosmid, cos11.1. Hybridisation of radiolabelled genomic DNA to a Southern blot of restriction enzyme digested cosmid DNA suggested that a 3 kb Bam HI fragment and an overlapping 3 kb Bgl II fragment were potentially non-repetitive (Fig. 3.6). These results were confirmed by other double digest analyses (not shown). A subclone of the 3 kb Bgl II fragment was used to prepare a 1.4 kb Bam HI/ Bgl II probe (probe A). This was hybridised to a Southern blot of genomic DNA cleaved with Bam HI. The results showed that the probe detected fragments of the correct size, as deduced from the restriction map of cos11.1, and gave no background smear characteristic of residual repetitive DNA (Fig. 3.6F).

Probe A was used to rescreen the cosmid library. It detected 56 positives, of which 3 did not contain any 21OH sequence and were selected for further analysis. For the second walk, a 4 kb Bam HI fragment (B) was isolated from the end of cos6.8 (Fig. 3.8). This was used to obtain one cosmid from the Mbo I library, and one from a Bam HI library (see below). The final cosmid in the cluster around 21OHB was

Fig. 3.8

Restriction map of the genomic DNA isolated by chromosome walking from 210HB. Cosmid genomic inserts were mapped using Bam HI (M), Bgl II (B), Hind III (H), Kpn I (or its isoschizimer Asp 718) (K), Eco RV (V), Xho I (X), Sal I (S) and Cla I (C). The 5' to 3' orientations of the known genes are indicated by horizontal arrows. The limits of the cosmid inserts are shown by the horizontal bars (⊢⊣). Regions of unique DNA sequence are defined below the cosmid inserts by the open boxes, and the probes used to isolate the cosmids, by the shaded boxes.


21OHB C4A
M
B
H
K
V
X
S
C
11.1
12.2
11.2
6.8
5.8
6.16
M4
11a
10b
C
B
A
21OH
100
50
0kb

detected using a 2.2 kb Nco I/ Bgl II probe (C) derived from a subclone of cosM4 (Fig. 3.8).

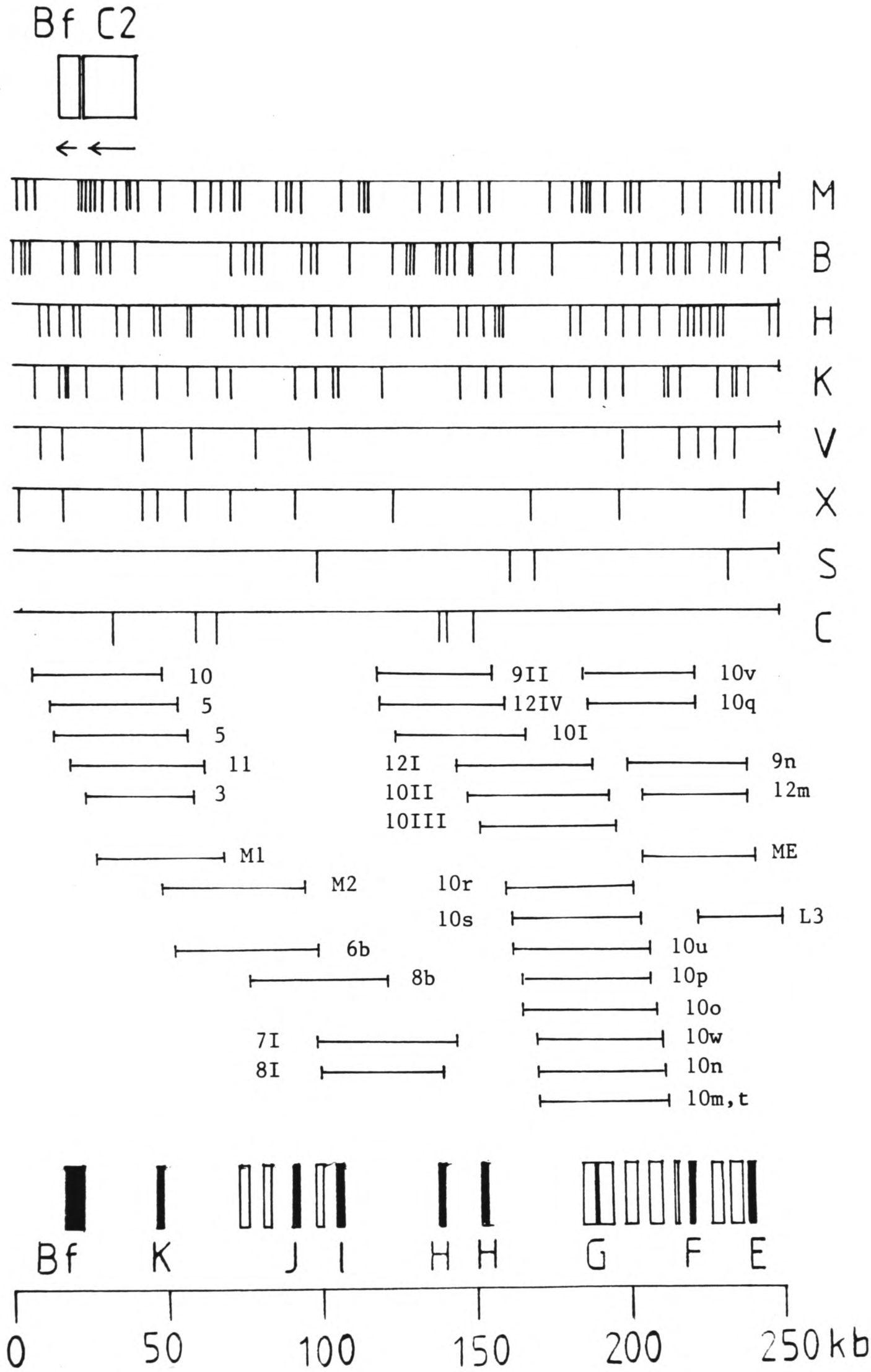
The six cosmids were characterised by restriction mapping, as described in 3.2.2, to extend the cloned region to 102 kb beyond the 21OHB locus (Fig. 3.8). During this time, the orientation of the class III genes within the human MHC was established using PFGE (Dunham et al., 1987; see section 3.2.4.1), and the 21OHB gene was found to lie at the centromeric end of the known class III gene cluster. The gap between the 21OH and DRA genes was estimated at 300-350 kb. As no DRA clones were isolated from the cosmid libraries, bidirectional walking between these markers was impossible. In contrast, isolation of cosmids containing the genes for tumour necrosis factors (TNF) α and β provided a second reference point from which to clone the sequences between C2 and TNF α, an estimated 390 kb. See section 3.2.3 (c).

(b) From the C2 locus

The extension obtained from C2 to the end of the original cosmid cluster was 22 kb. No unique sequences were isolated within the 15 kb beyond probe K, or from subclones of the last 10 kb of the cloned region. All 1.5x 10^6 recombinants from the Mbo I libraries had been screened with probe K to obtain the six positives analysed (3.2.2). Therefore, the 1.6 kb Hind III fragment was hybridised with filters from a small library of 1x 10^5 independent recombinants prepared by Bam HI partial digestion of DNA from an individual HLA typed as A1, B17, C4 A6B1, BfS, DR7/ A1, B8, C4 AQOB1, BfS, DR3 (M. J. Anderson, unpublished). Two positives (cosM1, cosM2; Fig. 3.9) were recovered. Restriction mapping showed that these cosmids overlapped the original cluster by 34.5 kb, and extended it by 28.5 kb. A new walking probe was

Fig. 3.9

Restriction map of the genomic DNA isolated by chromosome walking from the C2 gene. Cosmid genomic inserts were mapped using Bam HI (M), Bgl II (B), Hind III (H), Kpn I (or its isoschizimer Asp 718) (K), Eco RV (V), Xho I (X), Sal I (S) and Cla I (C). The 5' to 3' orientations of the known genes are indicated by horizontal arrows. The limits of the cosmid inserts are shown by the horizontal bars (⊢⊣). Regions of unique DNA sequence are defined below the cosmid inserts by the open boxes, and the probes used to isolate the cosmids, by the shaded boxes.


Bf C2
M
B
H
K
V
X
S
C
10
5
5
11
3
M1
M2
6b
8b
7I
8I
9II
12IV
10I
12I
10II
10III
10r
10s
10u
10p
10o
10w
10n
10m,t
10v
10q
9n
12m
ME
L3
Bf
K
J
I
H
H
G
F
E
0
50
100
150
200
250 kb

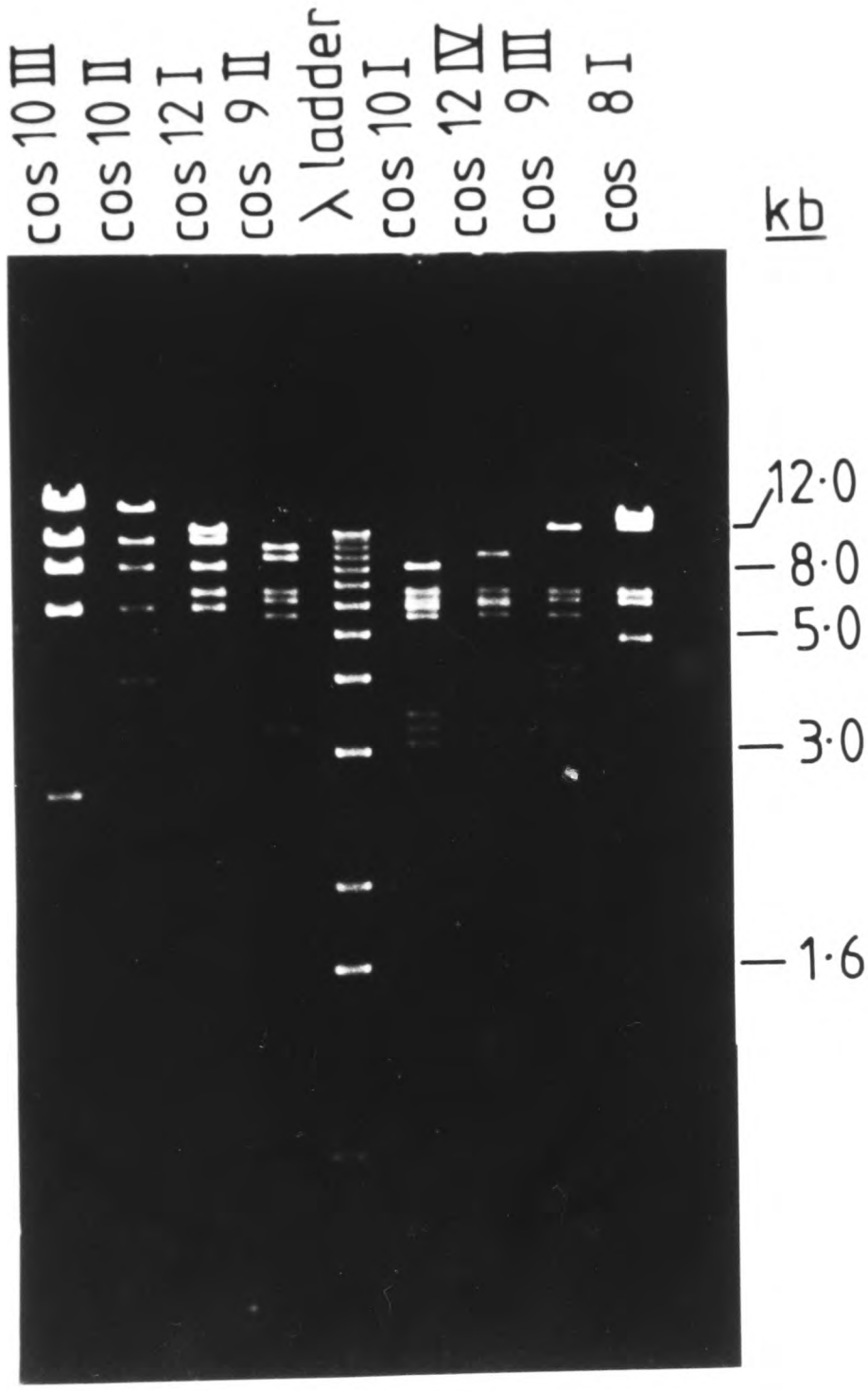
prepared from this extension. The 0.9 kb Bam HI/ Xho I genomic fragment (J) was located 48 kb from the 5'-end of the C2 gene. On rescreening the Mbo I library (5x 10^5 recombinants) probe J detected 15 positives. Two of these were taken for further characterisation on the basis of hybridisation to a restriction fragment from the region between probes J and K. These were independently mapped with the standard restriction enzymes to confirm that the genomic fragments cloned from the Bam HI library did not contain any polymorphic sites or chromosomal rearrangements. The restriction maps for the region in common between the two libraries were identical with all enzymes used. The insert of one recombinant was found to overlap the cosmid cluster isolated with probe K by 10 kb, and the other insert extended the DNA cloned from the Bam HI library by 27.5 kb.

A new probe, I, was isolated from a position ∿12kb telomeric to probe J. This was used to rescreen the Mbo I library, and to isolate two more cosmid clones. The last 0.8 kb from one of these, cos7I, was recovered from a Bgl II/ Sal I digest, and taken as the next walking probe (H). The new cosmids fell into three categories. One, like cosmids 7I and 8I, contained a single copy of probe H, and overlapped with the available restriction data by 34 kb. Three more cosmids also contained a single copy of probe H and a region which appeared to cross-hybridise with the probe, but could not be aligned with the molecular map. A final group of three cosmids was found to have restriction sites in common with both these clusters (Fig. 3.10). A more detailed characterisation of the 7 cosmids from this screen showed that probe H lay within a duplicated 7 kb Bam HI fragment, the copies (H_1 and H_2) being separated by 11 kb. H_1 could be distinguished from H_2 in a Bam HI/ Hind III double digest, as the 7 kb Bam HI fragment further from C2 was cut to give a hybridising fragment of 2.4 kb (see chapter VI). Owing to the extensive overlap

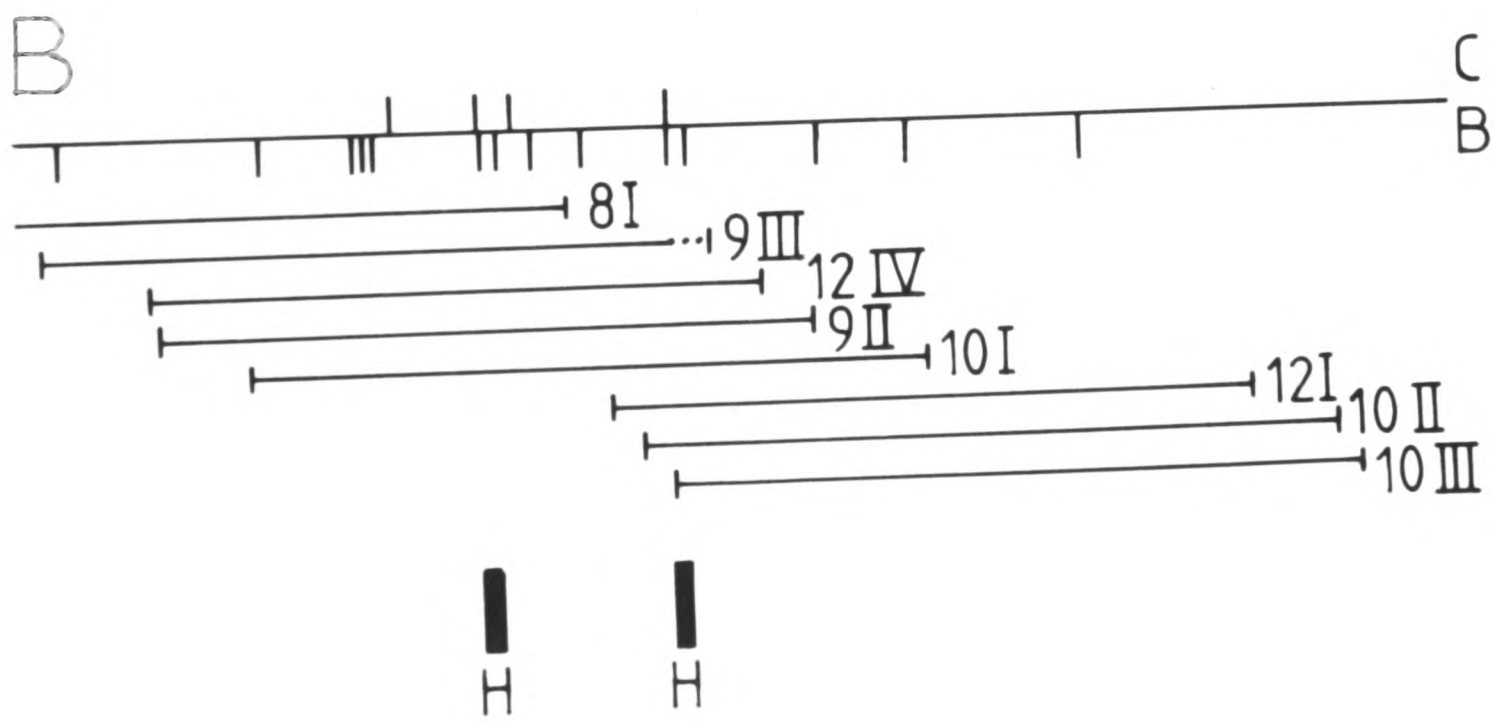
Fig 3.10

Clones isolated with probe H digested with Bgl II/ Cla I (Fig. 3.10A). Lane 1 = cos10III, 2 = cos10II, 3 = cos12I, 4 = cos9II, 5 = cos10I, 6 = cos 12IV, 7 = cos9III and 8 = cos8I. From the restriction data, the cosmids could be divided into three groups, depending upon the number of copies of probe H which they contained. The overlapping inserts are shown by the horizontal bars (⟷) in Fig. 3.10B.

A


cos 10 III
cos 10 II
cos 12 I
cos 9 II
λ ladder
cos 10 I
cos 12 IV
cos 9 III
cos 8 I
kb
12·0
8·0
5·0
3·0
1·6

B


C
B
8I
9III
12 IV
9II
10 I
12I
10 II
10 III
H
H

between these cosmids and the pre-existing molecular map, and between the cosmids within this cluster, it was unlikely that the results obtained with probe H represented a cloning artifact. Furthermore, probe H and a unique 0.6 kb Bam HI fragment (G) isolated from the telomeric end of the cluster were both linked to known class III region loci by hybridisation to common fragments on PFGE blots (see 3.2.4.2). Final confirmation that probe H represented a duplicated locus came from studies of cloned and chromosomal DNA, as discussed in chapter VI. In total, this single walk extended the map by 52 kb beyond the endpoint of cos7I.

Probe G was used to screen the library for the next set of positives. Twenty-eight were detected, of which 12 were taken for rescreening, and 11 were characterised. A 1.3 kb Bgl II fragment (F) from the most telomeric clone (cos10q) was used to obtain the last 2 cosmids from the Mbo I library.

The four walks made from this end of the cosmid cluster, brought the extension from C2 to 195.5 kb (Fig. 3.9), and the total of the cloned genomic DNA encompassing the 21OHB and complement loci to 381.5 kb. The final genomic probe (E) failed to detect new recombinants from either the Mbo I or the Bam HI libraries. Therefore, the last 11.5 kb was obtained from a Hind III genomic library prepared in a lorist 6 vector (a gift from T. Rabbitts). No suitable probe fragment was mapped within this region.

(c) From TNF α

Genes for TNF α and TNF β were mapped within the murine MHC by analysis of restriction fragment length polymorphisms in inbred mouse strains (Muller et al., 1987a; Gardner et al., 1987). Subsequent linkage

of cosmid clusters (Muller et al., 1987b) positioned the tumour necrosis factor genes 70 kb away from the mouse class I D locus. By analogy with the murine H-2 complex, the human genes would lie between the complement/ 21OH gene cluster and HLA-B. Analysis of human MHC deletion mutants by genomic Southern blot hybridisation appeared to confirm that TNF α lay within the HLA, but could not establish the exact position (Spies et al., 1986).

The Mbo I library was screened with a TNF α genomic probe (a gift from Mark Bodmer) to isolate 9 positives, of which 6 were fully characterised. The cluster covered 82.5 kb around the TNF α locus extending 40 kb from both the 3' and the 5' ends of the gene (Fig. 3.11). The position of the TNF β locus was deduced from comparison of the restriction data with the published maps of Nedospasov et al. (1985, 1986). The 30 kb around the TNF loci were found to be identical for the restriction enzymes Bam HI, Hind III, Kpn I and Xho I, which had been positioned in the bacteriophage clones.

To use the TNF cluster as a reference point from which to initiate a second walk, the orientation of the loci and the distances between the cloned genes were determined from PFGE (Dunham et al., 1987). A 1.4 kb genomic Bam HI fragment (L), located 10.5 kb 3' to TNF α was used in conjunction with probe J to show that C2 was telomeric to 21OHB, and that TNF α was centromeric to TNF β. The region from C2 to TNF α was estimated at 390 kb (see 3.2.4.1). To close the gap between the two clusters, a 0.4 kb Bam HI/ Hind III fragment (M) was isolated from the centromeric end of cos8.2 (Fig. 3.11). Six new recombinants were analysed as before, and a second probe (N) recovered for rescreening the library. The three cosmids detected by probe N only extended the previous cluster by ∿10kb (Fig. 3.12). A 0.6 kb Bgl II fragment (O) from the end of cos5A did not hybridise to any clones from the Mbo I

Fig. 3.11

Restriction map of the genomic DNA isolated with the TNFα probe. Cosmid genomic inserts were mapped using Bam HI (M), Bgl II (B), Hind III (H), Kpn I (or its isoschizimer Asp 718) (K), Eco RV (V), Xho I (X), Sal I (S) and Cla I (C). The 5' to 3' orientations of the known genes are indicated by horizontal arrows. The limits of the cosmid inserts are shown by the horizontal bars (⊢⊣). Regions of unique DNA sequence are defined below the cosmid inserts by the open boxes, and the probes used to isolate the cosmids, by the shaded boxes.

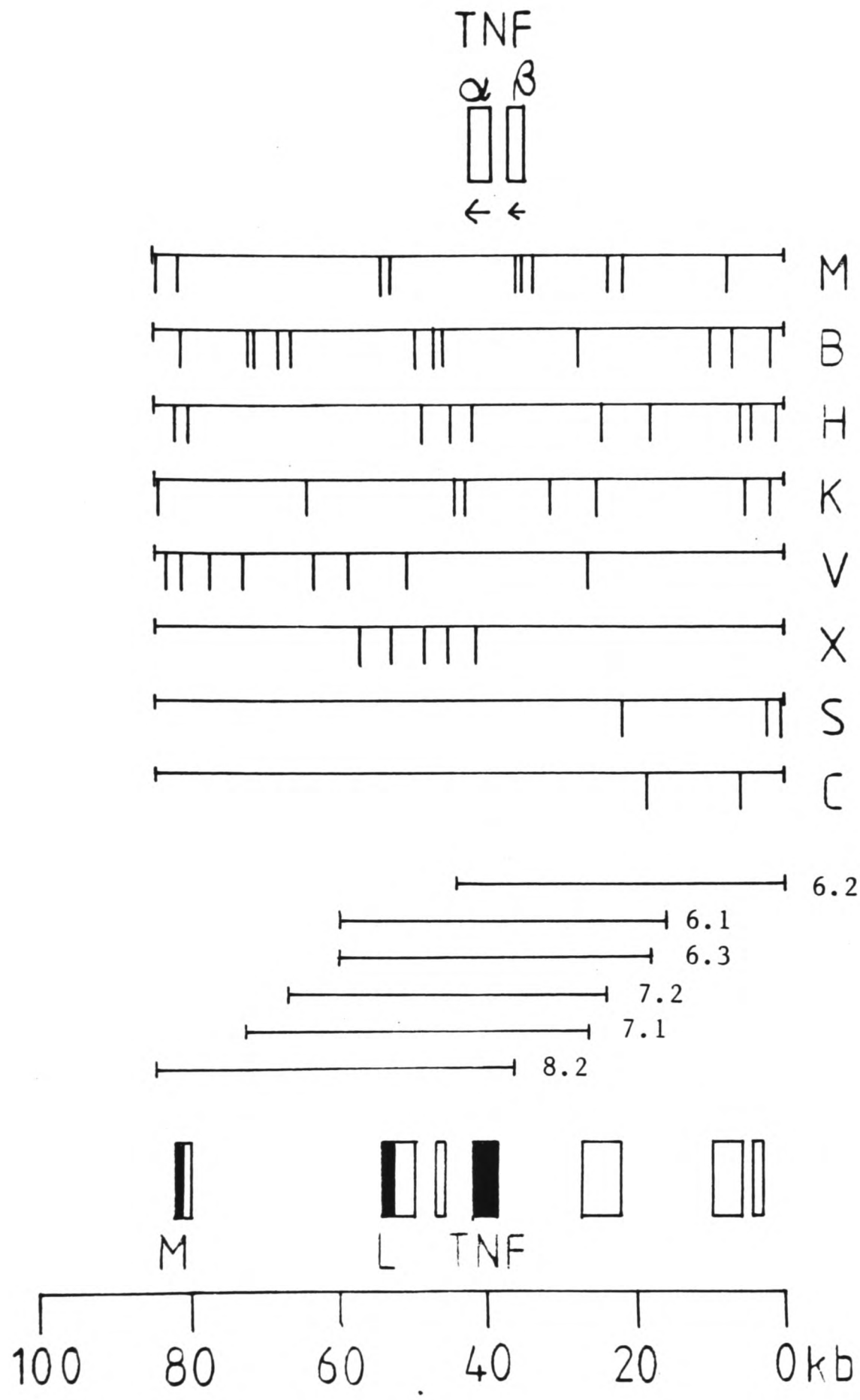
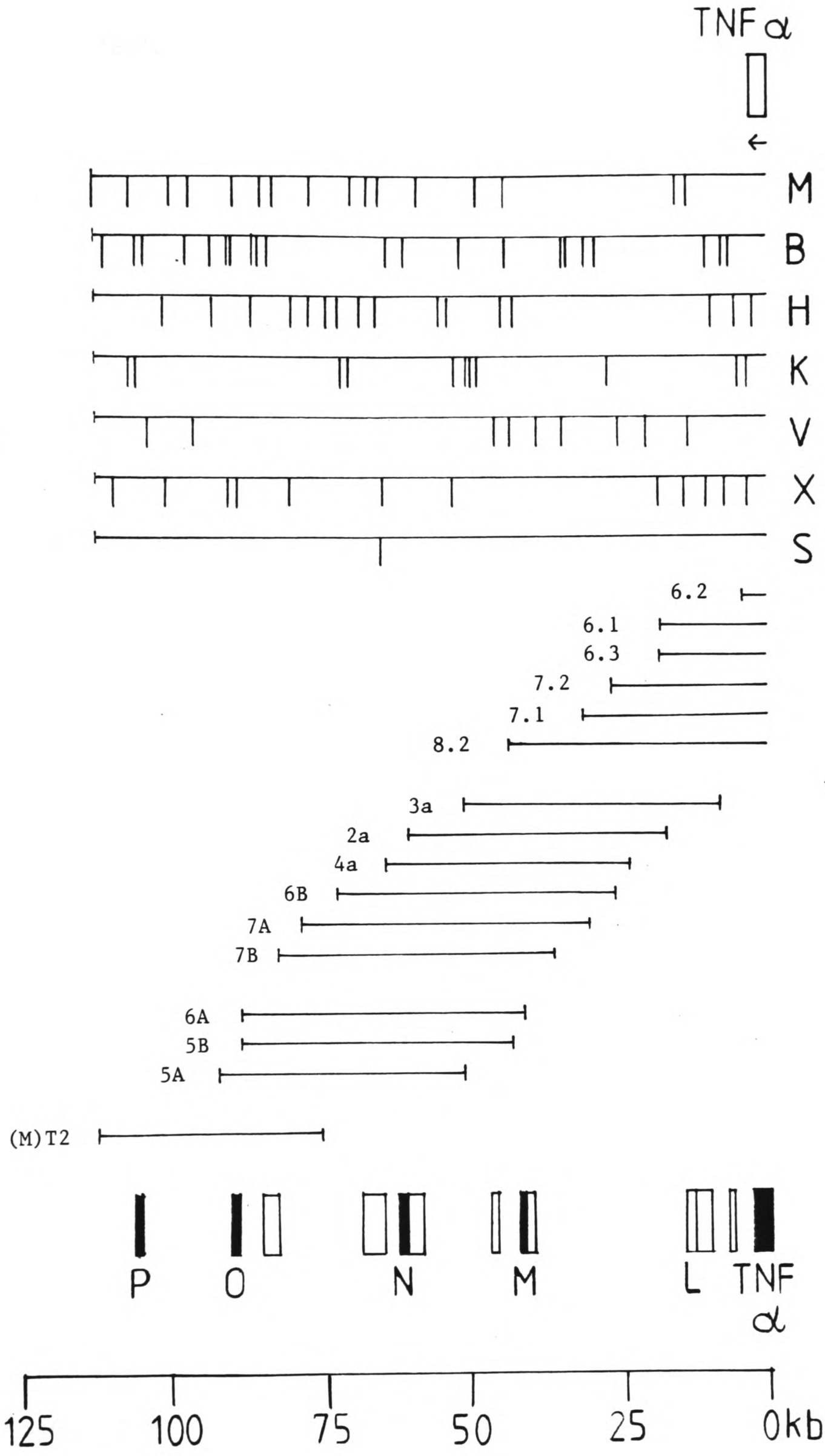

TNF
α
β
M
B
H
K
V
X
S
C
6.2
6.1
6.3
7.2
7.1
8.2
M
L
TNF
100
80
60
40
20
0kb

Fig. 3.12

Restriction map of the genomic DNA isolated by chromosome walking from the TNF gene cluster. Cosmid genomic inserts were mapped using Bam HI (M), Bgl II (B), Hind III (H), Kpn I (or its isoschizimer Asp 718) (K), Eco RV (V), Xho I (X), Sal I (S) and Cla I (C). The 5' to 3' orientations of the known genes are indicated by horizontal arrows. The limits of the cosmid inserts are shown by the horizontal bars (⟷). Regions of unique DNA sequence are defined below the cosmid inserts by the open boxes, and the probes used to isolate the cosmids, by the shaded boxes.


TNFα
M
B
H
K
V
X
S
6.2
6.1
6.3
7.2
7.1
8.2
3a
2a
4a
6B
7A
7B
6A
5B
5A
(M)T2
P
O
N
M
L
TNF
α
125
100
75
50
25
0kb

libraries. Instead, the final positive was obtained from the Bam HI library. Probes from this insert were found to hybridise to restriction fragments of identical size in both Bam HI and Bgl II digests of DNA from the individuals used to prepare the Bam HI and the Mbo I libraries. In addition, all of these probes were linked to the pre-existing MHC markers by PFGE, confirming their position within the class III region. However, no new recombinants could be detected by hybridisation with the Bam HI and Hind III lorist library filters. One clone was found on screening the 1.5x 10^6 cosmids of the Mbo I libraries, but this failed to rescreen.

In total, the walk from the TNF loci encompassed 148 kb, with an extension of 108 kb from the 3' end of the TNF α gene (Fig. 3.12).

3.2.4 Pulsed-field Gel Electrophoresis Analysis of Walking Probes

In addition to the standard enzymes used to characterise the cosmid inserts, DNA was cleaved with the rare cutters Mlu I, Not I, Nru I and Pvu I. All of these enzymes recognise sequences containing at least one CpG dinucleotide, and are methylation sensitive, thus cutting less frequently in genomic DNA than in the cloned DNA. (See chapter IV for further details.) Southern blots of restricted genomic DNA separated by PFGE were supplied by Ian Dunham.

3.2.4.1 Orientation of the Class III Genes within the MHC (Dunham et al., 1987)

The sites for Mlu I, Not I, Nru I and Pvu I were mapped in the cloned DNA by single digests and double digests in combination with the reference enzymes BamH I and Bgl II. By comparing which of the class III

region probes hybridised to common fragments on Southern blots from PFGE, the sites which cleaved in the chromosomal DNA could be correlated with those in the cloned DNA (see chapter IV).

A 980 kb Not I band in common with the class II subregions DR and DQ and the complement loci was found to extend 27 kb 5' to the C2 gene. A probe (J) which lay beyond the Not I site mapped in the cloned DNA hybridised to a separate fragment estimated at 210 kb. Therefore, C2 was shown to be telomeric to 21OH. Similarly, TNF α was linked to the class I subregions HLA-B and HLA-C by hybridisation to a common 780 kb Pvu I fragment. No digest linked TNF α to the complement genes. However, a probe (L) 10.5 kb 3' to TNF α was isolated from the cosmid cluster. This was shown to lie beyond an Nru I site ~3 kb from the TNF α locus. Probe L hybridised to a different Nru I fragment from TNF α, but detected the 640 kb fragment in common with the other known class III region probes. From these data, the TNF α locus could be positioned centromeric to TNF β by comparison with the published maps of Nedospasov et al. (1985, 1986). Furthermore, the distances between 21OHB and DRA, C2 and TNF α, and C2 and HLA-B were estimated at 300, 390, and 650 kb respectively (Fig. 3.13).

3.2.4.2 Linkage of Walking Probes by PFGE

To ensure that the clones isolated represented contiguous DNA from chromosome 6, all walking probes were linked to pre-existing class III markers by PFGE (Fig. 3.14). A blot of Not I, Nru I and Pvu I single and double digests of genomic DNA from the cell line used to prepare the library was hybridised with the unique sequences after they had been tested on standard genomic Southerns. The blot was stripped between each probing as described in 2.7.5.

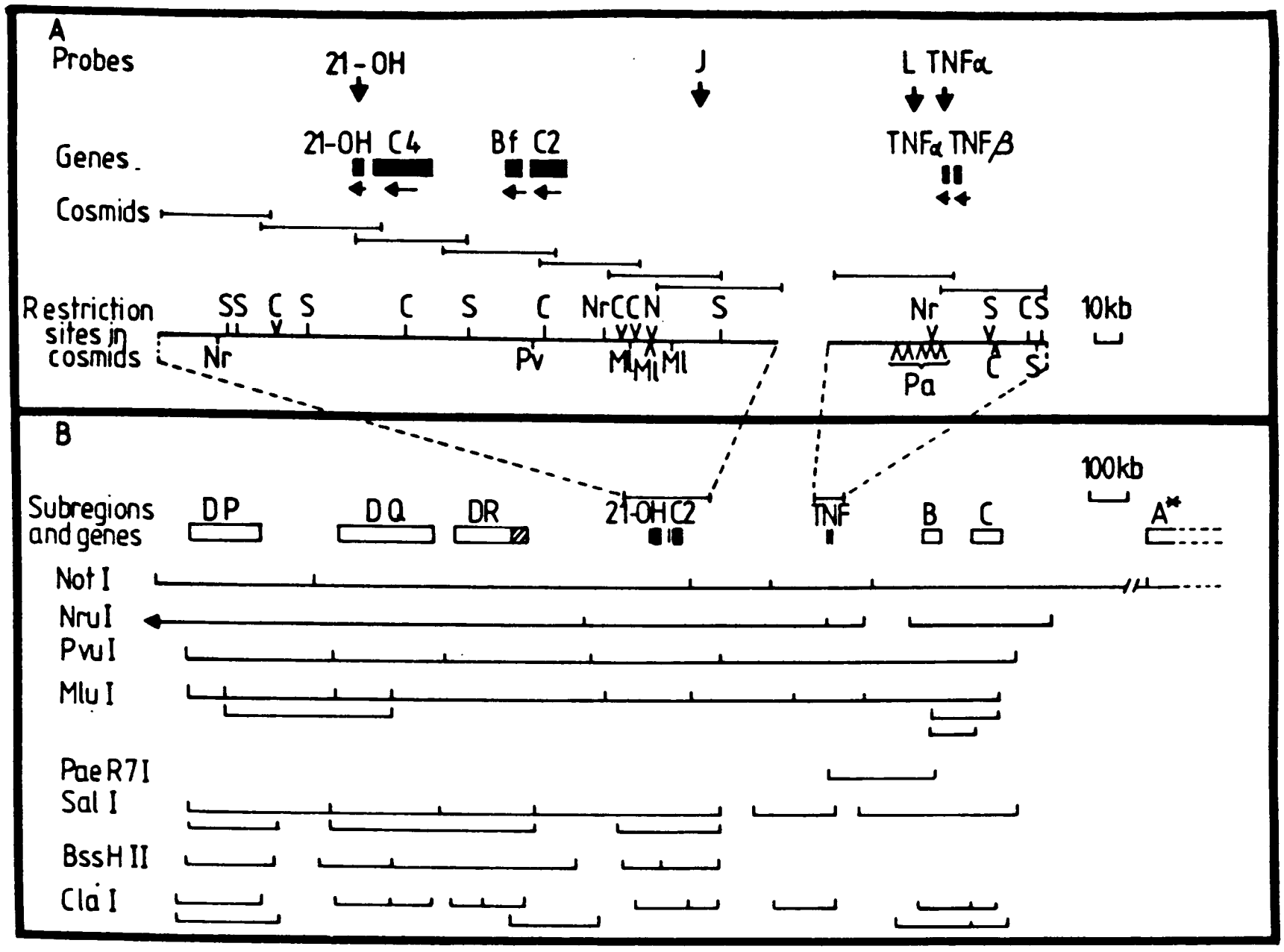


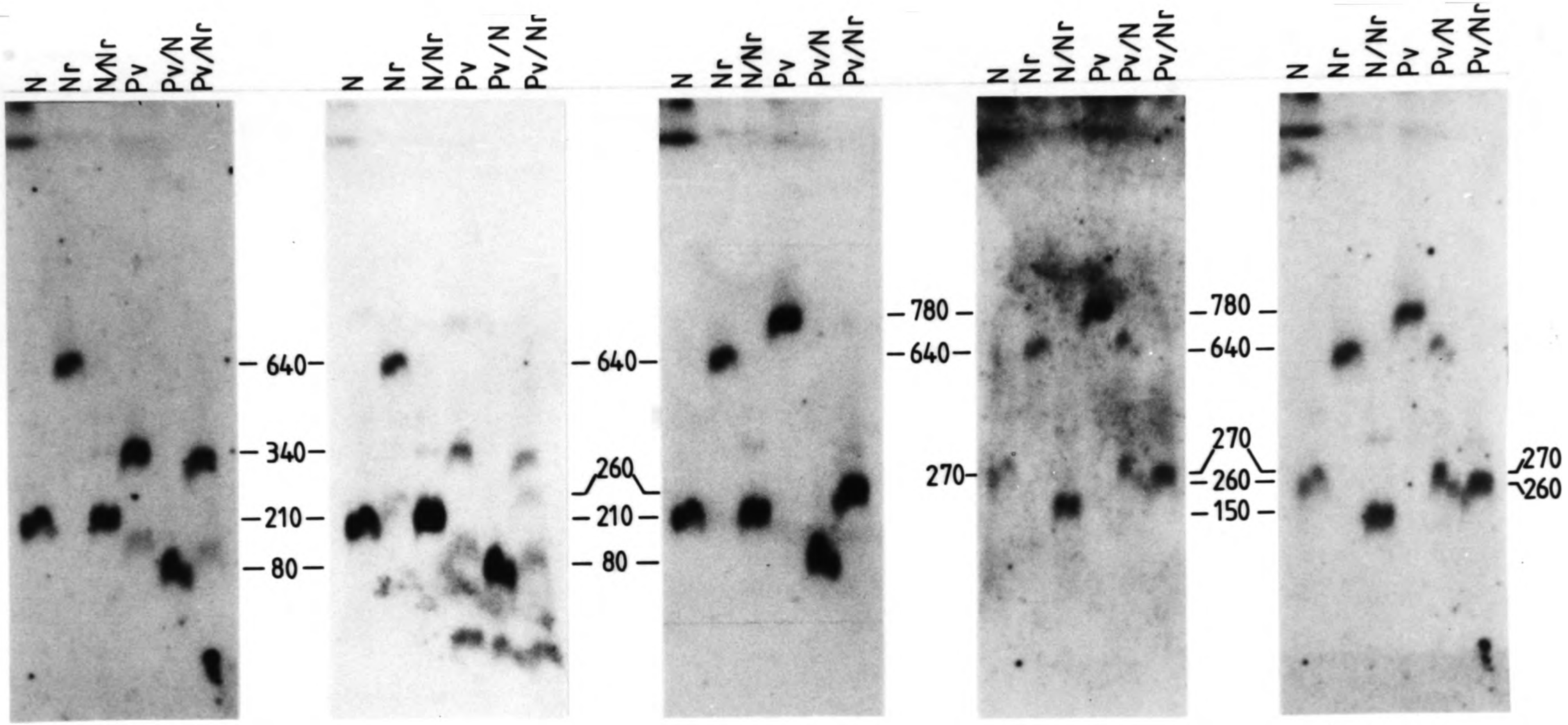
Fig 3.13

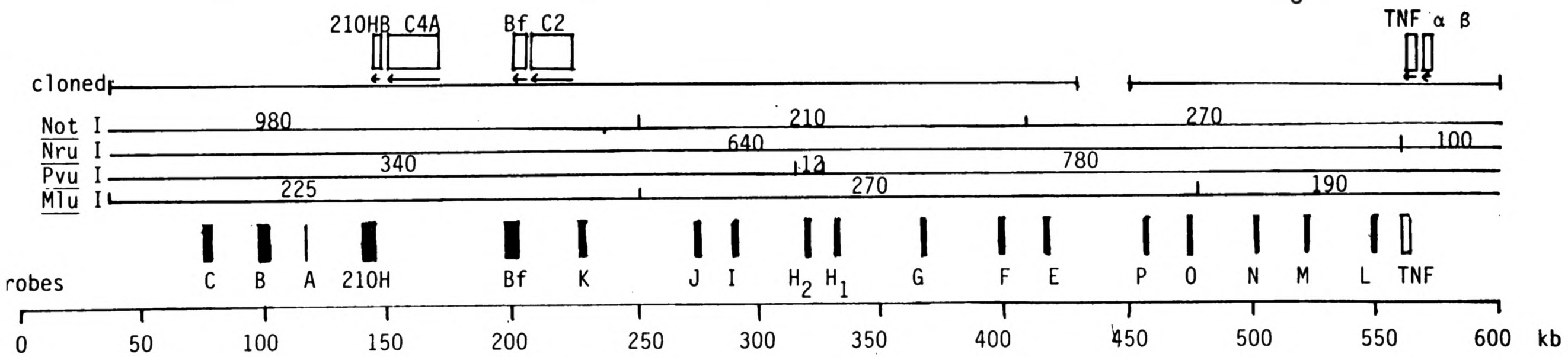
The long-range map of the human MHC (Dunham et al., 1987). Fig. 3.13A shows overlapping cosmid clones from the complement/ 21OH and TNF regions. The positions of the genes are indicated by the solid boxes, and the 5' to 3' orientation by the horizontal arrows. Enzyme sites known to cleave in chromosomal DNA are marked by arrow heads, other sites are indicated by vertical bars. (C = Cla I, S = Sal I, Nr = Nru I, N = Not I, Ml = Mlu I, Pa = Pae R71, Pv = Pvu I.) Fig. 3.13B shows the restriction map obtained from Southern blot analysis. The exact positions of genes are shown by solid boxes, and the open boxes define the limits of the region hybridising to the probes. The cross-hatched box shows the probable position of the DRA gene.

Fig. 3.14

Linkage of walking probes from across the class III region. The probes were hybridised to a Southern blot of genomic DNA digested with Not I (N), Nru I (Nr) and Pvu I (Pv) in single and double digest combinations. The samples were separated using a pulse time of 65 s. The probes which gave the results displayed are indicated on the first line below the autoradiographs. Probes which hybridised to identical fragments are designated below. The map in Fig. 3.14B shows the relative positions of the probes, and the fragments (kb) attributed to the class III region.

A


N
Nr
N/Nr
Pv
Pv/N
Pv/Nr
640
340
210
80
260
780
270
150
Probe J
I
Probe H
Probe F
G
Probe P
E
Probe L
M
N
O


210HB C4A
Bf C2
TNF α β
cloned
Not I
Nru I
Pvu I
Mlu I
980
210
270
640
100
340
12
780
225
270
190
Probes
C
B
A
210H
Bf
K
J
I
H2
H1
G
F
E
P
O
N
M
L
TNF
0
50
100
150
200
250
300
350
400
450
500
550
600
kb

All the walking probes hybridise to a common 640 kb Nru I fragment, the end-point of which has been established at a position 2.5 kb 3' to TNF α. The probes centromeric to 21OHB also hybridise to a Not I fragment of 980 kb, a Pvu I fragment of 340 kb, a Not I/ Nru I fragment of 280 kb and a Not I/ Pvu I fragment of 270 kb. The telomeric end-points of all of these restriction elements have been mapped within the cloned cosmid inserts. All of the probes hybridise to single and double digest products which are consistent with these sites.

Probes J and I hybridise to the same Nru I and Pvu I fragments of 640 and 340 kb as the 21OH probe. However, they detect a different Not I fragment of 210 kb. Analysis of the cloned DNA shows that the Not I site cleaved in chromosomal DNA is 16 kb centromeric to probe J and 27 kb telomeric to the 5' end of the C2 gene.

Probe H, a 0.8 kb Bgl II/ Sal I fragment recovered from the end of a cosmid insert, maps to a position 92 kb telomeric to C2. This probe, which is duplicated, contains the Pvu I site which links the observed fragments of 340 kb, containing the complement genes, and 780 kb, containing the TNF genes. In addition, it detects a band of 12 kb, the distance between the duplications. Probe H also hybridises to the 210 kb Not I fragment and the 640 kb Nru I fragment in common with probes J and I.

Probes G and F both hybridise to the same fragments of 210 kb in the Not I and Not I/ Nru I digests and 640 kb in the Nru I digests. They are linked to TNF α by a common Pvu I fragment of 780 kb.

Probes E, P, O, N, M and L all hybridise to identical fragments in single and double digests with Not I, Nru I and Pvu I. The Pvu I fragment of 780 kb is in common with probes H, G, F and TNF α. The Not I fragment is distinct from that detected by probes J to F, but identical to that of 270 kb which contains the TNF genes.

The Not I fragment at 210 kb has been cloned from the cosmid library, and is estimated at 155 kb. The discrepancy between the two figures appears to be caused by anomalous migration of the fragment in the OFAGE system because:

(a) resizing of the fragment on a "Waltzer" gel system gives an estimate of 150 kb,

(b) comparison of haplotypes shows no variation in the size of this fragment from different cell lines,

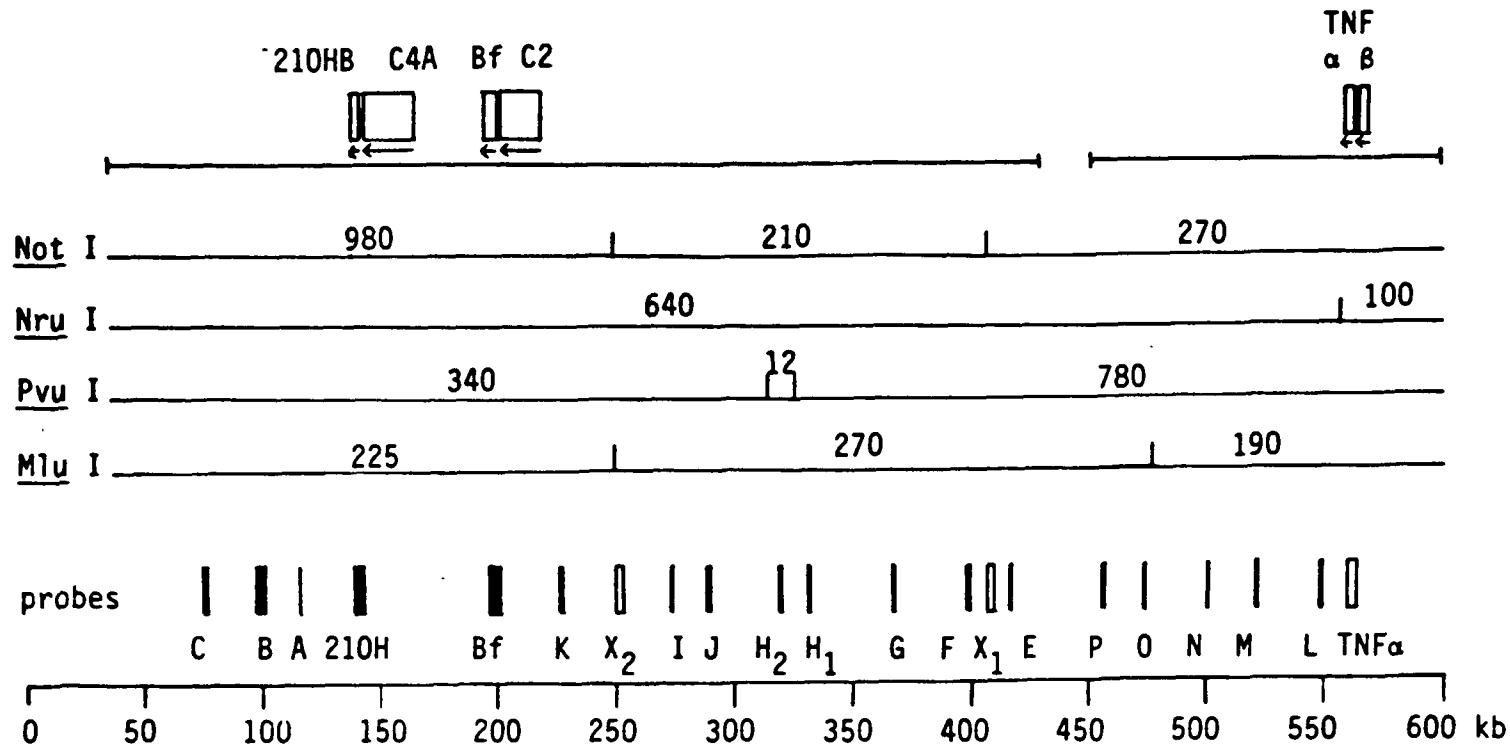
(c) fine-detail pulsed field mapping of the class III region with the enzymes Bss HII, Eag I and Sac II do not reveal any deletions in the cloned DNA with respect to the long range restriction map.

All of the Not I, Nru I and Pvu I single and double digest products have been resized using the "Waltzer" system. A comparison of the long-range maps shows that the sizes are at least 10% smaller than with the OFAGE apparatus, and are in closer agreement with the fragment lengths estimated from the cloned DNA (Fig. 3.15).

3.2.4.3 Linkage of the Cosmid Clusters by PFGE

From the Nru I/ Pvu I and Not I/ Nru I double digests, the fragments hybridising to probes E and P, which flank the gap between the cosmid clusters, are 260 and 150 kb. From the positions of the sites in the cloned DNA, the differences between the observed fragment sizes and the amount of DNA encompassed by the recombinants are 55 kb (260 - 205 kb) for the Nru I/ Pvu I digest and 22 kb (150 - 128 kb) for the Not I/ Nru I digest.

To obtain a more accurate estimate, the sites for the enzymes Bss HII, Eag I and Sac II were also mapped at a chromosomal level and within the cloned DNA. By using probes from the adjacent ends of the two cosmid



B

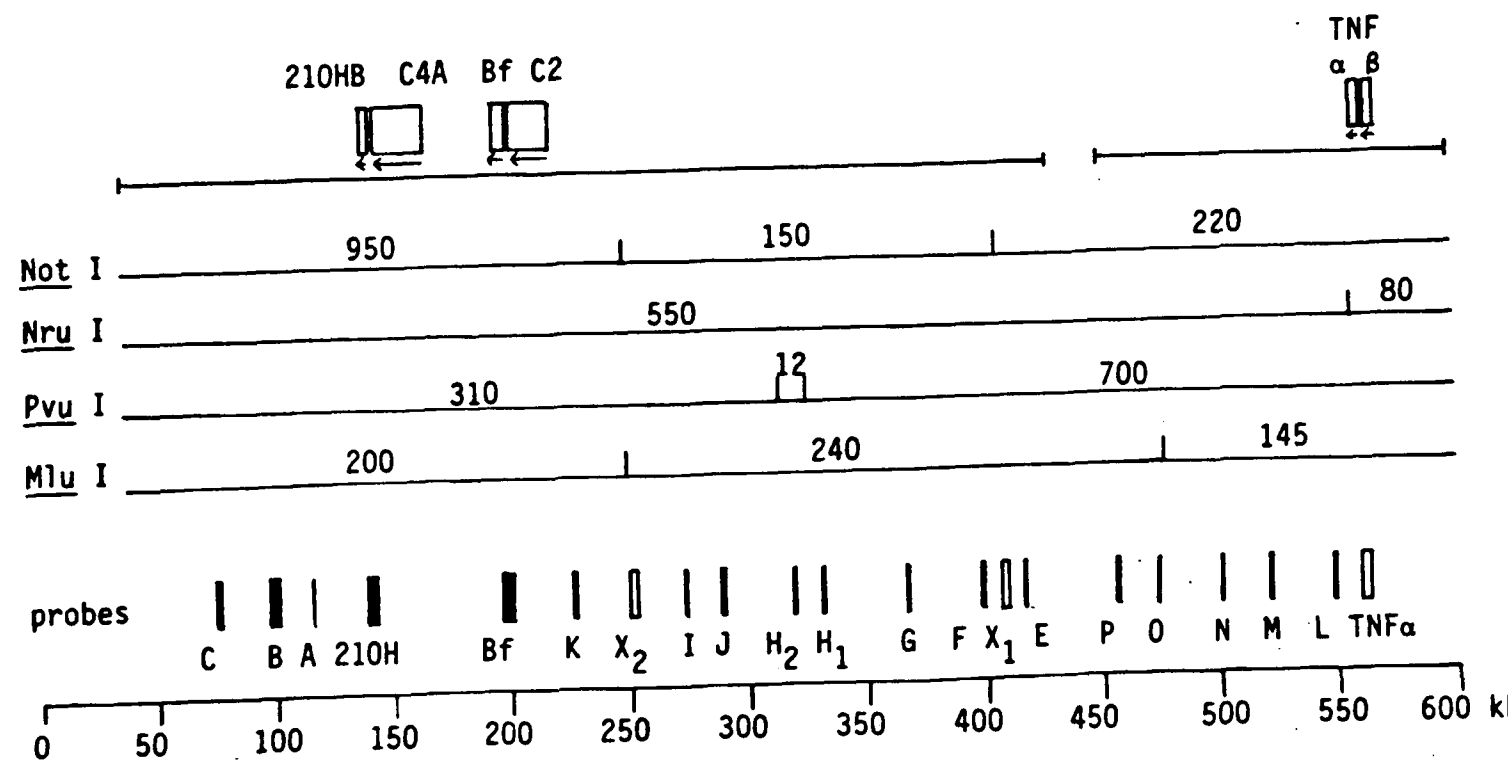


Fig. 3.15

Long-range maps of the class III region of the human MHC constructed using the OFAGE (A) and "Waltzer" (B) gel systems. The positions of the genes are shown by the open boxes, and the direction of transcription by the horizontal arrows. The sizes of the fragments are indicated above the horizontal bars.

clusters with PFGE Southern blots the lengths of the restriction fragments spanning the uncloned portion could be calculated. Chromosomal fragments of 70 kb (Bss HII), 110 kb (Eag I) and 118 kb (Sac II) were detected with probes P and E (Fig. 3.16), which predicts a gap of 22-23 kb. Both end-points of the observed fragments could be positioned in the cloned DNA by comparison with blots hybridised to other probes from the class III region. Within the TNF cosmid cluster the Bss HII, Eag I and Sac II sites are 20, 64 and 0.5 kb, respectively, from the centromeric end. Within the complement/ 210H cosmid cluster the sites are 28, 23 and 92kb from the telomeric end, respectively (see chapter IV).

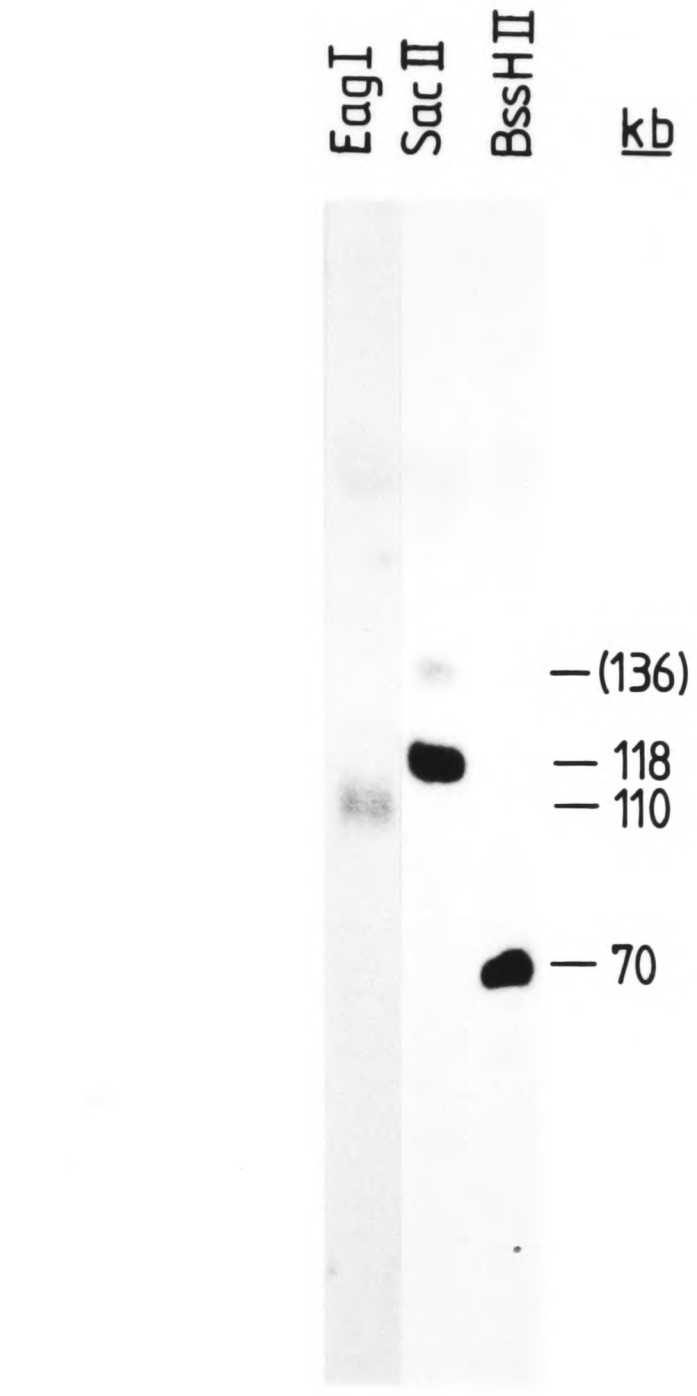
3.2.5 Summary

In total, 45 cosmids have been isolated by walking from the complement/ 210H gene cluster, and 16 cosmids have been isolated in a second walk initiated from TNF α. The inserts of all the recombinants have been mapped with the restriction enzymes Bam HI, Bgl II, Cla I, Eco RV, Hind III, Kpn I, Sal I and Xho I in single and double digest combinations. Together they encompass 541 kb of DNA, out of a possible maximum of 563 kb from the most centromeric to the most telomeric end-point, allowing for the 22-23 kb gap.

The joint map of the cosmid clusters is shown in Fig. 3.17, and includes all the walking probes. Only the Bam HI restriction sites are illustrated. For a complete restriction map, see appendix I.

Fig. 3.16

Estimation of the size of the gap between the cosmid clusters. The endpoints of the restriction fragments were defined in the cloned DNA. Probe E (shown) hybridised to the same Eag I (not shown) and Bss HII fragments as probe P. Probe P detected the same 136 kb partial product in the Sac II digest. The gap between the cosmid clusters, indicated by the break in the solid line (⊣ ⊢) is approximately 22-23 kb.


EagI
SacII
BssHII
kb
(136)
118
110
70

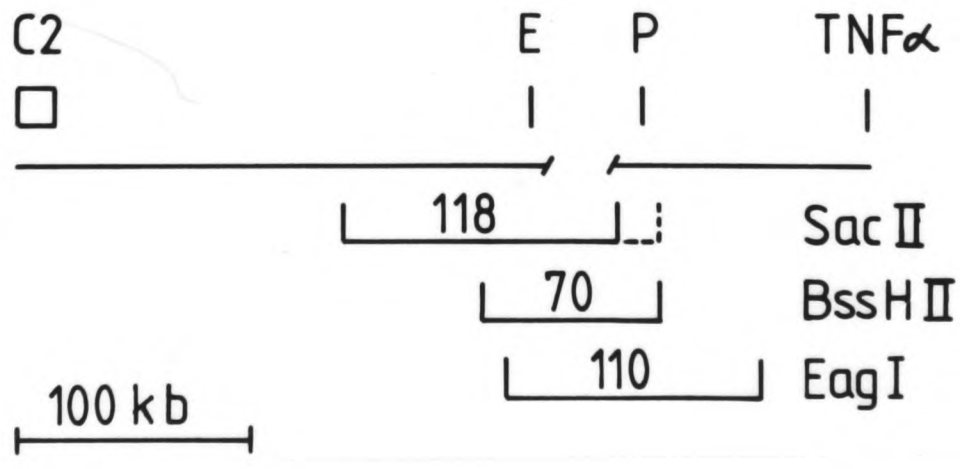
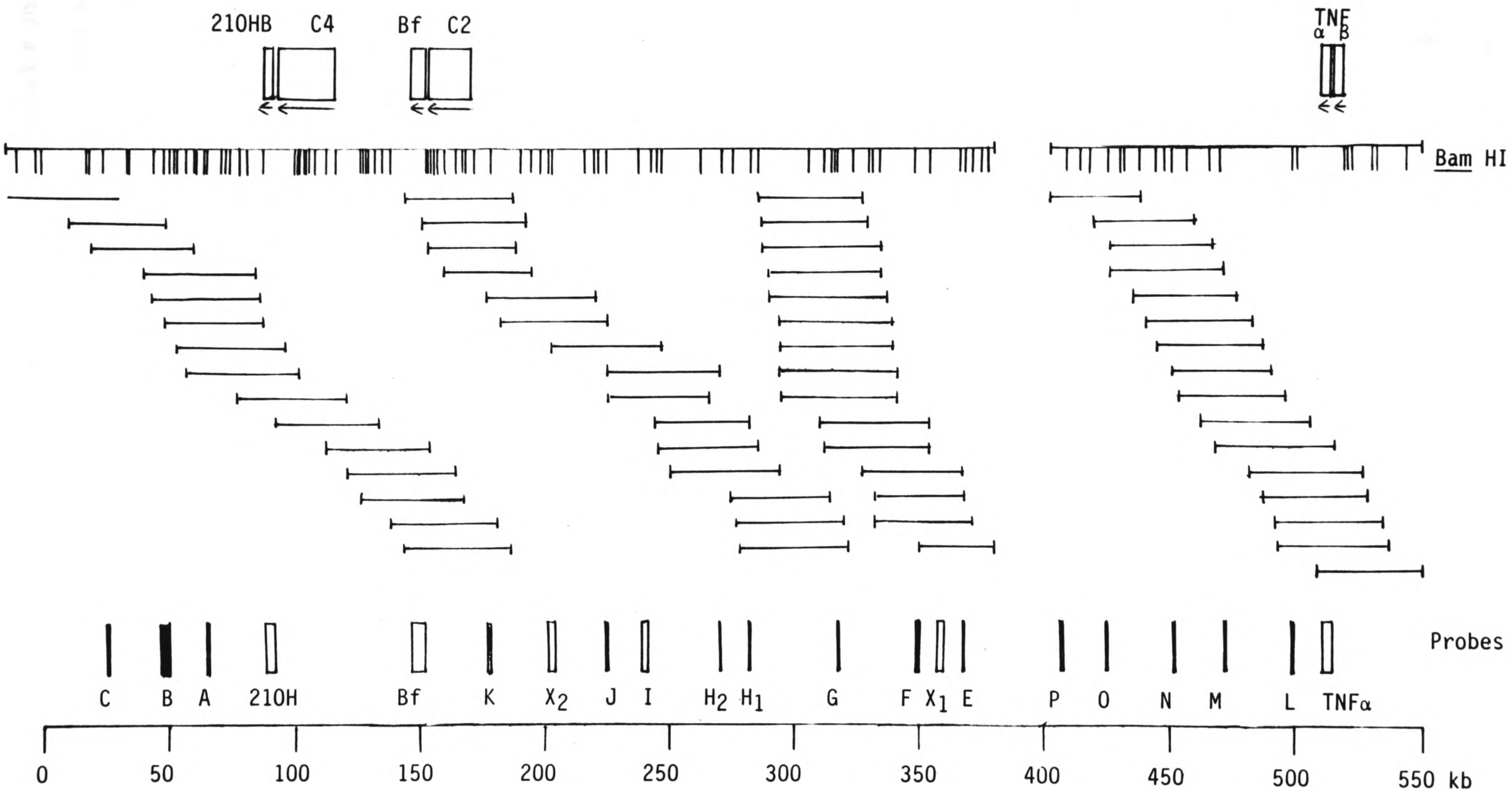

C2
E
P
TNFα
118
SacII
70
BssHII
110
EagI
100 kb

Fig. 3.17

A joint map of the cloned portion of the class III region of the human MHC. Only the restriction sites for Bam HI are shown. The 5' to 3' orientations of the known genes are indicated by the horizontal arrows. The limits of the cosmids are defined by the horizontal bars (⟝⟞), and the walking probes are shown as shaded boxes.


21OHB
C4
Bf
C2
TNF
α
β
Bam HI
Probes
C
B
A
21OH
Bf
K
X2
J
I
H2
H1
G
F
X1
E
P
O
N
M
L
TNFα
0
50
100
150
200
250
300
350
400
450
500
550 kb

3.3 DISCUSSION

3.3.1 Insert size and Library Complexity

The 55 cosmids isolated from the Mbo I library had an average insert size of 43 kb. The smallest insert was 38.2 kb, and the largest was 46.8 kb. This gives a total recombinant DNA molecule of 42.7 to 51.3 kb in length, or, 88% to 106% of the wild type bacteriophage λ genome (48.5 kb). This is within the constraints of the in vitro packaging system which lie at 79% to 109%.

The five cosmids from the Bam HI library (designated M in the figures) contained inserts of 36.7 to 44.6 kb with an average of 41 kb. The only Hind III lorist library cosmid (L3) was 28.1 kb. The smaller size of the latter arose from the restraints imposed by:

(a) the distribution of the Hind III sites within the genome, and, (b) the necessity to include two copies of the 5.1 kb lorist vector, to provide two cos sites for packaging.

Only 2 cosmids out of the 61 characterised appeared to have inserts with identical endpoints within the genomic DNA. On closer investigation, one of these was found to have reformed Bam HI at the vector cloning sites. Therefore, all the cosmids isolated and characterised arose from independent recombinant events.

Apart from the cosmid isolated with probe K, two other clones were found to contain small non-contiguous regions of DNA at one end of the insert. The frequency of these clones, 5%, is within the expected range for a library prepared with phosphatased genomic DNA (2-5%, Steinmetz et al., 1986).

Assuming an average insert size of 43 kb, and a haploid genome size of 3×10^9 base pairs, the complexity of a library which would allow the isolation of a given sequence can be calculated. For a 99% probability

of finding the sequence, the complexity is about 321,000 independent recombinants, as estimated from

$$N = \frac{\ln(1-P)}{\ln(1-i/h)}$$

where N= the number of clones per complexity, P= probability, h= haploid genome size, and, i= average insert size (Clarke and Carbon, 1976).

Therefore, with a library of 1.5×10^6 recombinants, there should be at least 4-5 positives for a given unique sequence probe. Failure to isolate a linking cosmid is unfortunate, as the probes are separated by only 40-41 kb, and may reflect an unusual structure within this region.

3.3.2 Growth of Cosmid Cultures and Representation of Genomic Sequences

The growth rates of cosmids appeared to vary considerably depending upon the region of the DNA from which the insert was cloned. Most cultures were dense to the eye after overnight incubation, but a few did not become cloudy until after 20 h of incubation. Those probes which were under-represented in the libraries often correlated with the most slowly growing cosmids. These clones were also the least stable of those isolated, and fresh stocks were always taken to prevent the isolation of deletion mutants. The frozen stocks of unstable cosmids were recovered in L-broth in the absence of antibiotic for 30 min to 1 h to allow expression of the $^{\beta}$-lactamase gene prior to setting up overnight cultures. Small amounts of DNA were retained at $-20^{o}C$ in case any recombinant required repackaging and transducing into NM 554 to replace a deleted stock.

The decreased growth rate of particular inserts could also be correlated with the failure to recover positives from the same region of genomic DNA on secondary screens. This can be explained by considering the relative growth rates of cosmids within the primary screen stock. If the positive recombinant replicates more slowly than other clones, it is "amplified out" during the 2 h recovery period which follows removal from the master plate. This prediction was confirmed in certain cases by rescreening at a very high density. Even with > 5,000 colonies per 9 cm plate, only one or two were positive for the under-represented probe. In contrast, sequences from around the C4A and 21OHB loci detected 60-65 positives for every 5×10^5 recombinants probed. With these cosmids, the first rescreen plates were 90-95% positive, and single colonies grew rapidly in liquid culture (visibly cloudy after 7 h).

The differential growth rates observed with cosmid inserts from specific genomic regions has been attributed to DNA sequences which suppress or enhance the host replication. Regions which are easier to clone may contain sequences which share homology with the bacterial replication origin, thus giving a selective advantage over other inserts. The inhibition of growth rate is harder to explain, although it has been suggested that certain sequences, such as direct or inverted repeats, are disadvantageous to the infected bacterium and are not propogated well within RecA$^-$ hosts (Wyman and Wertman, 1987). In addition, the expression of genes encoded by the insert, or the presence of strong promoter sequences, may interfere with the replication origin (Gibson et al., 1987).

Sequences which failed to clone in any of the libraries may have been lost owing to the method of encapsulating the recombinant DNA. E.coli K12 derived bacterial strains contain endogenous restriction endonucleases which recognise a heptomeric sequence containing

5-methylcytosine (Raleigh and Wilson, 1986). The system has been shown to operate *in vitro* with standard λ packaging extracts (Rosenberg, 1985 and 1987; Rosenberg *et al.*, 1985). Therefore, although the *RecA*$^-$ host strains contain mutations within the genes encoding for these endonucleases, cleavage of DNA prior to incorporation into phage heads would render the fragments unpackageable.

Finally, different restriction enzymes can alter the representation of the DNA sequences within the clones isolated. A summary of the distribution of the walking probes in table 3.2 shows that K was over-represented by approximately sixfold in the *Bam* HI library, whereas only one copy per complexity was isolated from the *Mbo* I library. The difference between the results may arise from the distribution of restriction sites within genomic DNA, and their accessibility to enzyme (Seed *et al.*, 1982). If there is a region of DNA which is particularly rich in sites accessible to enzyme, there is a bias towards isolating cosmids with end-points clustered in this region. Alternatively, in constructing a library, there may be portions of the genome which give partial digest products too small or too large to package with a λ system. This latter problem is overcome, in part, by choosing a range of partial digests around the optimal conditions (Seed *et al.*, 1982), but may also be alleviated by making several libraries prepared with enzymes that have 6 bp recognition sites, such as *Bam* HI, *Bgl* II or *Hind* III.

3.3.3 The Completed Cosmid Map

Two cosmid clusters have been isolated from within the class III region of the HLA by chromosome walking. These have been linked together by PFGE to establish the gene orientation and the distances between these loci and other MHC markers. Together, the inserts encompass 541 kb

Table 3.2

Representation of the probe sequences within the cosmid libraries.

Probe	*1° screen*	*2° screen*	*recombinants/ C*
C	*2*	*2*	*1.3*
B	*3 (Mbo I)*	*3*	*1.9*
	1 (Bam HI)	*1*	*3.4*
A	*56*	*56*	*35.9*
21OH	*64*	*64*	*41.1*
Bf	*3*	*3*	*1.9*
K	*12 (Mbo I)*	*5*	*1.1*
	3 (Bam HI)	*2*	*6.7*
J	*15*	*15*	*9.6*
I	*6*	*2*	*1.3*
H	*12*	*7*	*4.5*
G	*28*	*11/12*	*16.5*
F	*5*	*2*	*1.3*
O	*1 (Bam HI)*	*1*	*3.4*
N	*3*	*3*	*1.9*
M	*10*	*6*	*3.9*
TNF	*10*	*9*	*5.8*

C= one complexity.

With the exception of probe K, 5x 10^5 recombinants of the Mbo I library, and 1x 10^5 recombinants of the Bam HI library, were screened . For probe K, all 1.5x 10^6 Mbo I library recombinants were screened.

out of a possible maximum 563 kb from the most centromeric to the most telomeric endpoints. This represents about 50% of the estimated 1 to 1.1 Mb between HLA-DR and HLA-B (Dunham et al., 1987; Carroll et al., 1987).

The complement and 21OH genes have been mapped in 12 cosmids spanning 140 kb. Independent analysis by restriction enzyme digestion and Southern blotting has confirmed the order of genes in this haplotype (A2, B7, DR2, C2 C, Bf S, C4A 3, C4B QO) to be the same as that determined by Carroll et al. (1984), allowing for the deletion of the 21OHA and C4B loci. Expansion by cosmid chromosome walking has brought the portion cloned from around the complement cluster to 393 kb in 45 independent overlapping recombinants. From the TNF α locus, 16 clones encompassing 148 kb have been characterised in a second chromosome walk. The restriction map of the 30 kb around the TNF loci is identical with the published bacteriophage λ clones of Nedospasov et al. (1986) for the enzymes Bam HI, Hind III, Kpn I and Xho I. By comparison of the data, the position of the TNF β gene has been established.

The unique sequence probes isolated during chromosome walking have been used with PFGE to orientate the known class III genes within the MHC (Dunham et al., 1987). From the centromeric side the order is DRA, 21OHB, C4A, Bf, C2, TNF α, TNF β, HLA-B, which is identical to that determined for the analogous genes in the murine H-2 complex (Muller et al., 1987b). The distances between the loci were calculated to be 350 kb from DRA to 21OHB, 390 kb from C2 to TNF α and 250 kb from TNF β to HLA-B. In addition, PFGE has been used to show that the gap between the cosmid clusters cannot be greater than 22-23 kb. From the genomic clones, this would predict a separation of 334 kb between the C2 and TNF α genes. The overestimate of 55 kb with an intervening Not I fragment (210 kb observed against 155 kb cloned) accounts for the discrepancy between the figures.

Failure to obtain a linking clone may reflect an unusual sequence within the uncloned region of DNA. Three different libraries, with genomic inserts prepared from separate restriction enzymes, were hybridised with probes from both sides of the gap in an attempt to complete the molecular map. Only one positive was detected, and this was not recovered on the secondary screen. Future cloning strategies which may allow the isolation of this region could involve the use of an *in vitro* packaging system derived from a host strain deficient in the K12 restriction endonuclease system (Rosenberg, 1985 and 1987; Rosenberg *et al.*, 1985).

The joint maps of the cosmid clusters are shown in Fig. 3.15. The large expanse of DNA between the known loci could encode a number of novel gene products. Precedents include the genes for B144 and RD (Tsuge *et al.*, 1987; Levi-Strauss *et al.*, 1988) which have recently been mapped to the murine MHC, and to the human and murine MHCs, respectively. Analysis of the cloned DNA for the presence of new genes could help to elucidate the molecular basis of HLA associated diseases, where the true genetic locus could be in linkage disequilibrium with the known MHC markers. Characterisation of the cosmids and genomic DNA for the presence of coding sequences by methods such as HTF island distribution and the conservation of DNA sequence between animal species is described in the following chapters.

CHAPTER IV

MAPPING NOVEL TRANSCRIPTS WITHIN THE CLASS III REGION.

4.1 INTRODUCTION

Until recently, the only genes which were mapped to the class III region of the MHC were those for the complement components C2, factor B and C4, and the cytochrome P-450 steroid 21-hydroxylase (21OH). Subsequently, loci for tumour necrosis factor (TNF) α and β were found to lie within the MHC of man and mouse (Spies et al., 1986; Muller et al., 1987a; Gardner et al., 1987) and were positioned between the complement gene cluster and the class I region (Dunham et al., 1987; Carroll et al., 1987; Inoko and Trowsdale, 1987; Ragoussis et al, 1988; Muller et al., 1987b).

The physical distances between the class III and the flanking class I and class II genes have been established from PFGE as 350 kb from DRA to 21OHB and 650 kb from C2 to HLA-B (Dunham et al., 1987). These distances are large enough to harbour a number of loci. This may be of major importance in understanding the relationship between the HLA region and human disease. Although the direct involvement of alleles of the class I, II or III has been implicated in some cases, for many, such as narcolepsy, coeliac disease and the systemic rheumatic diseases, the precise link has remained elusive.

Techniques which allow the detection of potential coding sequences include looking for the evolutionary conservation of single-copy sequences, and mapping of novel transcripts within cloned DNA. These methods have been used to describe two new products of the MHC; the RD gene, so called because the predicted protein has an unusual repeating unit (Levi-Strauss et al., 1988), and a B cell and macrophage specific

transcript, B144 (Tsuge et al., 1987). RD has been positioned adjacent to the factor B gene in both man and mouse, whereas B144 has been located 10 kb centromeric to TNF α in mouse only.

An alternative method, which combines molecular mapping of cloned DNA and PFGE mapping of chromosomal DNA has been used to characterise the cosmid clusters isolated from the class III region by chromosome walking. This approach involves analysing DNA for the presence of CpG rich sequences called HTF islands (Brown and Bird, 1986; Lindsay and Bird, 1987).

The major fraction of vertebrate DNA has an unusual distribution of C+G, with the frequency of the CpG dinucleotide reduced to 25% of that predicted from the base composition (Josse et al., 1961). Analysis of the mouse genome with the restriction enzyme Hpa II (recognition sequence 5'-CCGG-3') has revealed a discrete DNA fraction which represents 1% of the total genome, but contains 15% of all CpGs. These Hpa II tiny fragment (HTF) islands have several other unique characteristics. The total C+G content is often greater than 50%, CpG and GpC occur at similar frequencies, and about 80% of the sequences tested by Southern blotting or cot curve analysis are present at a level of one or very few copies per haploid genome (Bird et al., 1985). Furthermore, the cytosine bases are unmethylated at the 5 position of the pyrimidine ring, unlike the majority (60-90%) of CpGs (Cooper et al., 1983; Bird et al., 1985). This appears to be true irrespective of the tissue of origin of the chromosomal DNA (Bird et al., 1985; Kolsto et al., 1986), and may be related to the association of many HTF islands with the transcriptional start sites of housekeeping genes (Bird, 1986; Gardiner-Garden and Frommer, 1987). The unmethylated status could be important for the binding of transcription factors (Bird, 1986), as methylation of islands on the inactive X chromosome has been shown to

inhibit expression of the adjacent gene (Toniolo et al., 1984; Wolf et al., 1984a and 1984b). In addition, methylation correlates with a high frequency of mutation, owing to the deamination of 5-methylcytosine to thymidine. If the mispaired T-G is not recognised by DNA repair mechanisms, the original C-G is replaced by T-A (Coulondre et al., 1978; Bird et al., 1987; Cooper and Youssoufian, 1988). For example, germline methylation of an ancestral gene at the α-globin locus in man has been proposed to explain the presence of the homologous pseudogene ψα1 (Bird et al., 1987).

The characteristic features of HTF islands, high C+G content and undermethylation, allow their identification with restriction enzymes that have methylation sensitive recognition sites and at least one CpG per site. Since sites for these endonucleases are infrequent in bulk genomic DNA, they should be preferentially located in islands, and most islands should contain sites for one or more of these enzymes. Analysis of the distribution of these sites in cloned and chromosomal DNA suggests that Bss HII, Eag I and Sac II are the enzymes preferentially located in island structures (Table 4.1) (Lindsay and Bird, 1987; Brown and Bird, 1986). These three endonucleases were chosen for characterising the class III region of the MHC for the presence of putative HTF islands. By combining cloning with PFGE the precise endpoints of the restriction fragments observed in digests of chromosomal DNA were positioned on the molecular map. Flanking probes were hybridised to Southern blots of genomic DNA from a range of animal species, to look for conservation of nucleotide sequence. This was used as an indication of the presence of coding regions. Potential islands were then analysed for associated transcripts by choosing probes for hybridisation to Northern blots. Using these criteria, 13 new genes could be assigned to the region between TNF α and C4A, in addition to

Table 4.1

Frequencies of rare cutting enzymes associated with HTF islands.

Enzyme	*Sequence*	*CpG/ site*	*% sites/ island*	*no. sites/ island*
Not I	*GCGGCCGC*	*2*	*89*	*0.12*
Eag I	*CGGCCG*	*2*	*74*	*1.2*
Sac II	*CCGCGG*	*2*	*74*	*1.2*
Bss HII	*GCGCGC*	*2*	*74*	*1.2*
Sma I	*CCCGGG*	*1*	*42*	*1.2*
Nae I	*GCCGGC*	*1*	*42*	*1.2*
Nar I	*GGCGCC*	*1*	*42*	*1.2*
Mlu I	*ACGCGT*	*2*	*27*	*0.34*
Pvu I	*CGATCG*	*2*	*27*	*0.34*
Hpa II	*CCGG*	*1*	*21*	*11.1*
Hha I	*GCGC*	*1*	*21*	*11.1*

From Lindsay and Bird (1987)

confirming the location of the human RD gene, and a possible candidate for the homologue of the murine B144 transcript.

4.2 RESULTS

4.2.1 Mapping Rare Sites in Cloned DNA

The sites for the enzymes BssH II, Eag I and Sac II were mapped in the cosmid inserts from the cloned portion of the class III region. Each enzyme was used singly and in double digest combinations with Bam HI and Bgl II. Since the positions of the Bam HI and Bgl II sites were already established, they provided reference points from which to place the rare enzyme sites. Sites for Mlu I, Not I, Nru I and Pvu I had already been mapped by the same method (3.2.4).

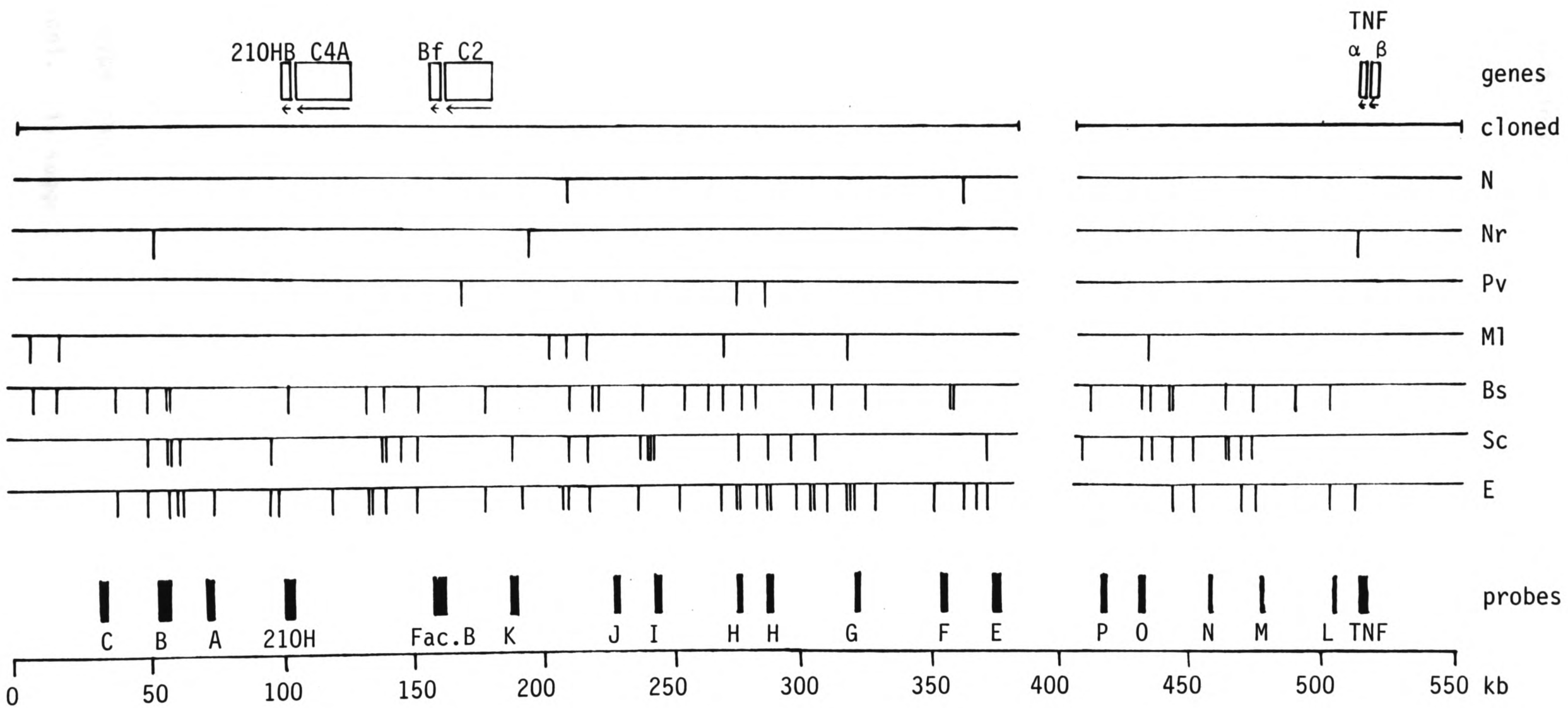
A surprisingly large number of sites were found within the cloned DNA; at least 33 for BssH II, 43 for Eag I and 31 for Sac II. Many of these occurred in clusters of two or more sites within less than 1 kb (Fig 4.1).

4.2.2 Mapping Rare Enzyme Sites in Genomic DNA

To establish which, if any, of the rare sites were cleaved at the level of chromosomal DNA, probes from across the cloned region were hybridised with Southern blots from PFGE . The blots consisted of DNA from the cell line used to prepare the cosmid library digested with BssH II, Eag I and Sac II. Successive probings of the same blots allowed the endpoints of the observed fragments to be positioned on the molecular map.

Fig. 4.1

Positions of restriction sites for infrequently cutting enzymes within the cloned portion of the class III region. Sites for the enzymes Not I (N), Nru I (Nr), Pvu I (Pv), Mlu I (Ml), Bss HII (Bs), Sac II (Sc) and Eag I (E) were mapped within the cosmid clones. The positions of the sites are indicated by the vertical bars.


21OHB C4A
Bf C2
TNF
α β
genes
cloned
N
Nr
Pv
Ml
Bs
Sc
E
probes
C B A 21OH Fac.B K J I H H G F E P O N M L TNF
0 50 100 150 200 250 300 350 400 450 500 550 kb

The region containing the complement/ 21OH gene cluster was found to have two potential islands between the factor B and C4A genes. Neither the complement nor 21OHB loci were associated with an island (Fig. 4.2). This is consistent with the tissue specific expression of these genes, as few tissue specific loci have HTF islands.

Between C2 and TNF α 17 single sites or clusters of sites were mapped in the chromosomal DNA (Fig. 4.3). The region 110 to 150 kb telomeric to C2 contained a large number of Eag I sites. The presence of small fragments between these points can only be inferred by the location of the ends of the flanking Eag I restriction elements, as no intervening probe is available.

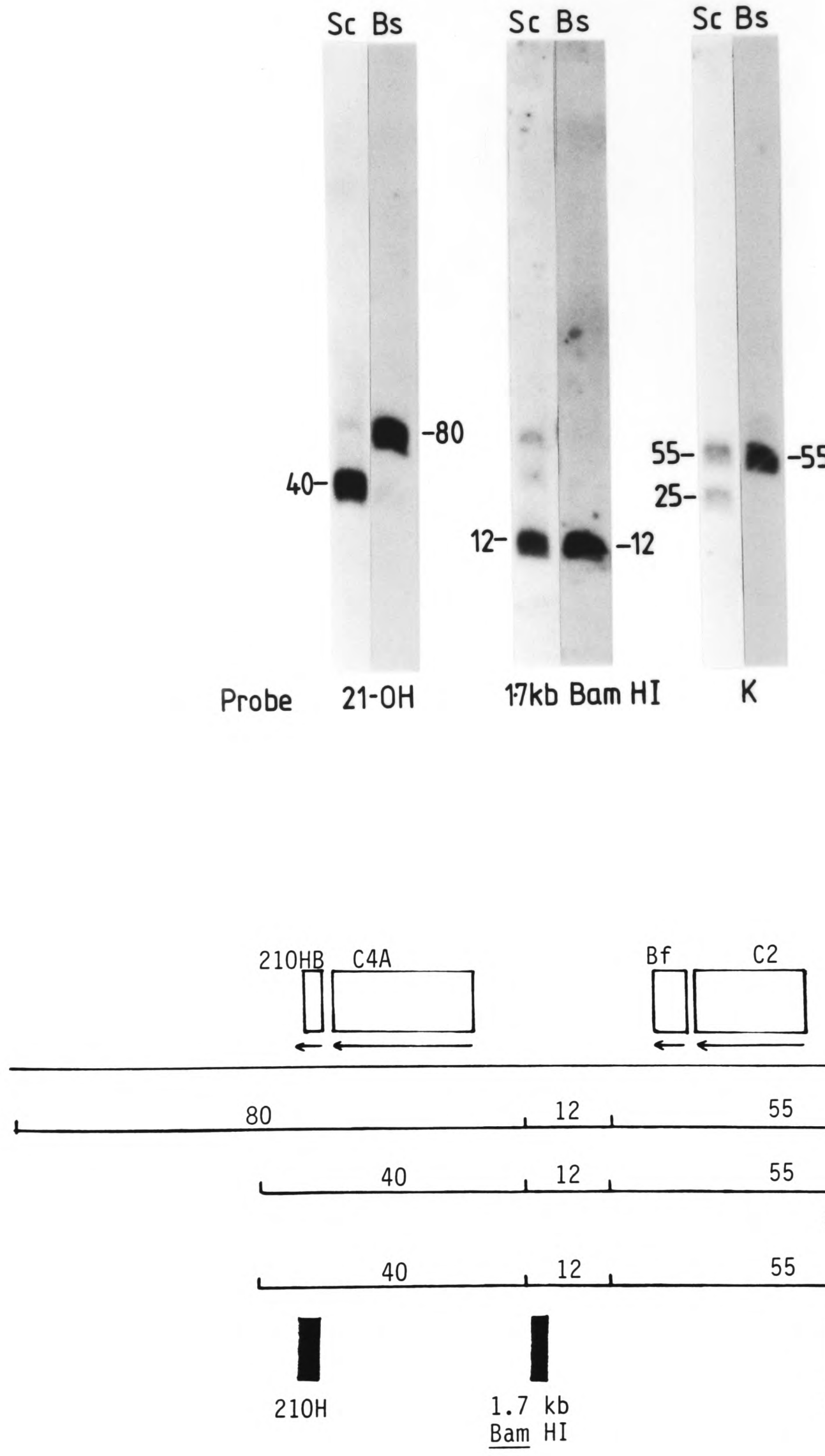
A fine-detail pulsed-field map of the cloned portion of the class III region was constructed from PFGE data (Ian Dunham, personal communication) (Fig. 4.3). A comparison of the fragment sizes from PFGE with the cloned DNA showed a high degree of concordance for the enzymes used (Table 4.2). The estimated sizes for the long-range map were within ±1-2 kb of the cloned sizes, with one exception. The Sac II fragment which spanned the gap between the two cosmid clusters was calculated at 118 kb as against 114 kb from the genomic clones. The difference between the figures is still less than 4%, and may reflect an overestimate owing to the larger size of this fragment with respect to the majority of the digest products.

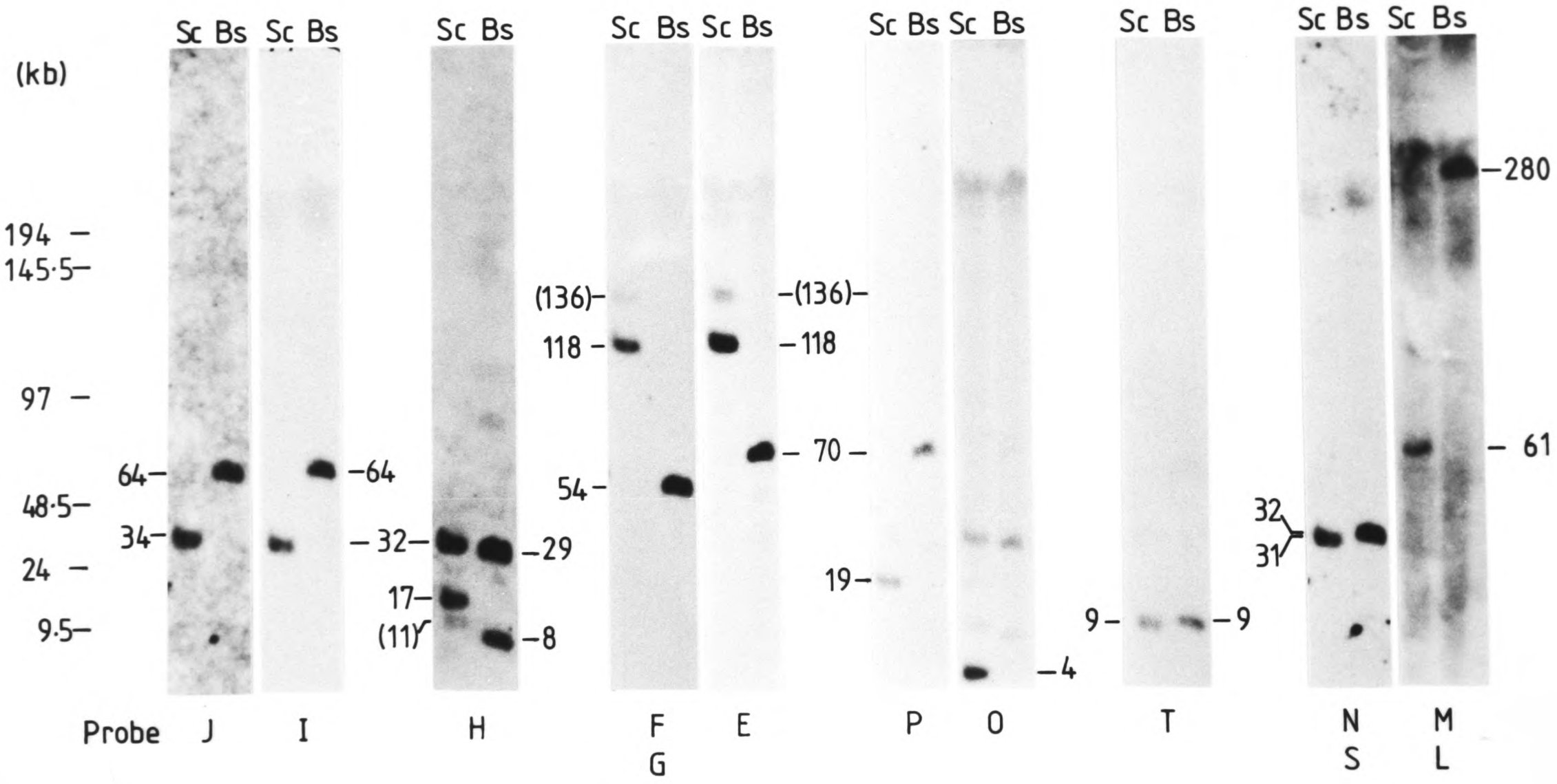
In total, 24 single sites or clusters of sites mapped in the cloned DNA were cleaved at the chromosomal DNA: 5 between the centromeric endpoint and 21OHB, 2 between C4A and Bf, and 17 between C2 and TNF α. No sites were found to lie in the 22 kb gap (Fig. 4.4).

The Bss HII fragments have been found to be identical in 7 other common haplotypes characterised by PFGE (Ian Dunham, personal communication). This suggests that the CpG rich clusters are

Fig. 4.2

Sites for infrequently cutting enzymes in the complement/ 21OH gene cluster. The upper panel shows the PFGE results for the enzymes Bss HII (Bs) and Sac II (Sc) using probes from the complement/ 21OH region. Results for Eag I (E) are not shown. From the data, the long-range map was constructed. Partial digest products are shown below the main digest products. All sizes are in kb.


Sc Bs
Sc Bs
Sc Bs
-80
40-
12-
-12
55-
25-
-55
Probe
21-OH
1·7kb Bam HI
K
210HB
C4A
Bf
C2
10 kb
80
12
55
Bs
40
12
55
Sc
25
40
12
55
E
probes
210H
1.7 kb
Bam HI
K


Sc Bs Sc Bs
(kb)
194 –
145·5–
97 –
64–
48·5–
34–
24 –
9·5–
Sc Bs
–64
– 32–
–29
17–
(11)
–8
Sc Bs Sc Bs
(136)–
118 –
–(136)–
– 118
54 –
Sc Bs Sc Bs
– 70 –
19–
–4
Sc Bs
9 –
– 9
Sc Bs Sc Bs
–280
– 61
32
31
Probe J I H F G E P O T N S M L

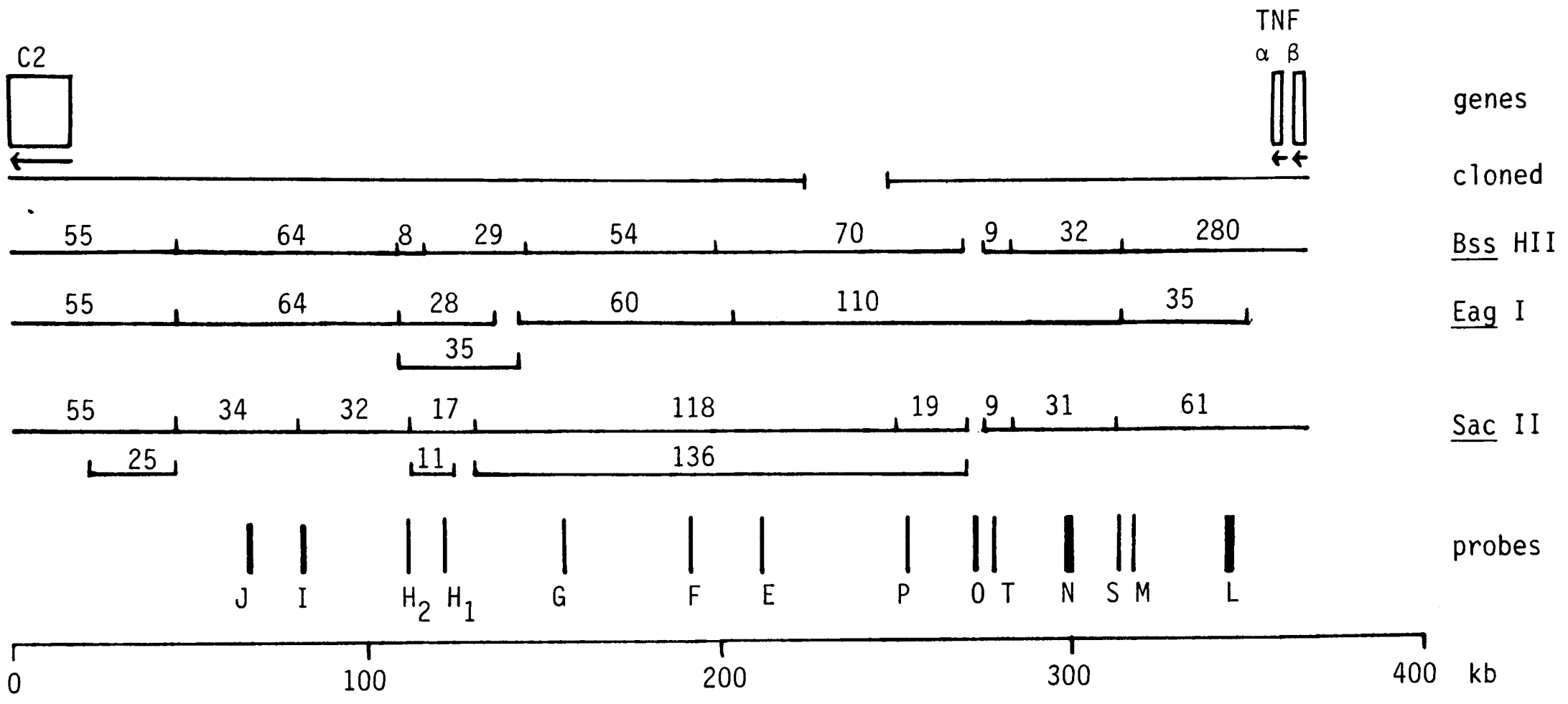


Fig. 4.3

Sites for infrequently cutting enzymes between C2 and the TNF genes. The panel of photographs shows the PFGE results for the enzymes Bss HII (Bs) and Sac II (Sc) using probes from across the class III region. Results with Eag I (E) are not shown. From the data, a long-range restriction map was constructed (lower panel). Partial digest products are shown below the main restriction digest products. All sizes are in kb.

Table 4.2

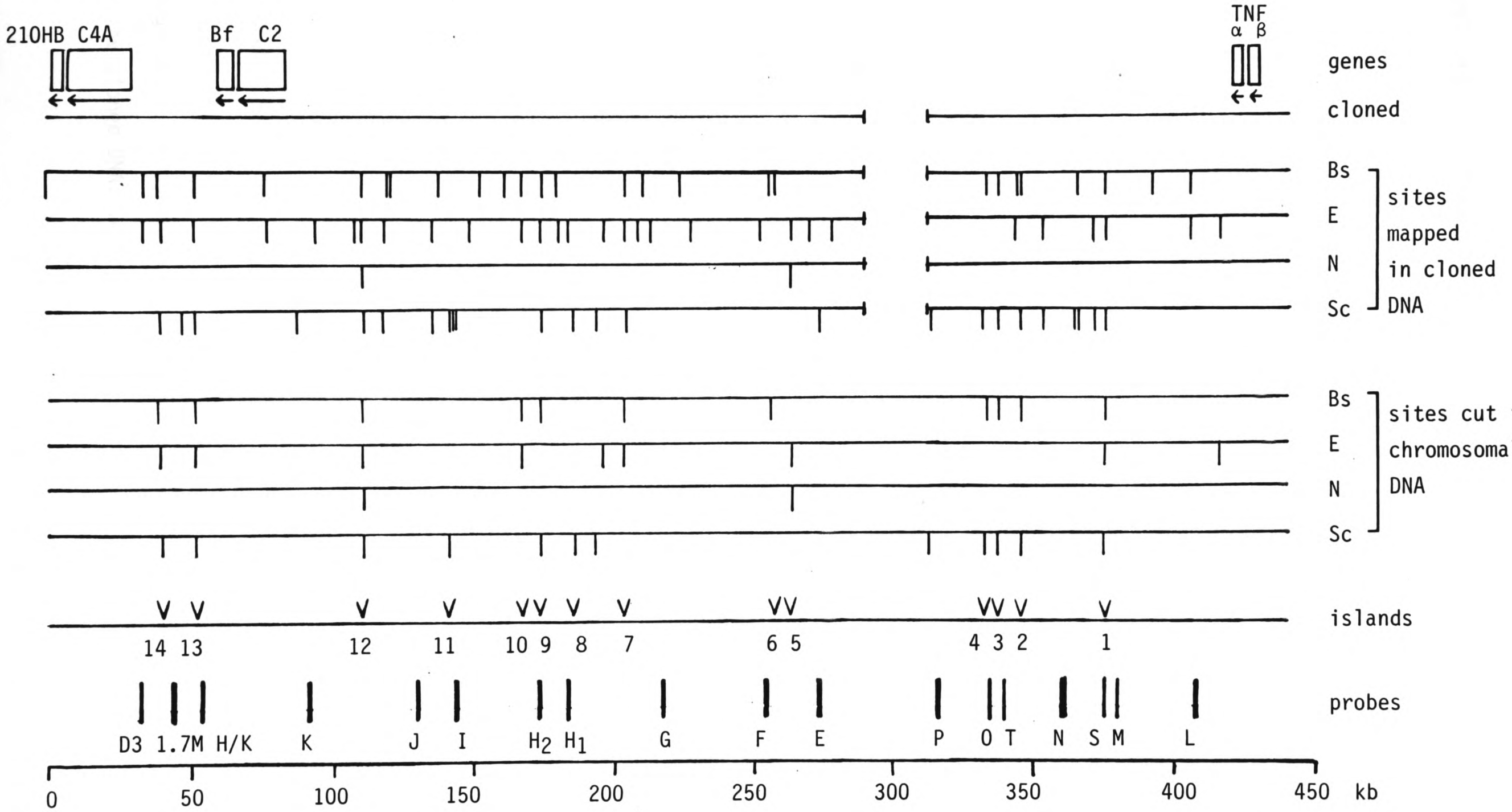
Comparison of the sizes of fragments as estimated from PFGE and cosmid cloning.

Probe	Bss HII		Sac II		Eag I	
	PFGE	cloned	PFGE	cloned	PFGE	cloned
J	64	62	34	35	64	62
I	64	62	32	32	64	62
H	29	29	32	32	28	29
	8	7	17	19	(35)	(35)
			(11)	(11)		
G						
F	54	53	118	114	60	59
			(136)	(133)		
E	70	70	118	114	110	108
			(136)	(133)		
P	70	70	19	19	110	108
			(136)	(133)		
O	NA	3.5	NA	4.5	110	108
T	9	8	9	8	110	108
N	32	31	31	31	110	108
S						
M	280	ND	61	ND	35	35
L						

All figures are in kb. Numbers in brackets indicate the sizes of partial digest products observed with some probes. Where different probes hybridise to the same fragments, they are positioned together in the table.
NA: the fragment size was not determined by PFGE.
ND: the fragment size was not estimated from the cloned DNA.

Fig. 4.4

Summary of the sites mapped at the cloned and chromosomal levels between 21OHB and TNF α. The positions of the sites for Bss HII (Bs), Eag I (E), Not I (N) and Sac II (Sc) are shown by the vertical bars. Islands 1 to 14 are indicated by the vertical arrow heads, and the probes used to define the chromosomal fragments are shown as solid boxes.


21OHB C4A
Bf C2
TNF
α β
genes
cloned
Bs
E
N
Sc
sites mapped in cloned DNA
sites cut in chromosomal DNA
islands
14 13 12 11 10 9 8 7 6 5 4 3 2 1
probes
D3 1.7M H/K K J I H_2 H_1 G F E P O T N S M L
0 50 100 150 200 250 300 350 400 450 kb

representative of true HTF island structures, since they are conserved in other cell lines.

4.2.3 Characterisation of Islands for Novel Transcripts

To characterise potential HTF islands for novel transcripts, adjacent unique DNA sequences were analysed by hybridisation to animal blots and Northern blots as described below. All of the probes, with the exception of probe H, gave hybridisation patterns consistent with single copy elements. Once the positions of the islands and conserved sequence probes were established, Northern blot analysis was used to confirm whether mRNA species could be assigned to the putative loci.

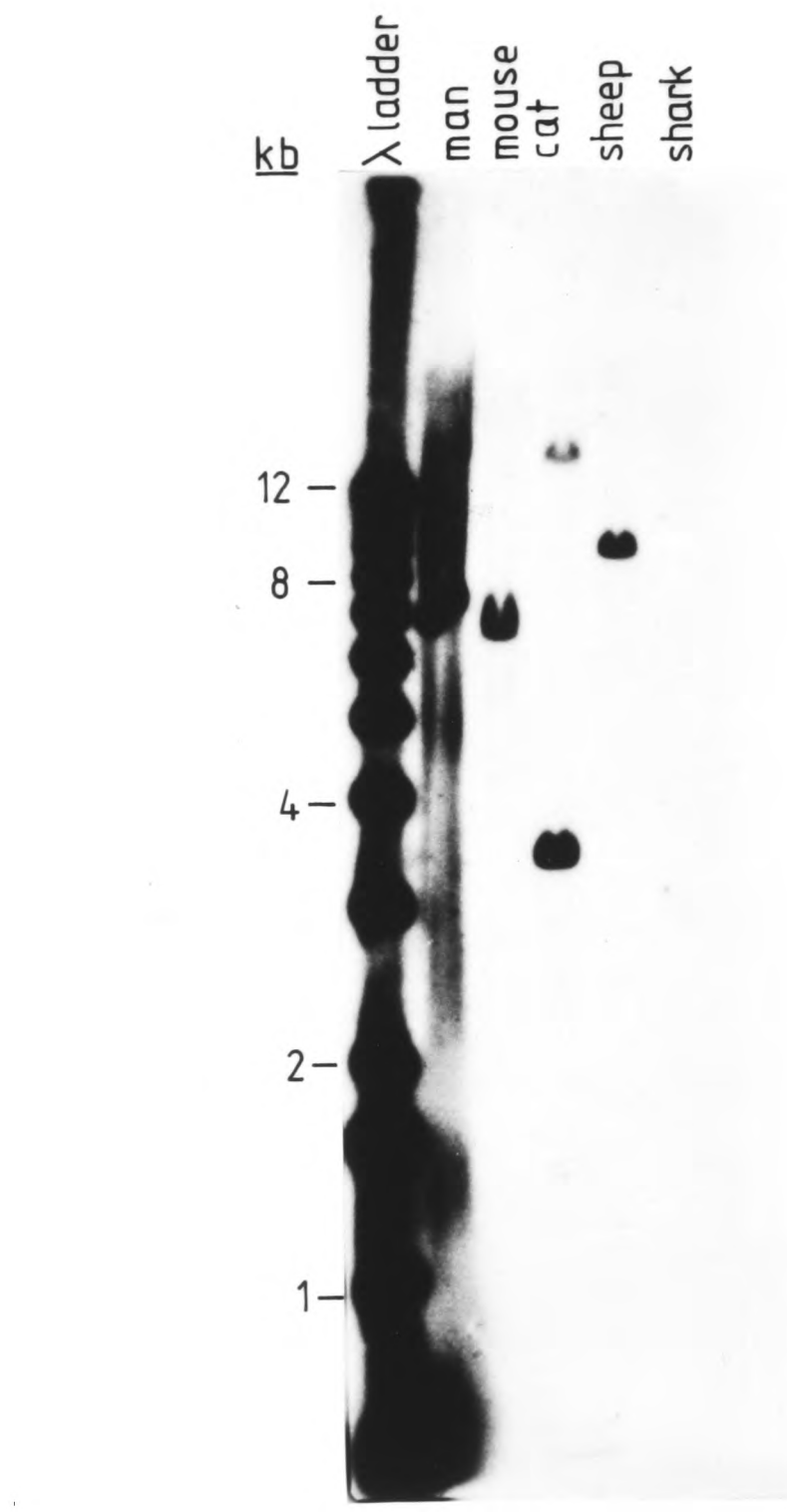
(a) conservation of probes

Unique sequence genomic probes were tested on Southern blots of Bam HI digested DNA from man, mouse, cat, sheep and shark. Coding sequences ought to be more highly conserved than non-coding, and cross-hybridise between related species (Monaco et al., 1986). Most of the walking probes and other genomic probes associated with HTF-islands were found to detect restriction fragments in the mammalian species under high stringency conditions (washing at $65^{o}C$, 0.1x SSC). Those unique sequences which did not failed to detect transcripts on subsequent Northern blot analysis, for example probe E, which is 8.3 kb from the Not I site characterised by probe X_1 (Fig. 4.4).

An example of an animal blot is shown in Fig. 4.5. The 1.6 kb Kpn I fragment, probe I, is situated 66 kb telomeric to the C2 gene. It is located 0.6 kb from a cluster of 3 Sac II sites, at least one of which is known to cleave in chromosomal DNA. Probe I hybridises to a 6.8 kb fragment in the human track, which is equivalent to the 6.8 kb fragment in the cloned DNA. In addition, it detects fragments at 6.5, 3.5 and 18,

Fig. 4.5

Southern blot of DNA from man, mouse, cat, sheep, and shark, hybridised with probe I. Fragments are seen in all the mammalian species, but not in shark.


kb
λ ladder
man
mouse
cat
sheep
shark
12
8
4
2
1

and 9.5 kb, in the mouse, cat and sheep, respectively, but not in the more distantly related shark. The Northern blot analysis of probe I (see below) shows that it is associated with a 1.9 kb RNA transcript, G9.

(b) Northern blot analysis

HTF islands between 21OH and TNF α were studied in more detail to establish whether they were associated with novel transcripts. Unique sequences from within 0-15 kb of the clusters of enzyme sites were hybridised to Northern blots. Total RNA species were obtained from U937, PMA stimulated U937, HepG2, Molt 4, and either Daudi or Raji cell lines; these represented monocyte, macrophage, liver, T lymphocyte and B lymphocyte lineages, respectively. All but one of the islands and two single sites were investigated; these could not be characterised as no suitable probes were available. Where possible, both 3' and 5' probes were taken for island analysis, and at least one of the two was found to detect a transcript (table 4.3).

The probes used in the analyses were shown to be unique by hybridisation to Southern blots of Bam HI and Bgl II digests of human genomic DNA, with the exception of probe X_1 and the 0.7 kb Bgl II fragment associated with G6 and G8, respectively. These were found to contain some repetitive elements which produced an intense background smear, above which restriction fragments of the correct size were just visible (Fig. 4.6A). The repetitive DNA was competed out of the radiolabelled probe prior to hybridisation by reannealing with total genomic DNA (2.7.3). This proved to be an effective method to reduce the background, and allowed detection of specific bands in both Southern and Northern blot analysis (Fig. 4.6B, and Fig. 4.8).

The messenger species ranged from 0.6 to 6 kb in size, and represent the products of 11 putative loci. Most of the transcripts were expressed in all of the cell lines, as predicted from previous studies of HTF

Table 4.3

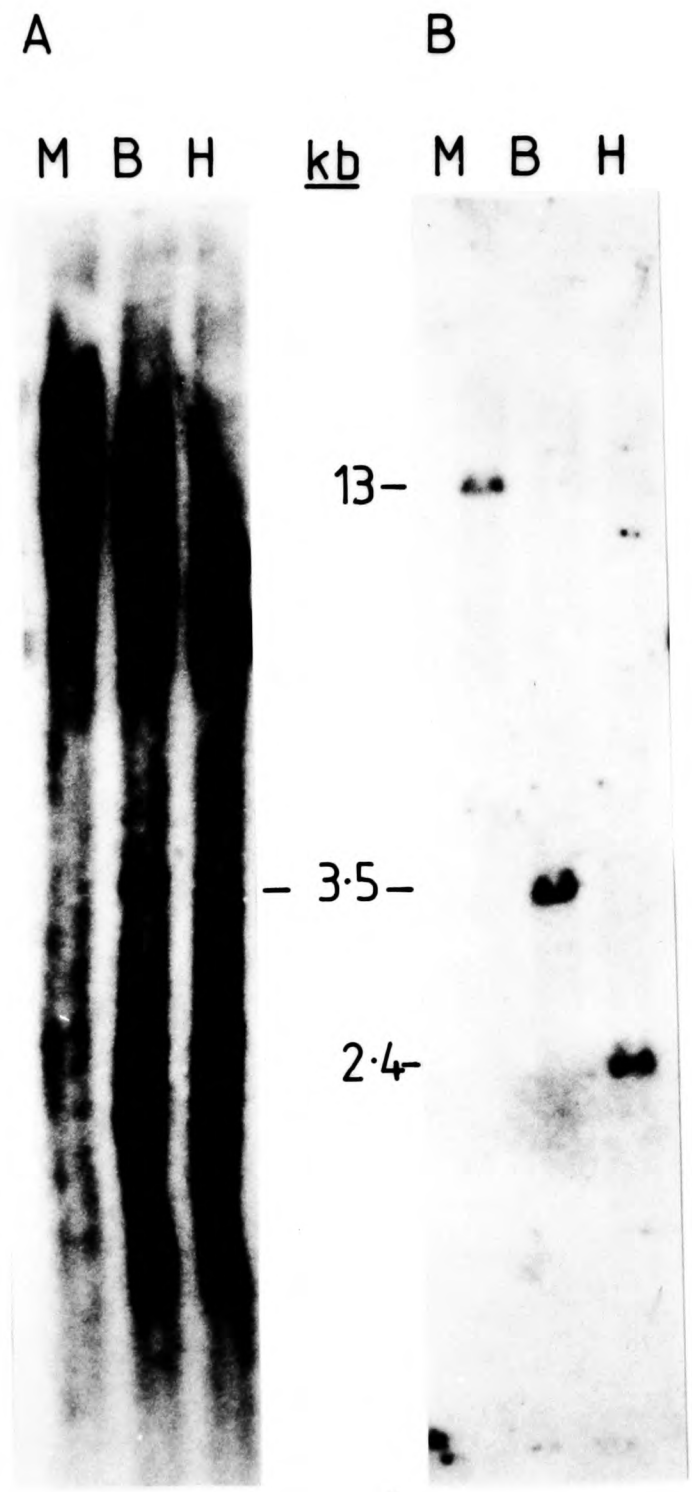
Probes used to analyse HTF islands for the presence of associated transcripts.

Island	Probe	Distance (kb) from island	Transcript ± size (kb)		Designation
1	M	3.2	+	0.6	G1
	S	0.1	+	6.0	G2
	N	15.0	+	6.0	G2
2	V	9.5	+	3.8	G3
	T	6.0	+	2.5	G4
3	T	0.4	+	2.5	G4
	O	3.5	+	1.7	G5
4	O	0.3	+	1.7	G5
	P	13.7	-		
5	X_1	contains I	+	1.4/1.5	G6
6	F	3.0	+	3.0	G7
7	N.D.				
8, 9	H	0.3	N.D.		HSP
10	W	2.5	+	1.0	G8
11	I	0.6	+	1.9	G9
12	M/X	0.3	+	2.6	G10
13	H/K	0.3	+	1.5	RD
14	1.7 BamHI	2.1	-		
	Sal I	contains I	+	1.4	G11
	D3	6.0	+	1.4	G11

N.D.= not determined in this study

Fig. 4.6

Removal of repetitive sequences by reassociation with total genomic DNA. Part A shows probe X_1 hybridised to a genomic Southern blot of human DNA digested with Bam HI (M), Bgl II (B) and Hind III (H). Part B shows the same experiment following removal of repetitive sequences from the probe by reassociation with total genomic DNA (final concentration 10 µg/ ml of hybridisation solution).


A
B
M B H
kb
M B H
13–
3·5–
2·4–

island distribution. One exception was the message associated with G2. This 0.6 kb RNA was detected only in T cells, monocytes and macrophages. A summary of the transcripts and their tissue distribution is shown in table 4.4.

G1: the G1 transcript is 0.6 kb in size. It is expressed in U937, PMA+ U937, and Molt 4 cell lines, but not in HepG2 or Raji. The mRNA was detected with a 0.45 kb Bam HI/ Hind III fragment (probe M) 3.2 kb telomeric to island 1 (Fig. 4.7). As the orientation and size of the gene is unknown, G1 is not conclusively shown to be an example of a tissue specific HTF-island associated transcript.

G2: the G2 transcript is 6 kb in size, and has been detected in all five cell lines used in Northern blot analysis. The mRNA hybridises with a 1.3 kb Bam HI/ Bgl II fragment (probe N), 15 kb centromeric to island 1 (Fig. 4.7). In addition, a 1.1 kb Kpn I fragment (probe S) adjacent to the rare enzyme sites hybridises weakly to a mRNA of the same size, and probably represents the 5' end of the gene.

G3: the G3 transcript is 3.8 kb in size, and is expressed in all five cell lines used in Northern blot analysis. The reduced intensisty of the message in the Daudi cell line may reflect a real difference in the level of expression, as there is not a significant decrease in the intensity of the 28S band with respect to the Molt 4 lane. Reprobing of the same blot also suggests that the amount of RNA per track is similar between these cell lines. The mRNA is detected with a 2.4 kb Bam HI fragment (probe V) 9.5 kb telomeric to island 2 (Fig. 4.7).

G4: the G4 transcript is 2.5 kb in size, and is expressed in all the cell lines used in Northern blot analysis. The increased intensity in the PMA+ U937 lane suggests that expression of this mRNA may be enhanced by the stimulation of the tissue culture cells. The mRNA is detected by

Table 4.4

Distribution of transcripts mapped within the cloned portion of the class III region of the human MHC. The five cell lines were analysed by Northern blot analysis, and are scored for expression (+) of the mRNA species.

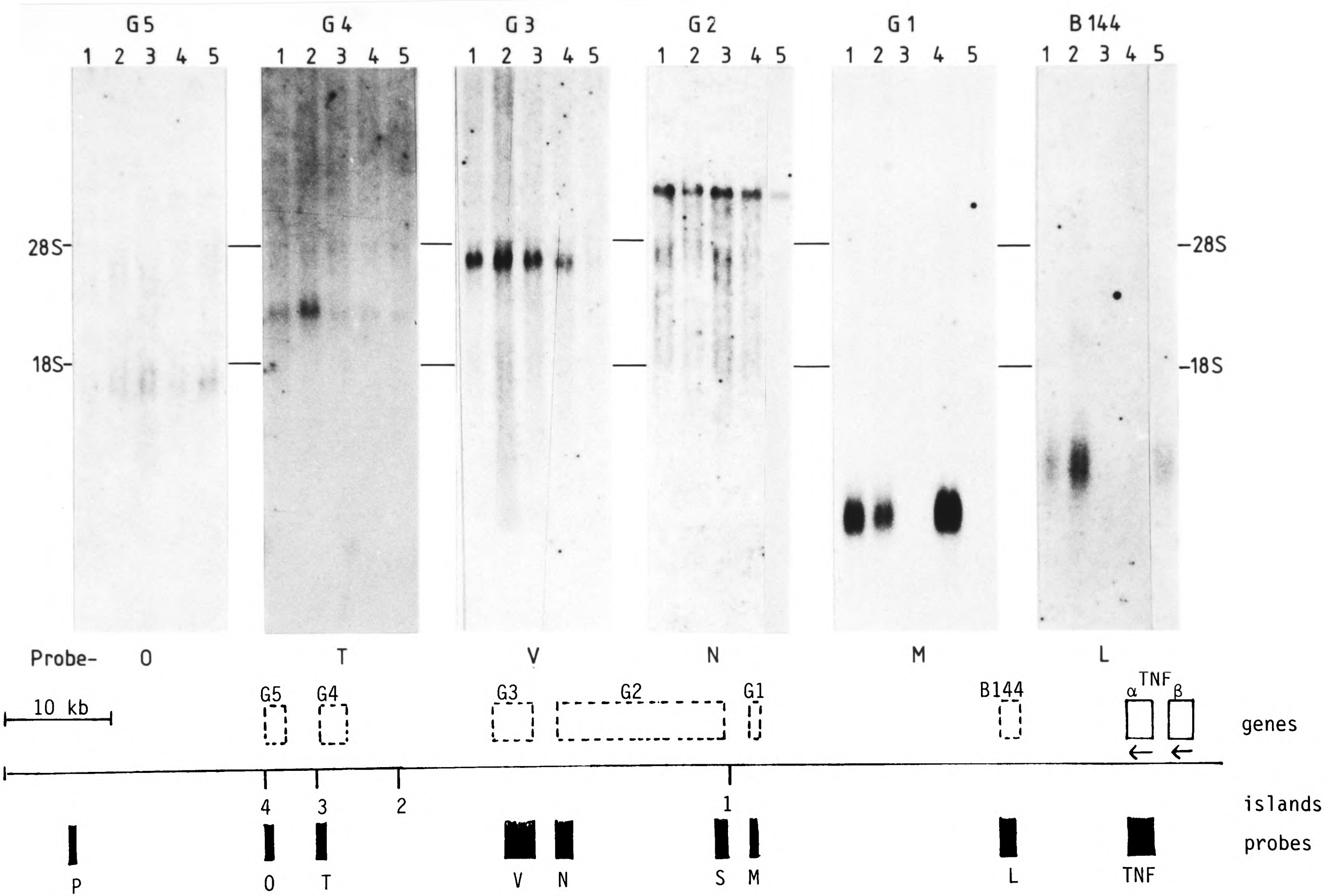
Gene	Transcript (kb)	Mo	Ma	T	B	H
TNF	1.1	-	-	+	+	-
TNF	1.1	-	+	-	-	-
B144	0.8	+	+	+	-	-
G1	0.6	+	+	-	+	-
G2	6.0	+	+	+	+	+
G3	3.8	+	+	+	+	+
G4	2.5	+	+	+	+	+
G5	1.7	+	+	+	+	+
G6	1.4, 1.5	+	+	+	+	+
G7	3.0	+	+	+	+	+
HSP 70-1	2.4	N.D				
HSP 70-2	N.D.	N.D.				
G8	1.0	+	+	+	+	+
G9	1.9	+	+	+	+	+
G10	2.6	+	+	+	+	+
C2	2.9	-	-	-	-	+
Factor B	2.6	-	-	-	-	+
RD	1.5	+	+	+	+	+
G11	1.4	+	+	+	+	+
C4A	5.5	-	-	-	-	+
21OHB	2.2			*		

Mo= monocyte
Ma= macrophage
B= B lymphocyte
T= T lymphocyte
H= HepG2

N.D.= the expression of this transcript has not been determined in this study.
* = expressed in adrenal gland.

Fig. 4.7

Northern blot analysis of island associated probes from the TNF cosmid cluster. The photograph shows a panel of blots hybridised with a series of probes from across the cloned region. The RNA samples are derived from U937 (1), PMA stimulated U937 (2), HepG2 (3), Molt 4 (4) and Daudi or Raji (5) cell lines. The approximate positions of the genes, based upon transcript sizes and identifying probes, are shown below.


G5
G4
G3
G2
G1
B144
1 2 3 4 5
28S
18S
28S
18S
Probe- O T V N M L
10 kb
G5
G4
G3
G2
G1
B144
TNF
α
β
genes
4
3
2
1
islands
probes
P
O
T
V
N
S
M
L
TNF

a 0.8 kb Bgl II fragment (probe T) 0.4 kb telomeric to island 3 (Fig. 4.7).

G5: the G5 transcript is 1.7 kb in size, and is expressed in all the cell lines used in Northern blot analysis. The mRNA is detected by a 0.6 kb Bgl II fragment (probe O) 0.3 kb telomeric to island 4 (Fig. 4.7).

G6: the G6 transcripts are 1.4 and 1.5 kb in size in U937, PMA+ U937, HepG2 and Molt 4, but 1.4 and 1.45 kb in Raji. The mRNA species are detected by a 2.4 kb Hind III fragment (probe X_1) which spans the Not I site, and contains at least part of the island structure (Fig. 4.8). The transcripts may be the products of alternative splicing, or of two loci, perhaps arranged 5' to 5', as in the murine surfeit locus (Williams et al., 1988). The slightly smaller mRNA in the Raji track could be a cell specific derivative of one of these putative products associated with the Not I island.

G7: the G7 transcript is 3.0 kb in size, and is expressed in all of the cell lines used in Northern blot analysis. The mRNA is detected with a 1.3 kb Bgl II fragment (probe F) located 3.0 kb centromeric to island 6 (Fig. 4.8).

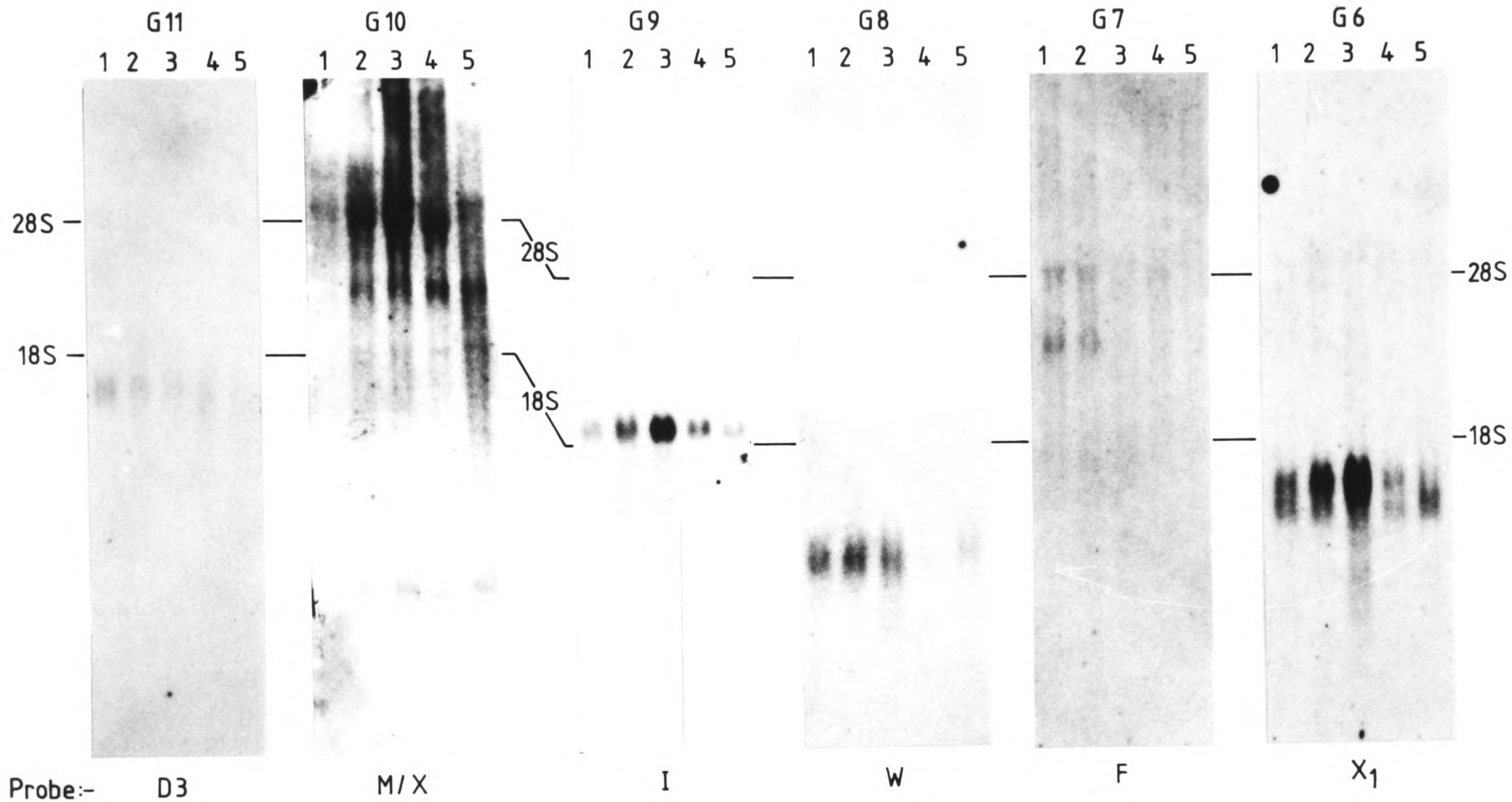
H_1 and H_2: No Northern blot analysis was carried out for islands 8 and 9, as the associated probe (H) was found to be derived from a member of the HSP 70 family. (See chapter VI.)

G8: the G8 transcript is 1.0 kb in size. It is expressed strongly in U937, PMA+ U937, HepG2 and Raji cell lines, but only weakly in Molt 4. This probably reflects a real difference in the level of expression, as probing of the same blot with another genomic fragment suggests that the quantity of RNA per track is similar. The mRNA is detected by a 0.7 kb Bgl II fragment (probe W) 2.5 kb centromeric to island 10 (Fig. 4.8).

G9: the G9 transcript is 1.9 kb in size, and is expressed in all of the cell lines used in Northern blot analysis. It is detected by a 1.6 kb

Fig. 4.8

Northern blot analysis of island associated probes from the complement/ 21OH cosmid cluster. The photograph shows a panel of blots hybridised with a series of probes from across the cloned region. The RNA samples are derived from U937 (1), PMA stimulated U937 (2), HepG2 (3), Molt 4 (4) and Daudi or Raji (5) cell lines. The approximate positions of the genes, based upon transcript sizes and identifying probes, are shown below.


G11
G10
G9
G8
G7
G6
1 2 3 4 5
28S
18S
Probe:- D3
M/X
I
W
F
X1
210HB
C4A
G11
RD
Bf
C2
G10
G9
G8
HSP 70
2
1
G7
G6
10 kb
genes
islands
probes
14 13
12
11
10 9 8
7
6 5
D3
SalI
H/K
K
M/X
J
I
W
H2
H1
G
F
X1
E

Kpn I fragment (probe I) located 0.6 kb telomeric to island 11 (Fig. 4.8).

G10: the G10 transcript is 2.6 kb in size, and is expressed in all of the cell lines used in Northern blot analysis. It is detected by a 1.3 kb Bam HI/ Xho I fragment (probe M/X) which is just 3' of island 12 (Fig. 4.8).

RD: the RD transcript is 1.5 kb in size, and is expressed in all of the cell lines used in Northern blot analysis. It is detected by a 1.2 kb Hind II/ Kpn I fragment (probe H/K) located 0.3 kb telomeric to island 13 (Fig. 4.8)

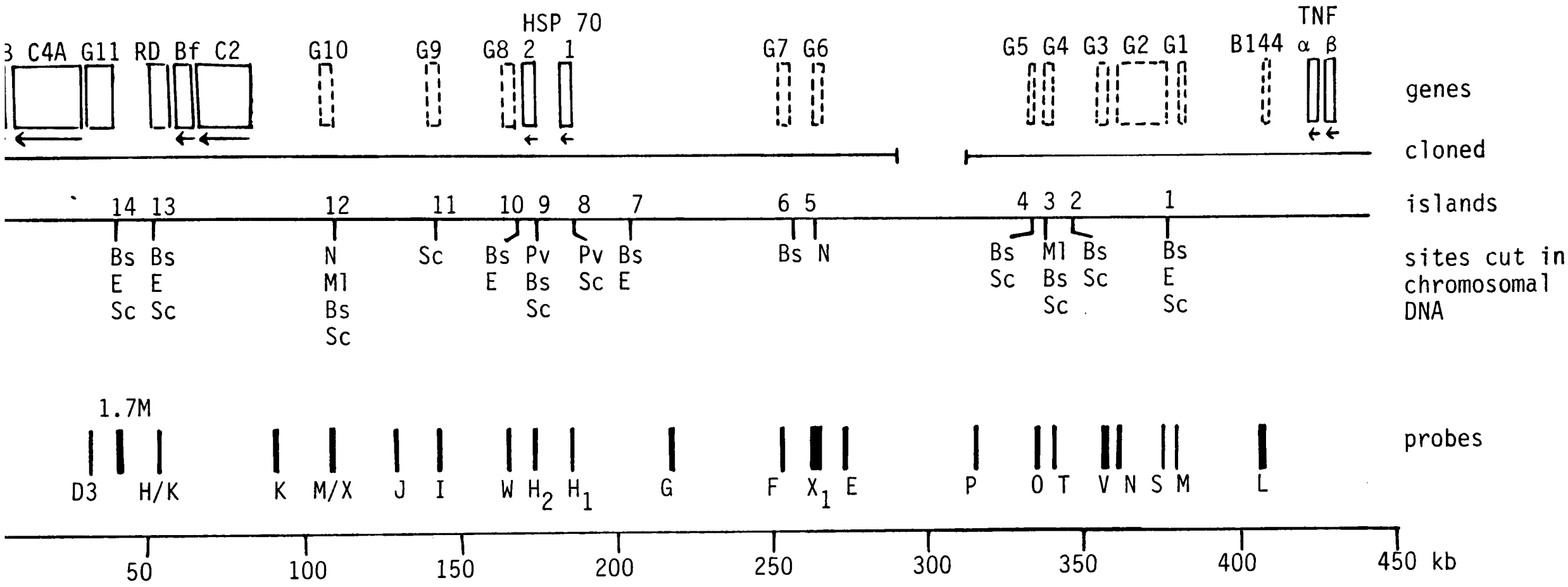
G11: the G11 transcript is 1.4 kb in size, and is expressed in all of the cell lines used in Northern blot analysis. It is detected by a 0.5 kb Nhe I/ Bst EII fragment (probe D3) located 6.0 kb centromeric to island 14 (Fig. 4.8), and is one of two novel transcripts to be mapped to the 30 kb between the factor B and C4 loci (chapter V).

A summary of the transcripts positioned between 21OH and TNF α is given in Fig. 4.9.

4.2.4 Possible Identification of the Human Equivalent of the Mouse B144 Transcript

The B144 gene has been positioned approximately 10 kb centromeric to the TNF α locus in mouse (Tsuge et al., 1987). The mRNA is an 800 bp transcript expressed only in B cells and macrophages; no protein product has been identified. In man, a region of unique DNA sequence is located 7 to 11 kb centromeric to TNF α. If man and mouse share a similar arrangement of loci, this may be the human analogue of the B144 gene.

A 1.4 kb Bam HI fragment, probe L, was chosen for hybridisation with a Southern blot of Bam HI digested chromosomal DNA from man, mouse, cat,



. 4.9

summary of the genes mapped between 21OHB and the TNF loci. Open boxes show the precise positions of the known genes, the directions of transcription are indicated by the horizontal arrows. The broken boxes indicate the probable positions novel genes. HTF islands, and the enzymes which characterise them are numbered 1 to 14 (Bss HII=Bs, Eag I=E, Sac II=Sc, I=N, and Mlu I=Ml). Probes used in Northern blot analysis of the class III region are shown as solid boxes.

sheep and shark. Following washing at high stringency (65^{o}C, 0.1x SSC) and autoradiography, the DNA sequence of L was found to be sufficiently conserved to allow cross-hybridisation to the other mammalian species, but not to shark (not shown). Conservation of nucleic acid sequence is believed to be diagnostic of coding regions of the genome (Monaco et al., 1986). Therefore, probe L was used in Northern blot analysis with RNA isolated from U937, PMA stimulated U937, HepG2, Molt 4, and Raji cell lines. The fragment detected transcripts in the U937 and PMA+ U937; it hybridised faintly with RNA from the Raji cell line. No message was seen in HepG2 or Molt 4 (Fig. 4.8). The RNA was estimated at 800 bp in length, which is consistent with B144. The reduced intensity of the message in the Raji track may reflect a cell line specific difference in the level of transcription. In the absence of confirmatory sequence data, the evidence suggests that L may be derived from the human equivalent of the B144 gene.

Although a single Eag I site which cleaves in chromosomal DNA has been mapped 3 kb 3' to TNF α, it is not clear how this is related to B144. Therefore, it is possible that B144 is an example of a tissue specific gene which is island associated.

4.3 DISCUSSION

The large number of sites for the rare enzymes BssH II, Eag I and Sac II appears to reflect an overall C+G rich character of the MHC. Comparison of the observed frequencies of sites with those estimated by Drmanac et al. (1986) shows that Not I, BssH II, Sac II and Eag I are upto 11 times more common than expected. In contrast, Pvu I, Nru I and Mlu I (all of which contain A and T in addition to C and G in their recognition sites) occur close to or below the expected frequency (Table

4.5). A study of the distribution of rare restriction sites in randomly isolated cosmids from chromosome 3 (Smith et al., 1987) has shown that high frequencies are observed in other regions of the genome. As these cosmids all hybridised to Alu repeats, this may be related to the distribution of genes within C+G rich isochores (see below).

Similarly, the percentage of these sites which have been shown to be island associated is much lower than that predicted by Lindsay and Bird (1987). Both Bss HII and Sac II occur below the expected 74%, at 45.5% and 61.3%, respectively, but the most dramatic difference is seen with Eag I, where only 27.9% of the sites in this study were island associated. Full details are given in tables 4.6a,b.

The region between the TNF α locus and 210HB encompasses approximately 420 kb. In total, 13 new transcripts have been defined on the basis of the presence of HTF islands in genomic DNA, and the conservation of DNA sequence between animal species. They range from 0.6 to 6 kb in size: a summary of the distribution of the putative loci is shown in Fig. 4.8, and the probes used to detect the transcripts are described in table 4.3. All but one of these island associated transcripts are expressed in every cell line investigated by Northern blot analysis. The exception, G2, appears to be present only in monocyte, macrophage and T cell lines.

Probe X_1 is associated with two transcripts on Northern blot analysis. As X_1 contains an island structure, this could reflect either alternatively spliced mRNA species from the same gene, or two loci in a 5' to 5' arrangement. Flanking probes did not clarify the situation, so only one putative locus (G6) has been included in the molecular map.

The nature of the duplicated probe, H, which allowed identification of the HSP 70 loci is described in chapter VI.

The approach described is biassed towards the identification of such ubiquitous messages, since the distribution of HTF islands reveals a

Table 4.5

Enyzme	Sequence	N	F(kb)	E(kb)	E/F
Not I	GC'GGCCGC	2	270.5	3 000	11.09
Nru I	TCG'CGA	3	180.3	100.5	0.56
Pvu I	CGAT'CG	3	180.3	126.5	0.69
Mlu I	A'CGCGT	8	67.6	132.0	1.95
Bss HII	G'CGCGC	33	16.4	189.0	11.52
Eag I	C'GGCCG	43	12.6	148.5	11.79
Sac II	CCGC'GG	31	17.5	149.0	8.51

N= no. sites which can be mapped in cloned DNA on the basis of restriction digests: frequencies may be higher, as small fragments (0.4 kb or less), or overlapping recognition sequences identified in sequenced DNA are not included.
F= 541/N, where 541 kb = total cloned DNA. F is the average distance between sites.
E= the expected average distance between sites (Drmanac et al., 1986)

Table 4.6a The complete region cloned

Enyzme	N^c	N^g	$\%N^c/N^g$	$\%I^a$	$\%I^e$	I^e/I^a
Not I	2	2	100	100	89	0.89
Nru I	3	1	33.3	0	-	-
Pvu I	3	2	66.6	66.6	27	0.41
Mlu I	8	3	37.5	25.0	27	1.08
Bss HII	33	15	45.5	45.5	74	1.65
Eag I	43	13	30.2	27.9	74	2.65
Sac II	31	19	61.3	61.3	74	1.21

Table 4.6b 21OH to TNF

Enyzme	N^c	N^g	$\%N^c/N^g$	$\%I^a$	$\%I^e$	I^e/I^a
Not I	2	2	100	100	89	0.89
Nru I	2	1	50	0	-	-
Pvu I	3	2	66.6	66.6	27	0.41
Mlu I	6	2	33.3	33.3	27	0.81
Bss HII	28	11	39.2	39.2	74	1.89
Eag I	36	8	22.2	19.4	74	3.81
Sac II	25	14	56.0	52.0	74	1.42

N^c= no. sites in cloned DNA
N^g= no. sites in genomic DNA
I^a= HTF island associated: the island is defined as a cluster of sites known to cut within genomic DNA. Single sites, unless shown to be transcript associated are not included.
I^e= expected frequency for island association: no figures are available for Nru I (after Lindsay and Bird, 1987).

predisposition towards association with housekeeping genes (Gardiner-Garden and Frommer, 1987; Bird, 1986). In addition, the ability to map islands and look for transcripts depends upon finding unique sequences within the cloned DNA. For example, one potential island (7) was not characterised by Northern blotting because suitable probes were not isolated from the cosmid inserts.

Studies of chromosome structure by staining with Geimsa or quinicrine have shown that G/Q bands correspond to regions which are A+T rich, whereas the Geimsa light, or R, bands correspond to regions which are C+G rich. Furthermore, the replication times of these bands are differentially controlled during the S phase of the cell cycle. The R bands are replicated early, whereas the Geimsa-dark bands are replicated late (Comings, 1978; Goldman et al.,1984). Analyses of the distribution of genes within early or late replicating bands, and within C+G or A+T rich isochores separated by caesium chloride density gradient centrifugation, have concluded that most housekeeping genes are encoded in early replicating (C+G rich) DNA (Goldman et al., 1984; Bernardi et al., 1985). The distribution of specific repeat elements within the genome seems to correlate with the observed banding patterns: Alu (56% C+G) follows R banding, and L1 (58% A+T) follows G/Q banding (Korenberg and Ryowski; 1988). This has led to the proposal that such repeat elements may have a normal role in DNA replication. The Alu family may also be involved in DNA exchange, creating hot spots for crossing-over between chromosomes, or allowing illegitimate recombinations that result in the deletion or insertion of genetic material. Detailed analysis of chromosomes carrying mutations in the α and β globin gene clusters has established that the majority of breakpoints occur within Alu repeats (Henthorn et al., 1986; Nicholls et al., 1987). As the p21 band of chromosome 6 is an R band these findings may have important consequences

in understanding the structure of the MHC. Firstly, there are implications with respect to the class II region gene duplications or deletions within extended haplotypes, and to the evolution of the MHC from a set of ancestral loci. Secondly, the prediction that most early replicating genes encode proteins which are universally expressed is consistent with:

(a) the high frequency of HTF islands within the class III region,

(b) the ubiquitous expression of the majority of HTF island associated transcripts within the cell lines studied,

(c) an essentially stable background to the class III region in which large scale chromosomal rearrangements would be deleterious.

The latter point appears to be confirmed by fine-detail PFGE mapping of the class III region in different extended haplotypes, where the only variation in the DNA content can be attributed to the number of C4 and 21OH loci present (Ian Dunham, personal communication).

Between 21OHB and TNF α there are a total of 20 loci; 11 of unknown function which are described here (G1-G11), RD, B144, two HSP 70 genes, C2, factor B, C4A, 21OHB and TNF α. This represents a gene density of 1 every 21 kb. For an estimated 20,000 to 50,000 genes per haploid genome, the expected average is 1/ 150kb to 1/ 60 kb. However, there may be significant variations in different regions of the genome.

No tissue specific gene previously described within the class III region (21OH, C4, Bf, C2) is found close to an island, although computer analysis suggests that class II loci may be preceded by C+G rich tracts (Tykocinski and Max, 1984). If other tissue specific genes map within the cloned region of the class III, then the observed gene density is an underestimate. To establish the true gene density, other techniques could be used in conjunction with island mapping. One approach, which was used to identify the putative human B144 gene, is to test unique

probes by hybridising to Southern blots of DNA from a range of animal species. Conservation of sequence could then be taken as a criterion for defining a putative coding region. This would fail to identify genes which were unique to man, or did not share enough homology to cross-hybridise under the stringent conditions described here. Again, this depends upon the ease with which new probes can be isolated from the genomic DNA. An alternative method is hybridisation of whole cosmids to Northern blots. Repetitive sequences can be removed by re-association with total genomic DNA prior to the addition of the probe to the hybridisation buffer. This technique has already proven sucessful with the genomic fragment X_1 (Fig. 4.6). Whichever method is chosen a large panel of RNA samples must be screened to ensure that no tissue specific transcript is overlooked. Tissue specific genes can also be mapped by hybridising cDNAs back onto blots of cosmid DNA. This has been used to locate the position of the B144 gene in the murine MHC (Tsuge *et al.*, 1987), and, albeit with an oligonucleotide rather than a transcript, to define the HSP 70 loci in the human MHC (see chapter VI).

The next step towards determining whether there is an association between the novel loci and HLA linked diseases is the identification of the potential products. The functions of housekeeping proteins, especially those which may be important for maintaining the structural integrity of the cell, could be just as significant in understanding the pathogenesis of disease as the antigen presenting functions of class I and II molecules. In this respect, the roles of the HSP 70 family members in protein folding, protecting nuclear morphology under conditions of stress,and as potent immunogens are discussed in chapter VI. In contrast, another transcript mapping 5' to C4 (G11) has been sequenced, but no homologies have been found between its putative product and other proteins in the EMBL database (chapter V).

CHAPTER V

ANALYSIS OF AN HTF ISLAND ASSOCIATED GENE: G11

5.1 INTRODUCTION

Two HTF island structures were mapped within the 30 kb gap between the genes for factor B and C4A (see chapter IV). One of these, island 13, is situated 7 kb 3' to factor B, and is probably associated with the RD gene, which has been positioned adjacent to the factor B loci in both man and mouse (Levi-Strauss et al., 1988). The second island, 14, is 9.5 kb 5' to the C4A locus. Unique region probes were isolated from both sides of the cluster of rare sites and used to define an associated transcript of ∿1.4 kb which was expressed in all five cell lines investigated. Subsequent isolation of cDNA clones from a PMA stimulated U937 library have allowed the nucleotide sequence of the mRNA species to be determined. Alternative splicing appears to produce two different transcripts from the same gene. This may be of importance in the regulation of expression of the putative product, or could result in two different peptides from alternative open reading frames.

5.2 RESULTS

5.2.1 Analysis of the HTF Island

Two HTF islands were mapped to the region between the factor B and C4A genes by comparison of restriction data from PFGE with the rare enzyme sites positioned within the cosmid clones (chapter IV). One of these, 7 kb 3' to the factor B gene is probably associated with the RD gene. The second, 9.5 kb 5' to the C4A gene, was further investigated using unique probes from both sides of the island structure. One

probe,from a 5 kb subclone spanning the region from the Sal I site in the cloned DNA to the end of the 0.8 kb Bam HI fragment, encompasses part of the island. The second probe, D3, a 0.5 kb Nhe I/ Bst EII genomic fragment, maps 1 kb telomeric to C4A (Fig. 5.1). Both of these probes were hybridised with Southern blots of Bam HI and Bgl II digested genomic DNA from man, Rhesus monkey, mouse, rat, cat, rabbit, sheep, chicken and shark. Cross-hybridisation to all the mammalian species, possibly to chicken, but not to shark, was observed with both probes. The results of the digests are shown in Fig. 5.2, and illustrate a high degree of homology in the size of the fragments detected by the Sal I probe between man and monkey. In addition, the probes appear to cross-hybridise to common fragments in the samples from cat and mouse, and could be linked in these species. This suggested that there was a conserved sequence between the factor B and C4 loci, which could correspond to a novel gene.

The probes were used to search for an associated transcript by Northern blot analysis. The panel of total RNA species was prepared from U937, PMA stimulated U937, HepG2, Molt 4, and Daudi cell lines. Probe D3 detected a transcript of ∿1.4 kb. The Sal I probe hybridised weakly to the same transcript, but a 1.7 kb Bam HI fragment, derived from the middle of the Sal I probe, did not. The transcript, designated G11, was expressed in all the cell lines analysed (Fig. 5.3).

The results suggested that the novel gene lay between the cluster of rare sites and the C4A locus, rather than between the two islands. The Sal I probe probably contains only a fraction of the 5' end of the transcript, which explains the weakly hybridising band observed on the Northern blot.

Fig. 5.1

The region between the factor B and C4A genes. The restriction sites for the enzymes Bam HI (M), Bgl II (B), Sal I (S), Bss HII (Bs), Eag I (E) ans Sac II (Sc) are shown by the vertical bars. The sites which characterise islands 13 and 14, and cleave at the chromosomal level, are also marked. The open boxes indicate the positions of the known genes, and the transcriptional orientations are shown by the horizontal arrows. The solid boxes indicate the positions of the probes used to define the G11 transcript. The horizontal bar (⟷) shows the limits of the 1.7 kb Bam HI probe derived from the Sal I probe.

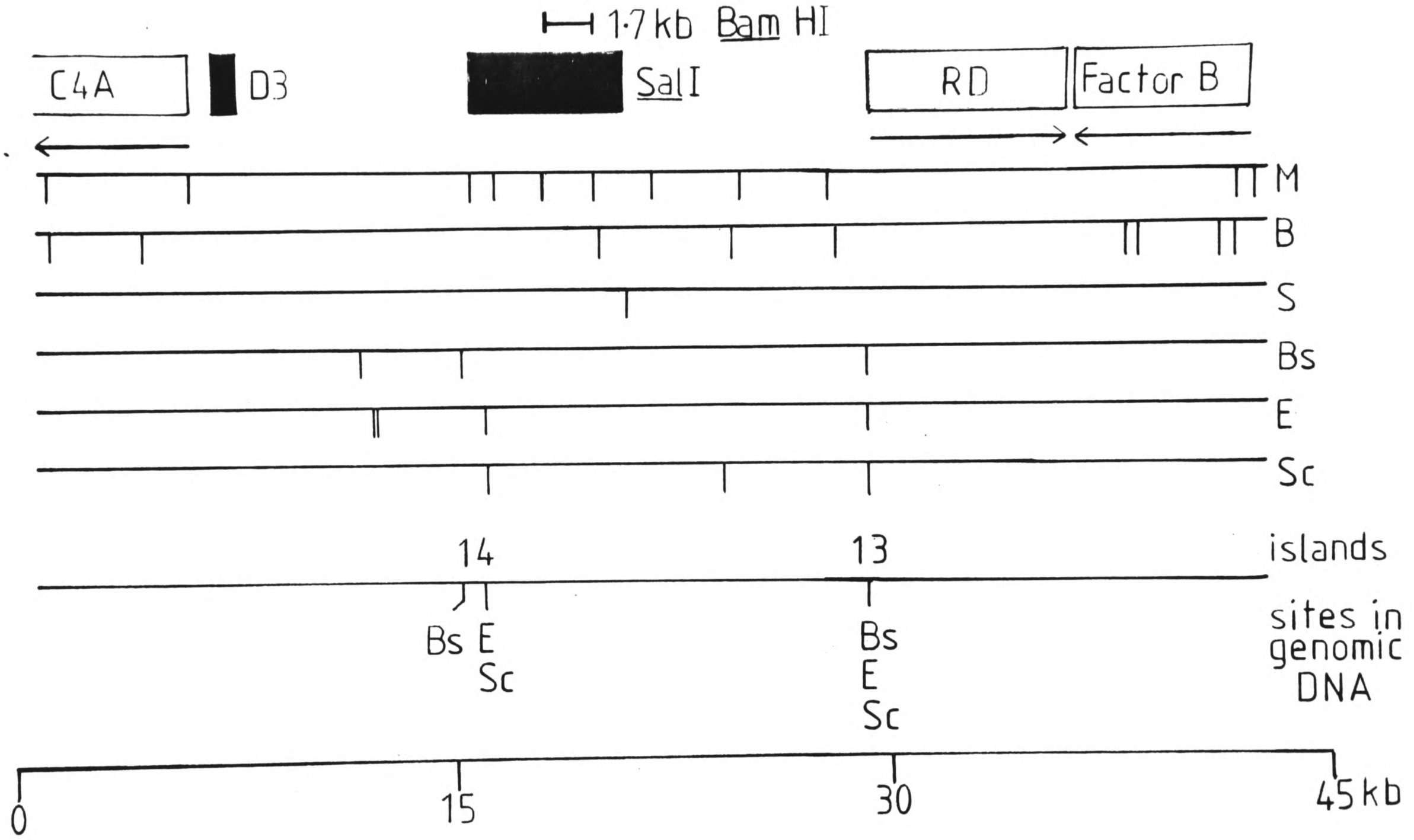
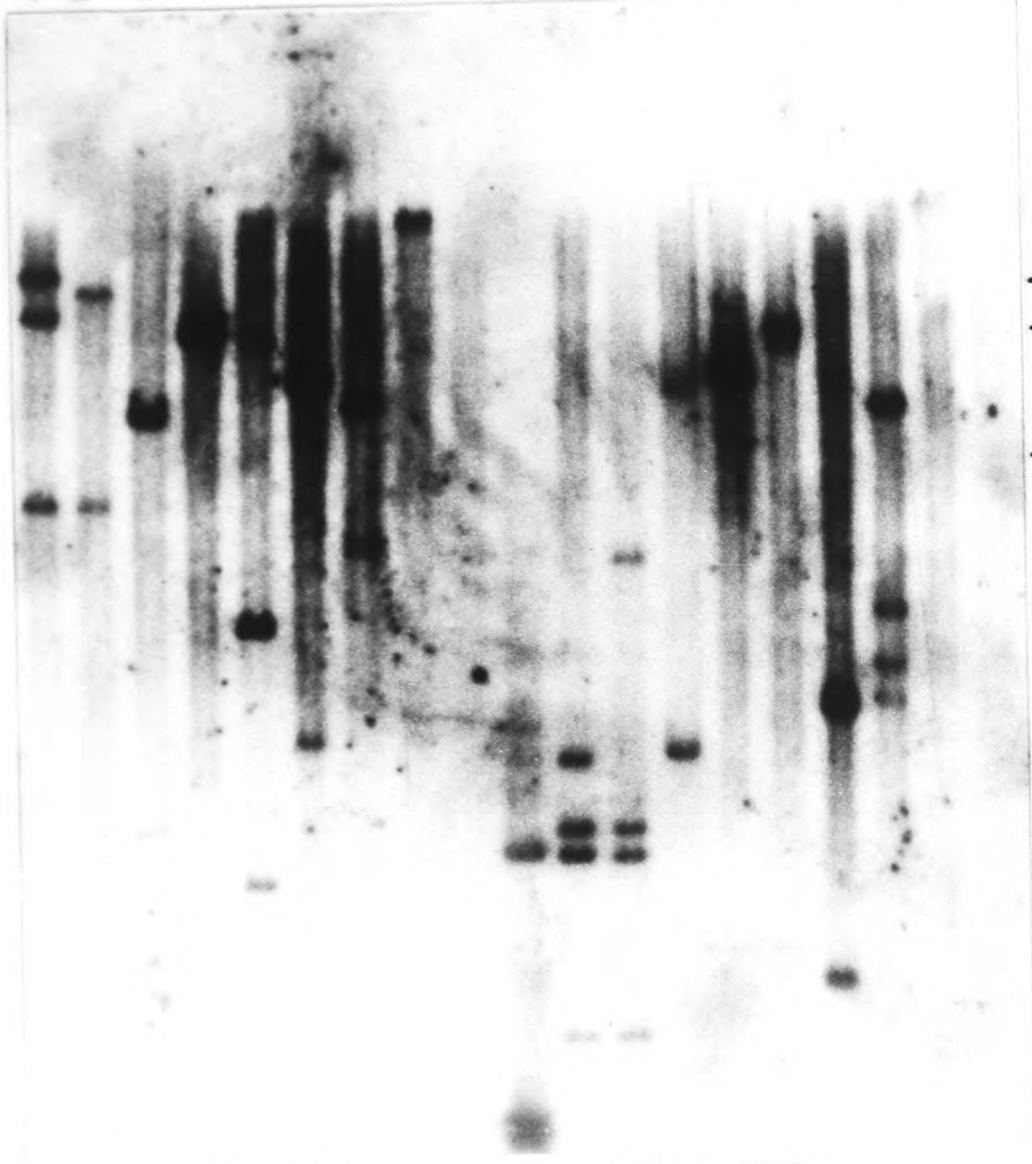

1·7 kb Bam HI
C4A
D3
Sal I
RD
Factor B
M
B
S
Bs
E
Sc
islands
14
13
Bs E Sc
Bs E Sc
sites in genomic DNA
0
15
30
45 kb

Fig. 5.2

Southern blots of genomic DNA from various animal species hybridised with probe Sal I (A) and probe D3 (B). Lanes contain samples from man (1), Rhesus monkey (2), mouse (3), hamster (4), cat (5), rabbit (6), sheep (7), chicken (8) and shark (9), digested with either Bam HI, or Bgl II. Common cross-hybridising fragments are observed in the mouse and cat, suggesting that the probes could be linked in these species.

A

1 2 3 4 5 6 7 8 9 S 1 2 3 4 5 6 7 8 9



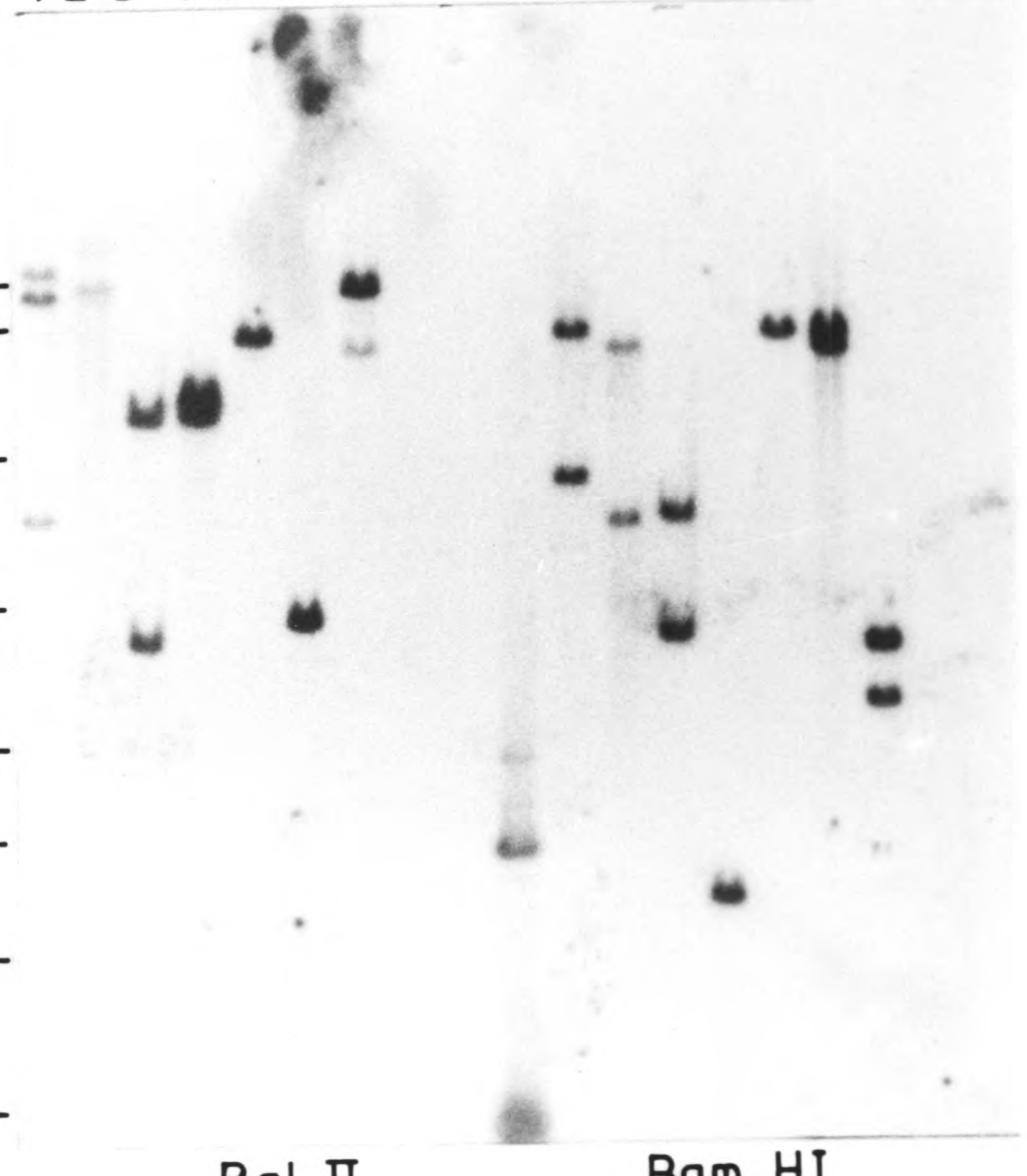
Bgl II Bam HI

kb

12·0
9·0
5·0
3·0
2·0
1·6
1·0
0·6

B

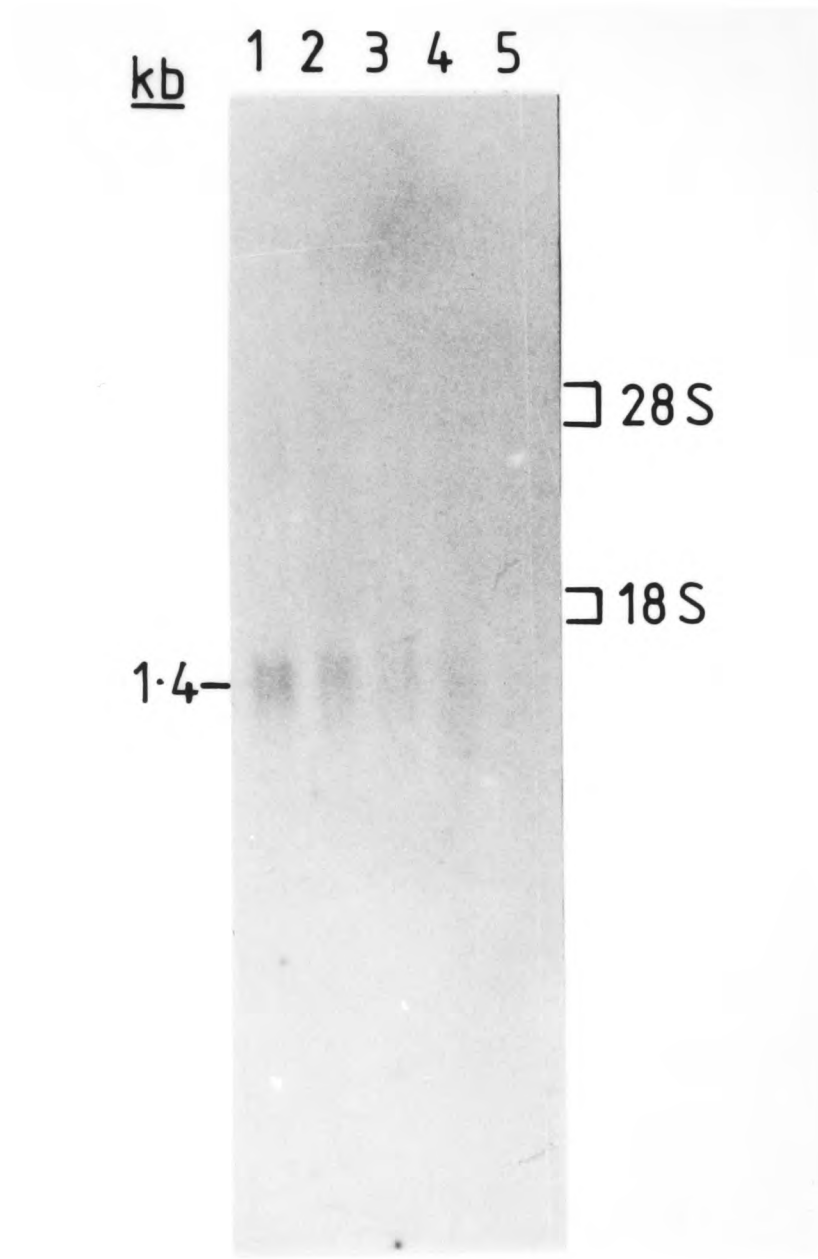
1 2 3 4 5 6 7 8 9 S 1 2 3 4 5 6 7 8 9



Bgl II Bam HI

Fig. 5.3

Hybridisation of probe D3 to a Northern blot of total RNA species isolated from U937 (1), PMA stimulated U937 (2), HepG2 (3), Molt 4 (4) and Daudi (5) cell lines. In all cases, a transcript of about 1.4 kb is visible.

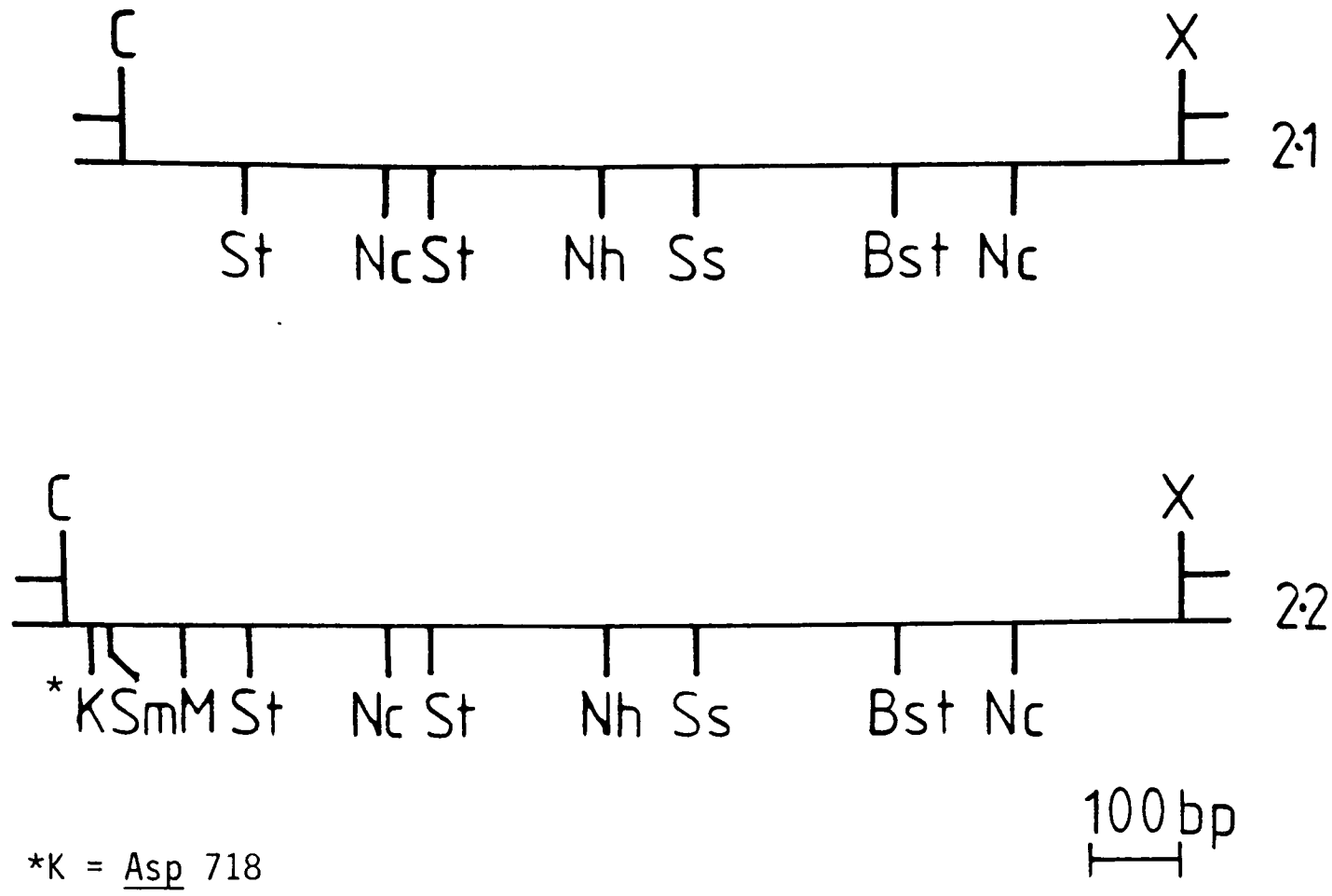

kb
1 2 3 4 5
28 S
18 S
1·4

5.2.2 Isolation of cDNA Clones

Probe D3 was used to screen an amplified cDNA library (a gift from S-K. A. Law) prepared from PMA stimulated U937 RNA and cloned into the Pvu II site of pATX. A total of 2.5x 10^5 colonies were screened, and four positives were isolated. These were mapped with the restriction enzymes Asp 718, Bam HI, Bst EII, Cla I, Nco I, Nhe I, Sma I, Sst I, Stu I and Xho I in single and double digest combinations (Fig. 5.4). Minigel analysis of the digests suggested that the positives represented two copies of each of two independent recombinants. One clone contained an insert of ∿1.2 kb (cDNA 2.1) and the other was ∿1.3 kb (cDNA 2.2). The larger cDNA differed from the smaller by the presence of internal Asp 718, Bam HI and Sma I sites, clustered at one end of the insert. In all other respects, the restriction maps appeared to be identical (Fig. 5.4).

5.2.3 Sequencing the cDNA inserts

Both the cDNA clones were subcloned into M13mp8 for sequencing by the dideoxy chain termination method. The sequencing strategy is shown in Fig. 5.5. Internal fragments from cDNA 2.1 were prepared from Nco I and Nhe I single and double digests. The ends were isolated using Nco I/ Cla I and Bst EII/ Xho I digests. Both strands of the fragments were sequenced. The restriction sites were overlapped using Msp I and Hinf I digests, once the basic sequence was available. cDNA 2.2 was subcloned and sequenced using a similar strategy (Fig. 5.5). For each cDNA, 100% of the nucleotide sequence was obtained for the 5' to 3' orientation, and ∿93% for the 3' to 5' orientation.



*K = Asp 718

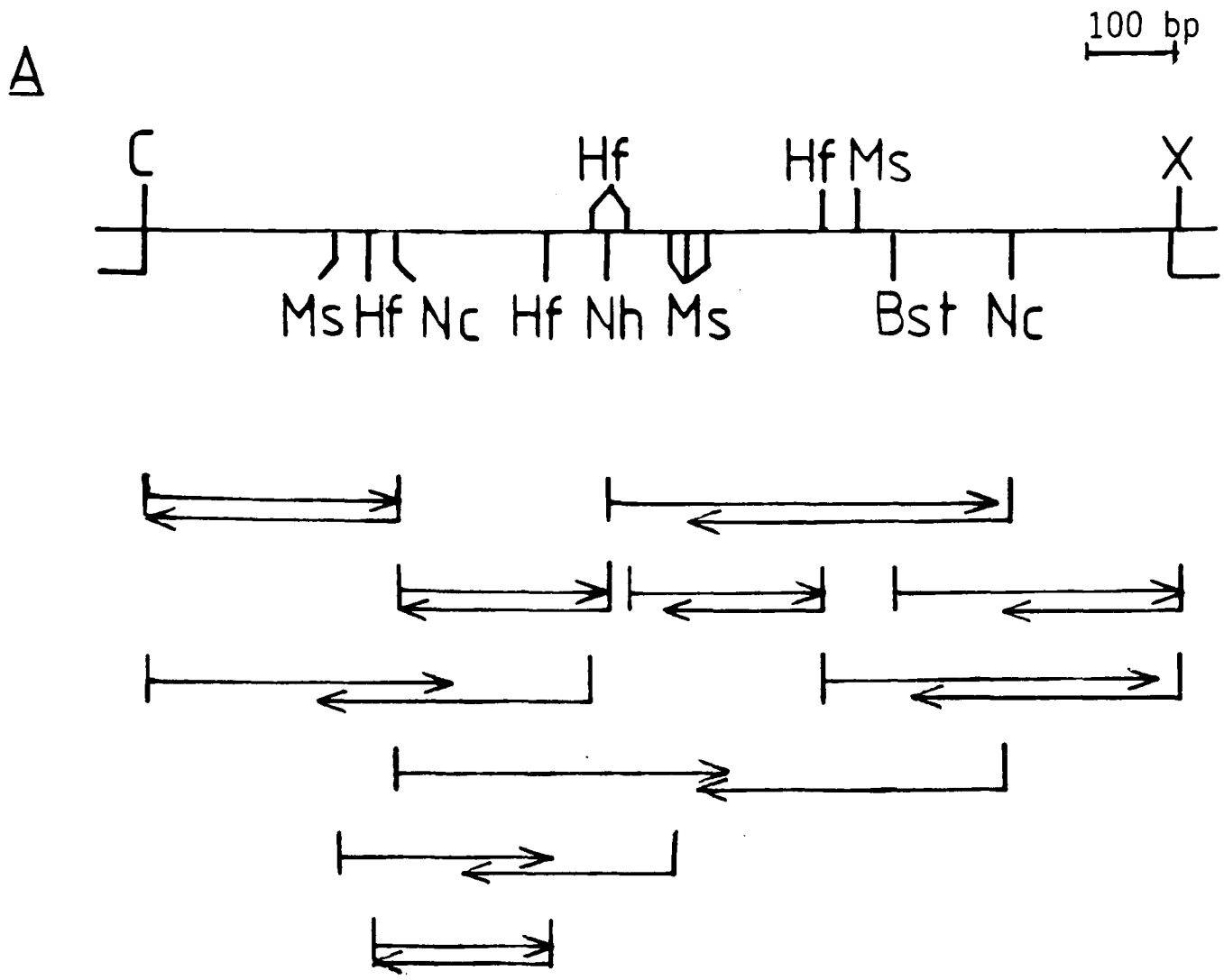
ig. 5.4

Restriction maps of cDNA 2.1 and cDNA 2.2, showing sites for the nzymes Bam HI (M), Bst EII (Bst), Kpn I (K), Sst I (Ss) and Stu I St). The sites for Cla I (C) and Xho I (X) in the vector are also arked. Vector sequences are indicated by the open boxes, and the nsert DNA by the intervening lines. The cluster of sites for M, K nd Sma I (Sm) distinguishes insert 2.2 from insert 2.1.

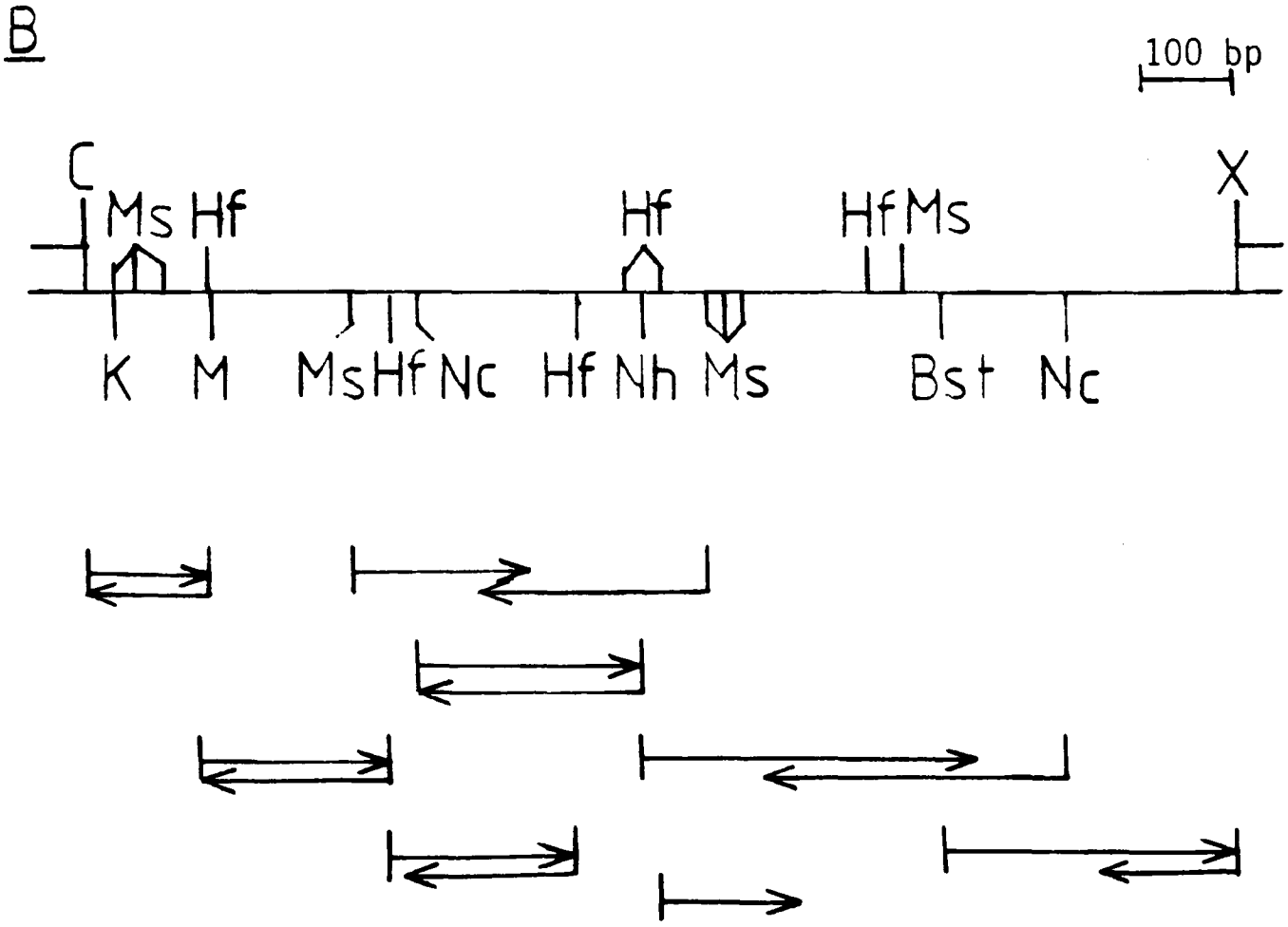
Fig 5.5

Sequencing strategy for cDNA 2.1 (A) and cDNA 2.2 (B). Restriction sites for Bst EII, (Bst), Kpn I (K), Hinf I (Hf), Nco I (Nc), Nhe I (Nh), Msp I (Ms), Bam HI (M), Cla I (C) and Xho I (X) are shown. The extent of the sequence obtained in each orientation is shown by the horizontal arrow.

A


100 bp
C
Hf
Hf Ms
X
Ms Hf Nc Hf Nh Ms
Bst Nc

B


100 bp
C
Ms Hf
Hf
Hf Ms
X
K M Ms Hf Nc Hf Nh Ms
Bst Nc

Comparison of the completed sequences revealed three differences between the clones (Fig. 5.6):

(1) cDNA 2.1 was 1137 nucleotides long, followed by 18 bases of the poly (A) tail, whereas cDNA 2.2 was 1225 nucloetides long, followed by 5 bases of the poly (A) tail,

(2) at the 5' end the sequences were identical for 200 bp 5' from the first Nco I site after which they diverged completely,

(3) the 240 bp Nco I/ Nhe I fragment from cDNA 2.2 contained an extra GpT dinucleotide at nucleotide positions 398 and 399 that was absent in cDNA 2.1.

Either, or both, of (2) and (3) could have arisen as cloning artifacts. Before analysing the cDNA sequences for open reading frames, the origins of the differences were clarified by investigating the appropriate genomic sequences. Subclones containing the 5' end of the gene for G11, and that part of the gene containing the GpT dinucleotide difference were obtained from cos2γ, as described in 5.2.4.

5.2.4 Analysis of the Genomic Sequences

The insert from cos2γ is 42 kb in length, and spans the region from the 5' end of the factor B gene to the first 5 kb of the C4A gene. The cosmid DNA was digested with Bam HI, and a 9.5 kb fragment covering the region from island 14 to the 5' end of the C4A locus was subcloned into pATX. Similarly, an Sst I digest was blotted and probed with the cDNA. Fragments of 3.3 kb and 2.6 kb were found to hybridise to the cDNA insert. These were recovered from agarose preparative gels (2.5.6) and ligated into pUC18.

The Bam HI and Sst I subclone inserts were mapped with a series of enzymes (Fig. 5.7). The digests were separated on agarose minigels,

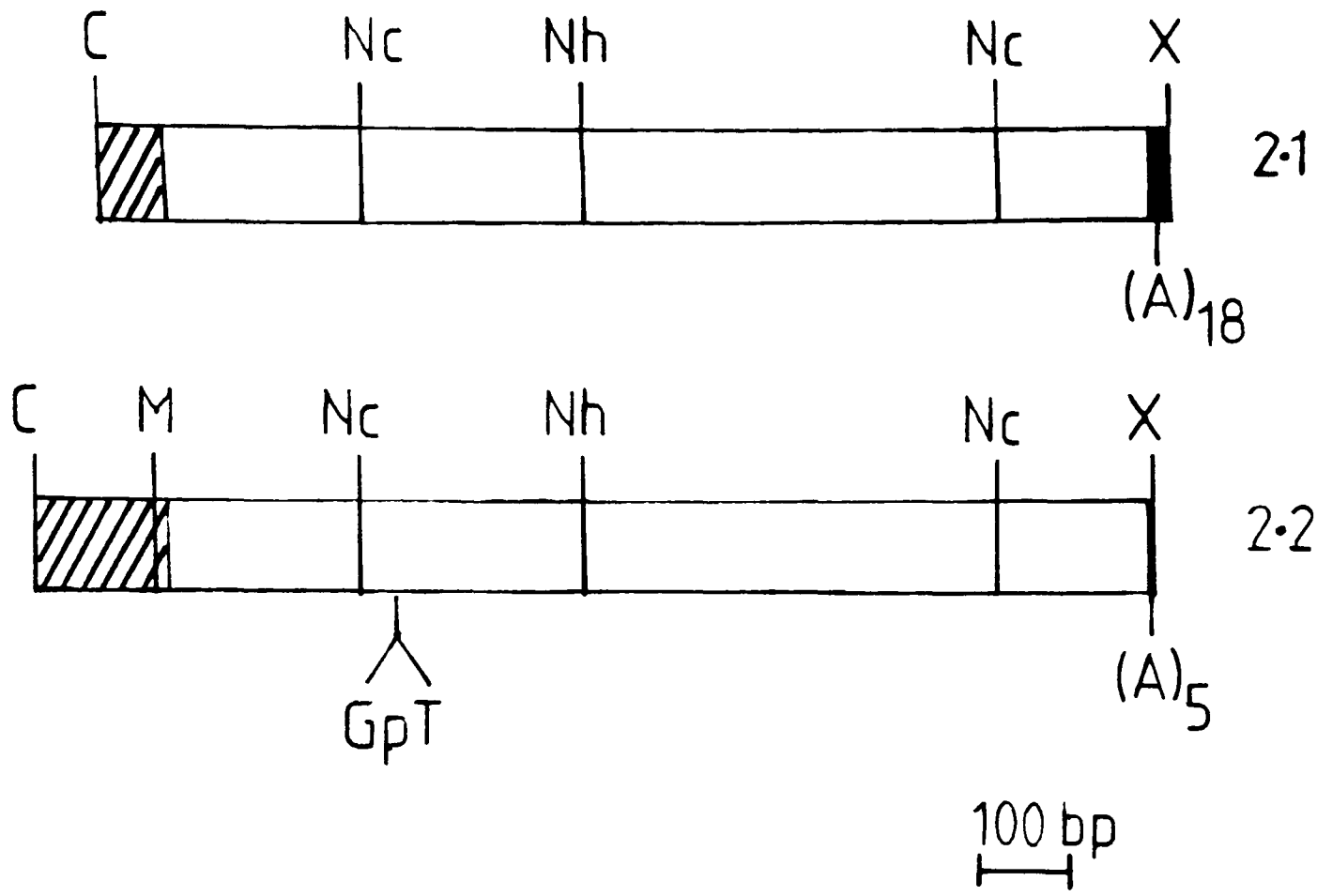


Fig 5.6

Schematic representation of the differences between cDNA 2.1 and cDNA 2.2. The 5' ends are shown as shaded boxes, and the poly (A) tails as solid boxes. The position of the extra GpT dinucleotide within the 240 bp Nhe I/ Nco I fragment of cDNA 2.2 is marked. Sites are shown for the enzymes Bam HI (M), Cla I (C), Nco I (Nc), Nhe I (Nh) and Xho I (X).

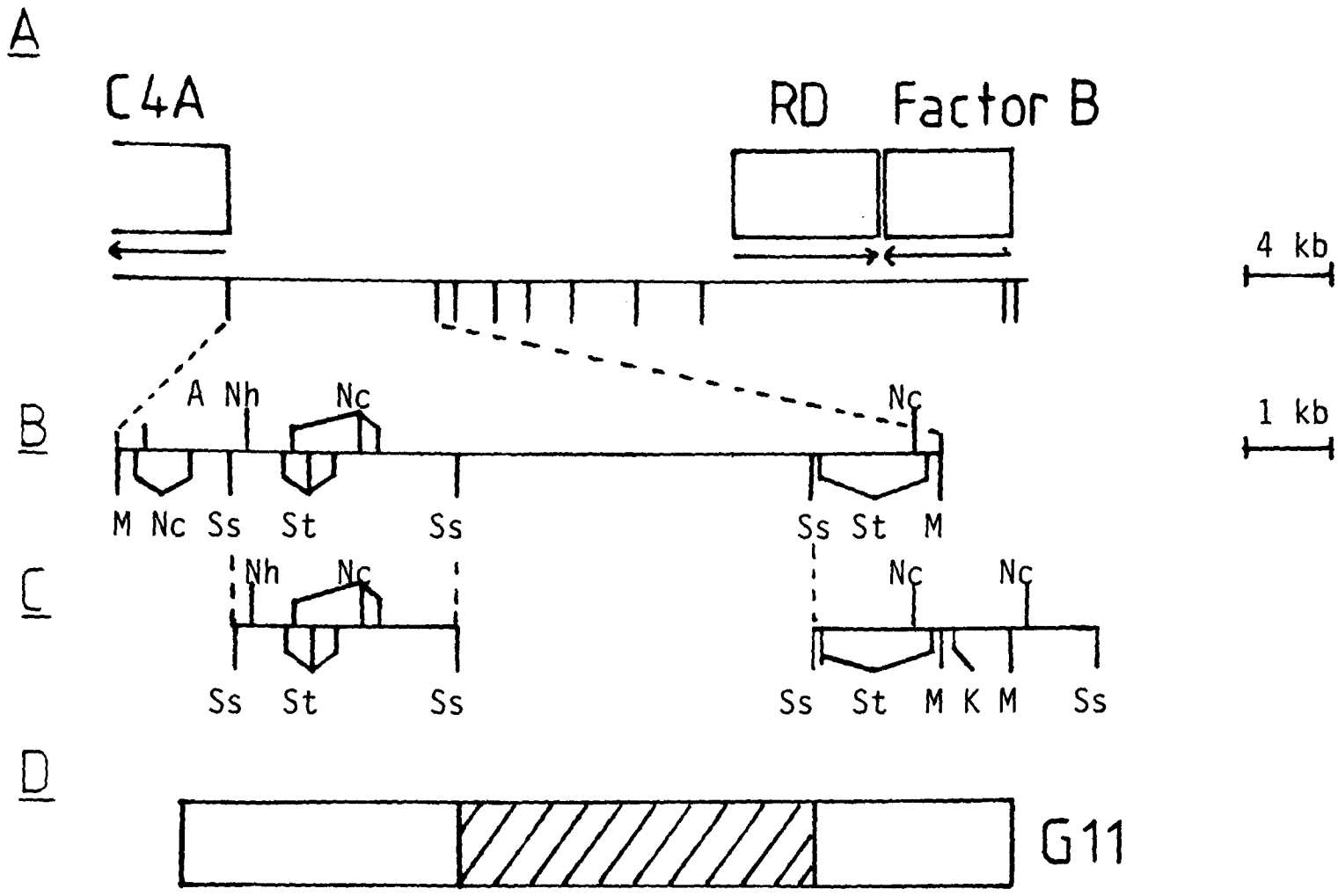


Fig 5.7

Mapping transcript G11 at the genomic level. Cos2γ spans the region from the 5' end of the factor B gene to 5 kb into the C4 gene. The extent of the insert with its Bam HI restriction sites is shown in A. The positions of the genes are indicated by open boxes, and the transcriptional orientations by horizontal arrows. The Bam HI and Sst I subclones (B and C) were mapped with Bam HI (M), Acc I (A), Nco I (Nc), Nhe I (Nh), Kpn I (K), Sst I (Ss) and Stu I (St). The limits of G11 (D) were determined from hybridisation experiments, and suggest the presence of a large intron (shaded).

* Kpn I is the isoschizimer of Asp 718.

Southern blotted, and hybridised either to the full length cDNA 2.1, or to a 280 bp Cla I/ Nco I fragment representing the 5' end of the cDNA insert. The Cla I/ Nco I probe detected both the 3.3 kb Sst I fragment, which came from the 5' end of the gene and encompassed part of the Sal I probe, and the 2.7 kb Sst I fragment. From the genomic map, this was known to lie 1.7 kb from the 5' end of the C4A gene (Fig. 5.7). Therefore, a large intron of at least 4 kb must be present in the 5' end of G11. The bulk of the coding sequence appeared to lie within a region 4 kb 5' to the C4A locus, suggesting a total gene size of 9-10 kb (Fig 5.7).

The 5' end of the genomic sequence was investigated by sequencing 750 bp Bam HI/ Asp 718 and 250 bp Asp 718/ Stu I fragments from the 3.3 kb Sst I subclone (Fig. 5.8). Both orientations of the fragments were obtained.

The nucleotide sequence was found to overlap the inserts from both cDNAs. The 5' end of cDNA 2.1 was found to consist of 66 bp from a 109 bp intron that had been spliced out of cDNA 2.2. Comparison of the genomic sequence with the longest open reading frame of cDNA 2.2 also showed that, in the absence of further introns, the codon at position 53 of the insert encoded the first in-phase methionine.

The internal GpT difference was defined by subcloning fragments of 200, 500 and 800 bp from Nco I and Nco I/ Nhe I digests of the 2.7 kb Sst I subclone (Fig. 5.8). It was shown to arise from alternative splicing at a 5' exon/ intron boundary with two possible donor splice sequences.

The exact position of the 3' end of the gene was established by sequencing ∿850 bp from the Bam HI site at the 5' end of the C4A gene. Nco I and Acc I/ Bam HI fragments of ∿600 and ∿300 bp, were subcloned from the 9.5 kb Bam HI fragment (Fig. 5.8). These were sequenced in both

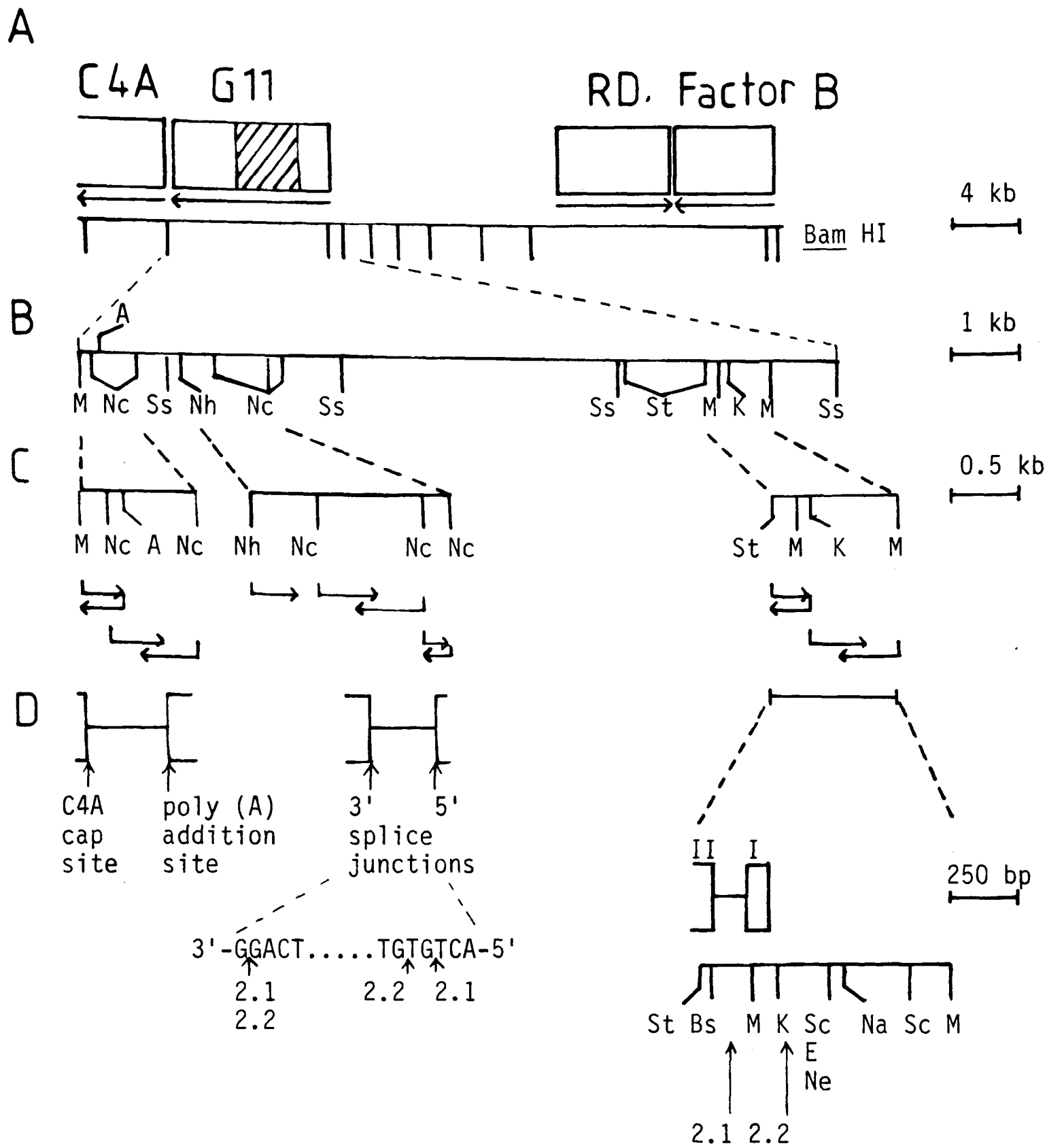


Fig. 5.8

Determination of the origins of the differences between cDNA 2.1 and cDNA 2.2. The extent of cos2γ is shown in A, with the positions of the known genes marked by open boxes. The 5' to 3' orientations are indicated by the horizontal arrows. The genomic fragments were mapped with the enzymes Acc I (A), Bam HI (M),*Kpn I (K), Nco I (Nc), Nhe I (Nh), Sst I (Ss) and Stu I (St), as shown in B. The appropriate sub-clones were sequenced (C) to define the positions of the differences (D). Sites for the rare enzymes Bss HII (Bs), Eag I (E), Nar I (Na), Nae I (Ne) and Sac II (Sc) were located in the 5' genomic sequence. Exons are shown by open boxes. *Kpn I is the isoschizimer of Asp 718.

orientations, and compared with the previous data for the region 5' to the C4A locus (Belt, 1985). The poly (A) addition site for the G11 transcript was found to lie only 613 bp upstream of the C4A mRNA cap site.

5.2.5 Analysis of the DNA Sequence

The cDNA sequences were analysed for open reading frames using the ANALYSEQ program. Putative polypeptides were then compared against known proteins in the NBRF and EMBL databases. No match at either the nucleotide or protein level was found.

cDNA 2.2 The complete sequence of cDNA 2.2 is 1230 nucleotides, including 5 bases of the poly (A) tail (Fig. 5.9). The longest open reading frame starts with an initiating methionine codon at position 53, and ends with a TGA stop codon at position 827. Thus, the 5' untranslated region is at least 52 bp in length, and the 3' untranslated region is 398 bp long. A polyadenylation signal, 5'-AATAAA-3', occurs at position 1205, 15 bases upstream of the poly (A) tail. Assuming an average poly (A) tail of 100-150 bp, the total length of the cDNA would be 1330-1380 bp, which is very close to the estimated size of 1.4 kb from the Northern blot analysis. Owing to the low relative abundance of the mRNA in total RNA, primer extension experiments failed to detect a product. Therefore, the precise position of the cap site has not yet been determined.

Comparison of the cDNA with the genomic sequence which overlaps the 5' end of cDNA 2.2 does not reveal further in-phase methionines. Although another 10 methionine codons occur within the genomic sequence, three of these are followed by stop codons (2 to 8 codons downstream). None of the methionines is associated with a putative hydrophobic leader

Fig. 5.9

The complete nucleotide sequence of cDNA 2.2. The positions of the nucleotides are marked below the sequence. The longest open reading frame is shown, with the amino acid residues encoded by the cDNA shown using the standard single letter code. The positions of the amino acid residues are numbered above the sequence. * marks the position of the stop codon.

```
                                                                  M  S  W
GTTCTCTTCCCTCCATTCCTACCCCTTCCCCCGGTACCATAAAATCCCGGGATATGAGCTG
         10         20         30         40         50         60

                    10                                  20
  K  R  H  H  L  I  P  E  T  F  G  V  K  R  R  R  K  R  G  P
GAAGAGGCATCACCTGATCCCGGAGACCTTTGGAGTTAAGAGGCGGCGGAAGCGAGGGCC
         70         80         90        100        110        120

                    30                                  40
  V  E  S  D  P  L  R  G  E  P  G  S  A  R  A  A  V  S  E  L
TGTGGAGTCGGATCCTCTTCGGGGTGAGCCAGGGTCGGCGCGCGCGGCTGTCTCAGAACT
        130        140        150        160        170        180

                    50                                  60
  M  Q  L  F  P  R  G  L  F  E  D  A  L  P  P  I  V  L  R  S
CATGCAGCTGTTCCCGCGAGGCCTGTTTGAGGACGCGCTGCCGCCCATCGTGCTGAGGAG
        190        200        210        220        230        240

                    70                                  80
  Q  V  Y  S  L  V  P  D  R  T  V  A  D  R  Q  L  K  E  L  Q
CCAGGTGTACAGCCTTGTGCCTGACAGGACCGTGGCCGACCGGCAGCTGAAGGAGCTTCA
        250        260        270        280        290        300

                    90                                 100
  E  Q  G  E  I  R  I  V  Q  L  G  F  D  L  D  A  H  G  I  I
AGAGCAGGGGGAGATCAGAATCGTCCAGCTGGGCTTCGACTTGGATGCCCATGGAATTAT
        310        320        330        340        350        360

                   110                                 120
  F  T  E  D  Y  R  T  R  V  C  D  C  V  L  K  A  C  D  G  R
CTTCACTGAGGACTACAGGACCAGAGTATGTGACTGTGTCCTCAAGGCCTGTGATGGCCG
        370        380        390        400        410        420

                   130                                 140
  P  Y  A  G  A  V  Q  K  F  L  A  S  V  L  P  A  C  G  D  L
ACCGTATGCTGGGGCAGTGCAGAAATTTCTAGCTTCAGTACTTCCAGCCTGTGGGGACCT
        430        440        450        460        470        480

                   150                                 160
  S  F  Q  Q  D  Q  M  T  Q  T  F  G  F  R  D  S  E  I  T  H
TAGTTTCCAGCAGGACCAAATGACACAGACCTTTGGCTTCAGGGACTCAGAAATCACGCA
        490        500        510        520        530        540

                   170                                 180
  L  V  N  A  G  V  L  T  V  R  D  A  G  S  W  W  L  A  V  P
TCTGGTGAATGCTGGAGTCCTCACCGTCCGAGATGCTGGGAGCTGGTGGCTAGCTGTGCC
        550        560        570        580        590        600

                   190                                 200
  G  A  G  R  F  I  K  Y  F  V  K  G  R  Q  A  V  L  S  M  V
TGGAGCTGGGAGATTCATCAAGTACTTTGTTAAAGGGCGCCAGGCTGTCCTTAGCATGGT
        610        620        630        640        650        660

                   210                                 220
  R  K  A  K  Y  R  E  L  L  L  S  E  L  L  G  R  R  A  P  V
CCGGAAGGCAAAGTACCGGGAACTGCTCCTATCAGAGCTCCTGGGCCGGCGGGCGCCTGT
        670        680        690        700        710        720

                   230                                 240
  V  V  R  L  G  L  T  Y  H  V  H  D  L  I  G  A  Q  L  V  D
CGTGGTGCGGCTTGGCCTCACCTACCATGTGCACGACCTCATTGGGGCCCAGCTAGTGGA
        730        740        750        760        770        780

                   250
  C  I  S  T  T  S  G  T  L  L  R  L  P  E  T  *
CTGCATCTCTACCACTTCAGGAACCCTCCTCCGCCTGCCAGAGACATGAAGATTCTGCTC
        790        800        810        820        830        840
```

```
ATCATTGCTCAGCTCCTCAGAGTGGGCCGGGAGGGGACTAGAAGAGCCTGCATGATGGTGG
        850         860         870         880         890         900

CTGAGACAGGGTCACCTTGGGAAGGCTTGGGAGCCAGGATGAGTGTCGGGCTCTCGTGTG
        910         920         930         940         950         960

TGCAAAAGGTCAGATGTGACTGCTGCTGTTTGCCTGGTTTCTGACCCAGTGGTGGGGTTT
        970         980         990        1000        1010        1020

GAGCAATGCTTCTCTGCCCTTCCATGGAAAGTGGAACCAGAAATGGTGCCAAGGCTGTGG
       1030        1040        1050        1060        1070        1080

CTGTTCCCTTTCGTGTAAAATGGTGCTGTTATTACTCTGTCTTGAAATAGGAAGGTGGGA
       1090        1100        1110        1120        1130        1140

TTTCTGGGGAGGCTGGTGAAGGAGGGCAGGGTTCTTTTCTCTATGTGTCATGTTAAAATT
       1150        1160        1170        1180        1190        1200

GCCAAATAAAGTACCTGTGCCTGTG(A)5
       1210        1220
```

peptide. Furthermore, all are found prior to, or within, the region which contains putative regulatory sequences that may be important for transcription of the G11 gene (see 5.2.5). Therefore, the putative open reading frame predicts a polypeptide of 258 residues. The product has no apparent leader sequence, which suggests that G11 could encode an intracellular protein. This is supported by the absence of any Asn-X-Thr/Ser sites for the addition of N-linked carbohydrate. One other unusual feature is the odd number of cysteine residues. Three of these are found very close together in the protein sequence (positions 113, 115 and 120), the fourth is at position 140 and the fifth is at position 244 (Fig. 5.9).

cDNA 2.1 The complete sequence of cDNA 2.1 is 1155 nucleotides, including 18 bases of the poly (A) tail (Fig. 5.10). It differs from cDNA 2.2 for the first 66 bases at the 5' end, and by the absence of a GpT dinucleotide between positions 311 and 312, which introduces a frame shift in the open reading frame. By comparison of the cDNA sequence with the genomic fragments isolated from cos2γ, the origins of these differences have been defined. The 5' end of cDNA 2.1 consists of 66 nucleotides from a 109 bp intron which has been spliced out of cDNA 2.2. The variation in the number of GpT dinucleotides can be mapped within 152 bp Nco I and 800 bp Nco I fragments, and shown to occur at an intron/ exon boundary. At the 5' side of the splice junction there are two possible donor sites, one of which is used in cDNA 2.1, and the other in cDNA 2.2.

The alternative splicing of the two mRNA species alters the reading frame of cDNA 2.1 with respect to cDNA 2.2. Assuming that the mRNA cap site for cDNA is in common with that for cDNA 2.2, the small intron contains an open reading frame which is in phase with the first methionine of cDNA 2.2 until the 3' splice junction. This could produce

Fig. 5.10.

The complete nucleotide sequence of cDNA 2.1. The polypeptide encoded by the open reading frame initiated from the first methionine is shown using the standard single letter code. The amino acid position is shown above the protein sequence.

Additional potential polypeptide products are listed below the nucleotide sequence. A is the polypeptide encoded by the longest open reading frame, and B is the polypeptide encoded from the first methionine in cDNA 2.2, reading through the 5' intron. The residues are numbered below the sequence data.

```
CGAGAGTTTTGTTAGAACCGCGTCCCCGCCCCAGTTCCCTGTCCGTGAGCCGATTTATCT
         10         20         30         40         50         60

                                                M  Q  L  F  P  R  G  L  F
GCCCAGGGTCGGCGCGCGGCTGTCTCAGAACTCATGCAGCTGTTCCCGCGAGGCCTGT
         70         80         90        100        110        120

   10                                  20
    E  D  A  L  P  P  I  V  L  R  S  Q  V  Y  S  L  V  P  D  R
TTGAGGACGCGCTGCCGCCCATCGTGCTGAGGAGCCAGGTGTACAGCCTTGTGCCTGACA
        130        140        150        160        170        180

   30                                  40
    T  V  A  D  R  Q  L  K  E  L  Q  E  Q  G  E  I  R  I  V  Q
GGACCGTGGCCGACCGGCAGCTGAAGGAGCTTCAAGAGCAGGGGGAGATCAGAATCGTCC
        190        200        210        220        230        240

   50                                  60
    L  G  F  D  L  D  A  H  G  I  I  F  T  E  D  Y  R  T  R  V
AGCTGGGCTTCGACTTGGATGCCCATGGAATTATCTTCACTGAGGACTACAGGACCAGAG
        250        260        270        280        290        300

   70
    C  D  C  P  Q  G  L  *
TATGTGACTGTCCTCAAGGCCTGTGATGGCCCGACCGTATGCTGGGGCAGTGCAGAAATTT
        310        320        330        340        350        360

CTAGCTTCAGTACTTCCAGCCTGTGGGGACCTTAGTTTCCAGCAGGACCAAATGACACAG
        370        380        390        400        410        420

ACCTTTGGCTTCAGGGACTCAGAAATCACGCATCTGGTGAATGCTGGAGTCCTCACCGTC
        430        440        450        460        470        480

CGAGATGCTGGGAGCTGGTGGCTAGCTGTGCCTGGAGCTGGGAGATTCATCAAGTACTTT
        490        500        510        520        530        540

GTTAAAGGGCGCCAGGCTGTCCTTAGCATGGTCCGGAAGGCAAAGTACCGGGAACTGCTC
        550        560        570        580        590        600

CTATCAGAGCTCCTGGGCCGGCGGGCGCCTGTCGTGGTGCGGCTTGGCCTCACCTACCAT
        610        620        630        640        650        660

GTGCACGACCTCATTGGGGCCCAGCTAGTGGACTGCATCTCTACCACTTCAGGAACCCTC
        670        680        690        700        710        720

CTCCGCCTGCCAGAGACATGAAGATTCTGCTCATCATTGCTCAGCTCCTCAGAGTGGGCC
        730        740        750        760        770        780

GGGAGGGGACTAGAAGAGCTGCATGATGGTGGCTGAGACAGGGTCACCTTGGGAAGGCTT
        790        800        810        820        830        840
```

```
GGGAGCCAGGATGAGTGTCGGGCTCTCGTGTGTGCAAAAGGTCAGATGTGACTGCTGCTG
        850        860        870        880        890        900

TTTGCCTGGTTTCTGACCCAGTGGTGGGGTTTGAGCAATGCTTCTCTGCCCTTCCATGGA
        910        920        930        940        950        960

AAGTGGAACCAGAAATGGTGCCAAGGCTGTGGCTGTTCCCTTTCGTGTAAAATGGTGCTG
        970        980        990       1000       1010       1020

TTATTACTCTGTCTTGAAATAGGAAGGTGGGATTTCTGGGGAGGCTGGTGAAGGAGGGCA
       1030       1040       1050       1060       1070       1080

GGGTTCTTTTCTCTATGTGTCATGTTAAAATTGCCAAATAAAGTACCTGTGCCTGTG(A)18
       1090       1100       1110       1120       1130
```

A

```
M P M E L S S L R T T G P E Y V T V L K A C D G R P Y A G A
                  10                  20                  30

V Q K F L A S V L P A C G D L S F Q Q D Q M T Q T F G F R D
                  40                  50                  60

S E I T H L V N A G V L T V R D A G S W W L A V P G A G R F
                  70                  80                  90

I K Y F V K G R Q A V L S M V R K A K Y R E L L L S E L L G
                 100                 110                 120

R R A P V V V R L G L T Y H V H D L I G A Q L V D C I S T T
                 130                 140                 150

S G T L L R L P E T *
                 160
```

B

```
M S W K R H H L I P E T F G V K R R R K R G P V E S D P L R
                  10                  20                  30

G E P G N H G N P G G G A S L P V A E S F V R T A S P P Q F
                  40                  50                  60

P V R E P I Y L P R V G A R G C L R T H A A V P A R P V *
                  70                  80
```

a truncated peptide consisting of 35 residues identical with the N-terminus of the predicted protein from cDNA 2.2, 36 residues encoded by the intron, and 18 residues from the +2 reading frame. Alternatively, if the mRNA cap site for cDNA 2.1 is different, initiation from the first methionine codon at position 96 (Fig. 5.10) would give a product with 72 residues in common with the polypeptide encoded by cDNA 2.2, plus 4 unique residues at the C-terminus. These occur after the position of the frame shift introduced by the loss of the GpT dinucleotide in cDNA 2.1. The longest open reading frame of cDNA 2.1 consists of 143 amino acids in common with the C-terminus of the product of cDNA 2.2, and a unique N-terminus of 17 residues, encoded prior to the internal frame shift. The first methionine in this reading frame is at nucleotide position 259.

Even assuming that cDNA 2.1 is a partially spliced precursor, and that the final mRNA species is devoid of the small intron, initiation from the same methionine as cDNA 2.2 would produce a truncated polypeptide of 119 residues.

5.2.5 5' Genomic Sequence (Fig. 5.11)

Taking the first in-phase methionine residue of cDNA 2.2 as the initiating methionine, the 5' genomic sequence was analysed for the presence of upstream transcriptional regulatory sequences. An exact match for a CAAT box (5'-GCCAAT-3') was found 239 to 244 nucleotides upstream from the methionine residue. The closest match for a TATA box (5'-TATGT-3') was found 44 bases 3' to the CAAT sequence. Other consensus sequences were found for the following transcription factors:

(a) three sites for Sp1 (5'-GGGCGG-3'), one of which is 165 nucleotides 3' to the initiating methionine,

Fig. 5.11

The 5' genomic sequence. The positions of exact matches for transcription factors are marked by the horizontal arrows. The start sites of the nucleotide sequences for cDNA 2.1 and cDNA 2.2 are indicated by the vertical arrow-heads. The amino acid residues encoded by cDNA 2.2 are shown using the standard single letter code.

```
CGATCCATGAGGTCCTAAGACAAGCAGGGGTACAGAGTTTCCATTCTACAGAGGAGGCCT
         10        20        30        40        50        60

GGAGAAGGATGACTGGTTTAGGACTAAGCGAGCCACCTGATCGCCAGGCTCTGGCCTTGA
         70        80        90       100       110       120

AACATTCAGGCCCCTCAGACGCCACCGCGGCCAAGCTCTCATCCTGCCTCTTTCCTTGCC
        130       140       150       160       170       180

CTTCACCCACCCTCCCTCCAGGTCCTCCAAATGCAGTGAGGTTAGGAAGGACGTCTGCGC
        190       200       210       220       230       240

                                                      CRE
                                                     <----
TCAGATCAAGAATCCAGTTACCTCAAAGCTCCCCAACTTCCACCTCCGCAGAGCTATGAC
        250       260       270       280       290       300

                                    OCT
                                   <----
GTCATGGCAGGCACGCCAGAGGCCGAAGGATGCAAAGTGGTTTCTGCTTTCGATGATGCA
        310       320       330       340       350       360

                                    Sp1
                                   ---->
ATCATTCAGCGACAGTGGCGGGCAAACCTCCGGGCGGGGAGGTGTGAGCTTCACGAAGGA
        370       380       390       400       410       420

                                                  AP2
                                                 ---->
GGTTGACACCAACGTGGCCACCGGCGCCCCTCCACGCCGCCAACGAGTCCCCGGGCGTGC
        430       440       450       460       470       480

                    CAAT
                   ---->
GTGCCCTTGGAGGGAGCCAATCCGCGGCCGGCGTGGGGCCCGGCCTGGCGGAGGTGAATG
        490       500       510       520       530       540

TATA                 Sp1
---->               <----
CTGGTATGTGCGTCGCCACCGCCCCTCCCAGCACTGACGGGCCTGAGGGACGACAAGTTG
        550       560       570       580       590       600

          CRE                                 AP2         CRE
         ---->                               <----       ---->
ACGCTCCTTTCGTCATCACCTGGTCTAGGAGGGACGCCCGGGGAGACCGTACGTCACTGC
        610       620       630       640       650       660

                                  2.2
                                   v
TCTGCGCCGGAAGACCCTATTTTCAGGGTTCTCTTCCCTCCATTCCTACCCCTTCCCCGG
        670       680       690       700       710       720

                                                  10
                          M  S  W  K  R  H  H  L  I  P  E  T  F  G
TACCATAAAATCCCGGGATATGAGCTGGAAGAGGCATCACCTGATCCCGGAGACCTTTGG
        730       740       750       760       770       780

                    20                                    30
   V  K  R  R  R  K  R  G  P  V  E  S  D  P  L  R  G  E  P
AGTTAAGAGGCGGCGGAAGCGAGGGCCTGTGGAGTCGGATCCTCTTCGGGGTGAGCCAGG
        790       800       810       820       830       840

                                        2.1
                                         v
TAACCATGGCAACCCCGGGGGTGGGGCCTCGCTTCCGGTAGCCGAGAGTTTTGTTAGAAC
        850       860       870       880       890       900

          Sp1                                   G  S  A  R  A
         <----
CGCGTCCCCGCCCCAGTTCCCTGTCCGTGAGCCGATTTATCTGCCCAGGGTCGGCGCGCG
        910       920       930       940       950       960

      40                                    50
   A  V  S  E  L  M  Q  L  F  P  R  G
CGGCTGTCTCAGAACTCATGCAGCTGTTCCCGCGAGGCCT
        970       980       990      1000
```

(b) two sites for AP2 (5'-CCCC[G/A]GGC-3'),

(c) a near match for CTF/NF1, which includes the CAAT box,

(d) near matches for AP1 and AP5.

Other motifs found within the genomic sequence were those for the cyclic AMP response element (CRE) and the immunoglobulin octamer. The positions of these elements are summarised in Fig. 5.12.

If the CATT and TATA boxes have been correctly identified, the cap site of the G11 transcript ought to lie 20-30 bp downstream of the TATA box, i.e., between positions 569 and 579 of the 5' genomic sequence (Fig. 5.11). Therefore, the majority of the regulatory elements defined occur at locations that are consistent with the predicted cap site.

HTF Island Structure: the sequences of the 5' genomic DNA fragments and the cDNA 2.2 were analysed for the distribution of CpG dinucleotides. The ratio of observed/ expected (O/E) CpG was calculated as described by Gardiner-Garden and Frommer (1987).

$$\text{O/E CpG} = \frac{\text{no. CpG}}{\text{no.Cx no.G}} \times N$$

where N is the total number of nucleotides in the sample analysed. Both the genomic and cDNA sequences were characterised in 100 bp windows, moving across the sequence in 20 bp intervals. The average %C+G and the O/E CpG were plotted and compared against the distribution of CpG and GpC dinucleotides within the same sample groups. CpG rich regions were defined where the moving average was >50% C+G and O/E CpG was >0.6 (Fig. 5.13; 5.14).

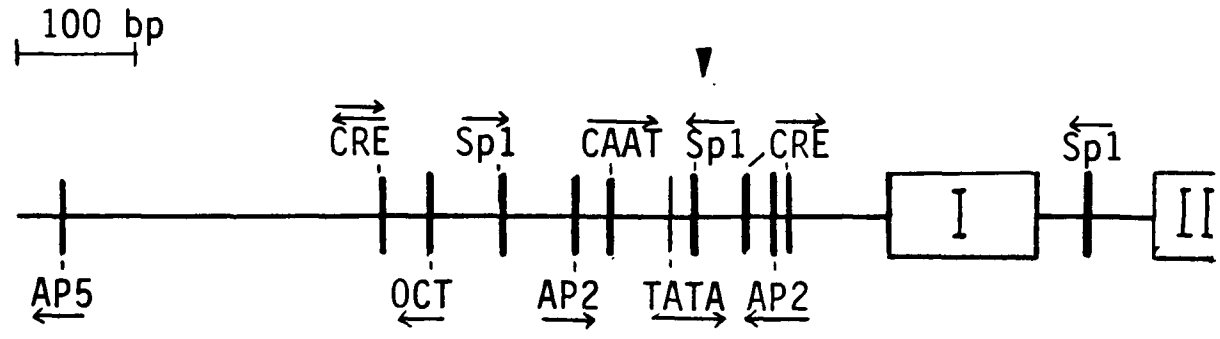
The HTF island structure was found to extend into the 5' end of the cDNA, and probably encompasses a total region of 0.8 to 0.9 kb, which includes part of the first exon and intron. A short CpG rich sequence was also detected between 0.7 and 0.85 kb in the cDNA (Fig. 5.14).

Fig. 5.12

TRANSCRIPTION FACTORS

Factor	*Recognition Sequence*	*5' genomic*	*position*
Sp1[a,b]	*GGGCGG*	*GGGCGG*	*392-397*
		CCGCCC	*559-564*
		CCGCCC	*908-913*
CRE[a]	*TT/GA CGTCA*	*TGACGTC*	*297-303*
		GACGTCA	*298-304*
		TTCGTCA	*609-615*
		TACGTCA	*650-656*
OCT[a]	*ATGCAAA T/G*	*ATGCAAAG*	*330-337*
AP2[a]	*CCCC G/A GGC*	*CCCCGGGC*	*469-476*
		GCCCGGGG	*636-643*
CTF/NF1[a]	*TGGCT(N)3 AGCCAA*	** ** TTGGAGGG*	*487-500*
		AGCCAA	
AP1[a]	*T T/G AGTCA*	** TTCGTCA*	*609-615*
		** GGACTAA*	*81-87*
AP5[a]	*CTGTGGAATG*	** CATTCTACAG*	*42-51*

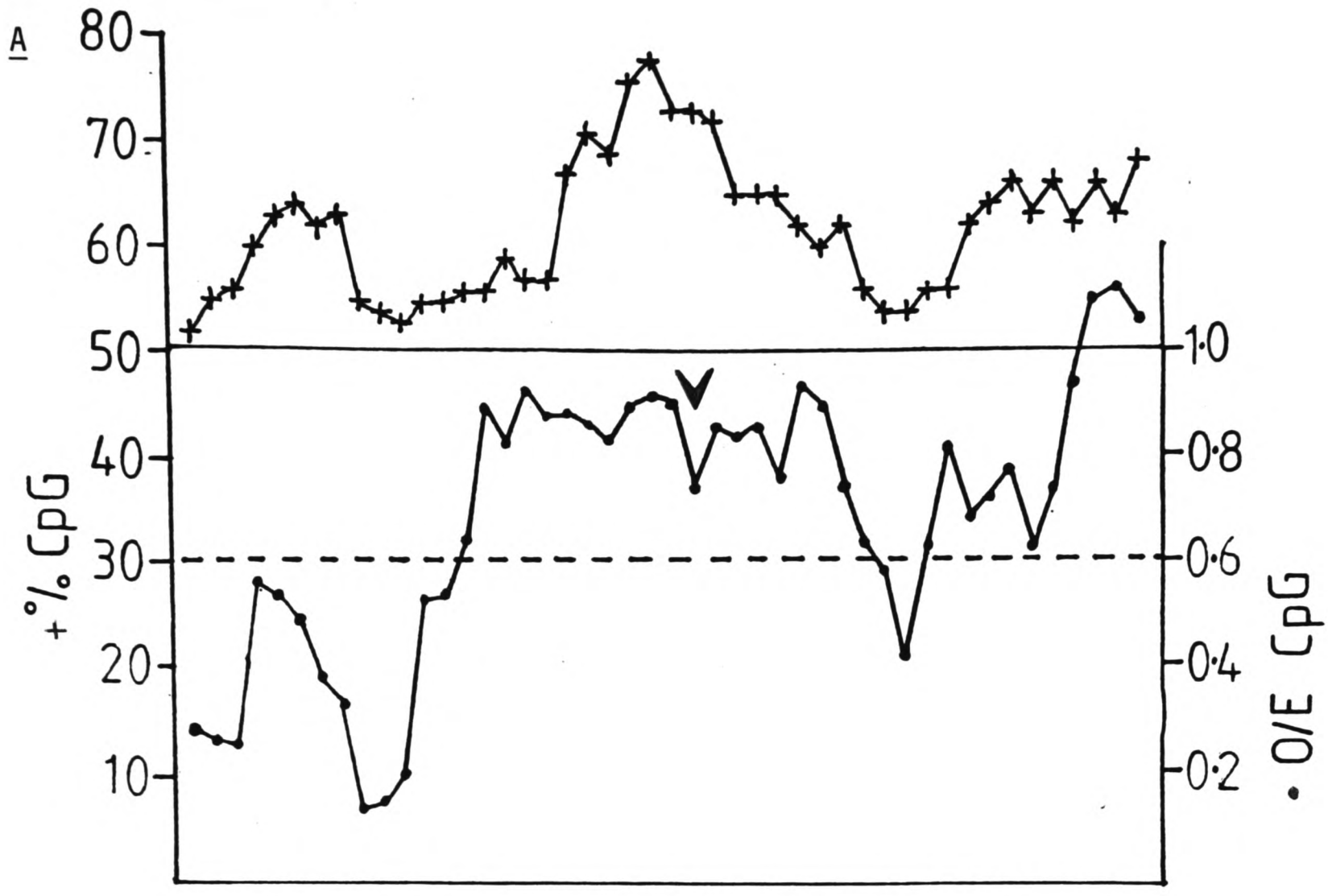
a, Jones, et al., 1988.
b, Dynan and Tjian, 1985.
**, indicates position of mismatch.*

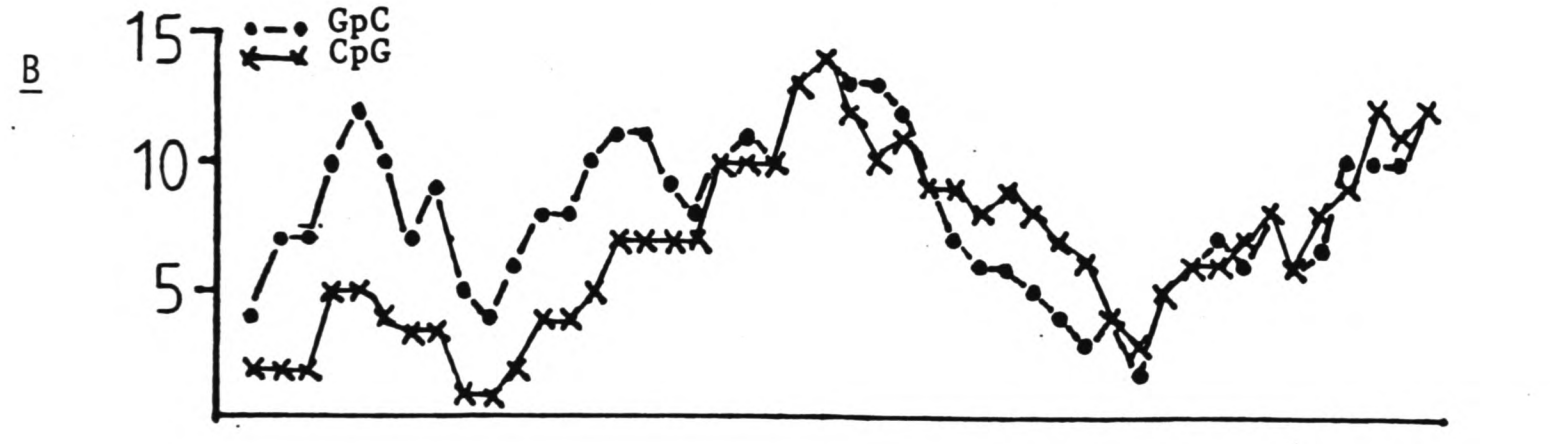


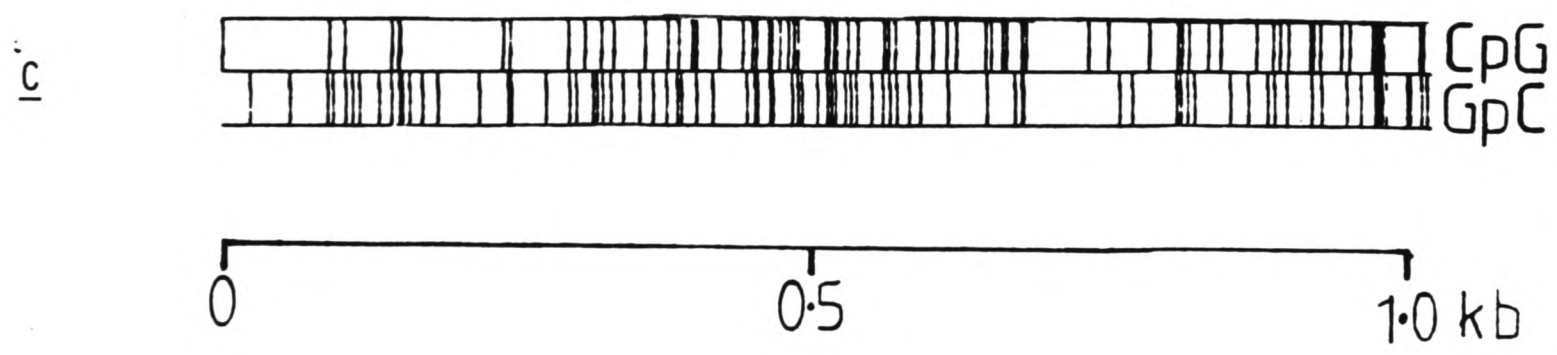
Positions of the recognition sequences relative to the putative coding regions of gene G11. The estimated position of the mRNA cap site is indicated by the arrow head.

Fig. 5.13

Analysis of the CpG content of the 5' genomic sequence. Part A shows the total CpG content and the observed/ expected CpG content (O/E), as calculated with a 100 bp window over 20 bp intervals (see text). The solid horizontal line represents a CpG content of 50%, and the broken horizontal line shows a O/E of 0.6. The arrow-head defines the position of the dip in the CpG content which could be indicative of the mRNA cap site. Part B shows the number of CpG and GpC dinucleotides using the same moving window, and part C shows their relative positions. Part D is a schematic representation of the 5' genomic sequence, with the positions of the sites for transcription factors, exons, and cDNA start sites indicated by the vertical bars.


A
80
70
60
50
40
30
20
10
+% CpG
1·0
0·8
0·6
0·4
0·2
• O/E CpG


B
15
10
5
GpC
CpG


C
CpG
GpC
0
0·5
1·0 kb

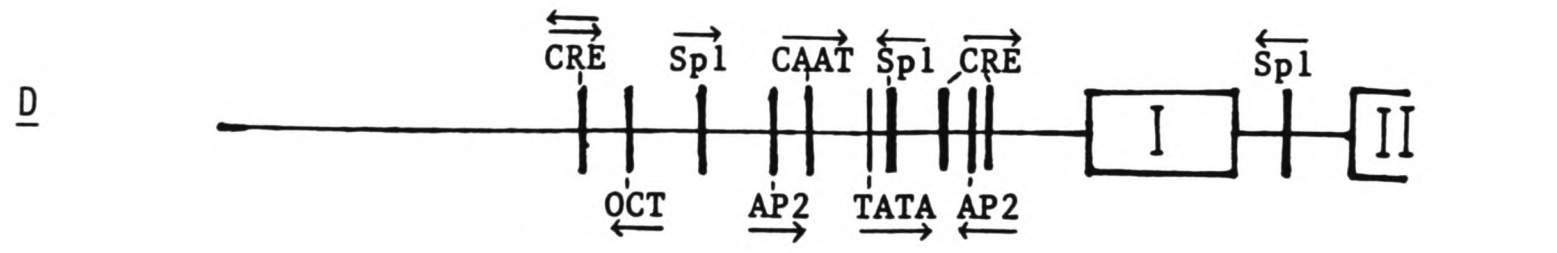
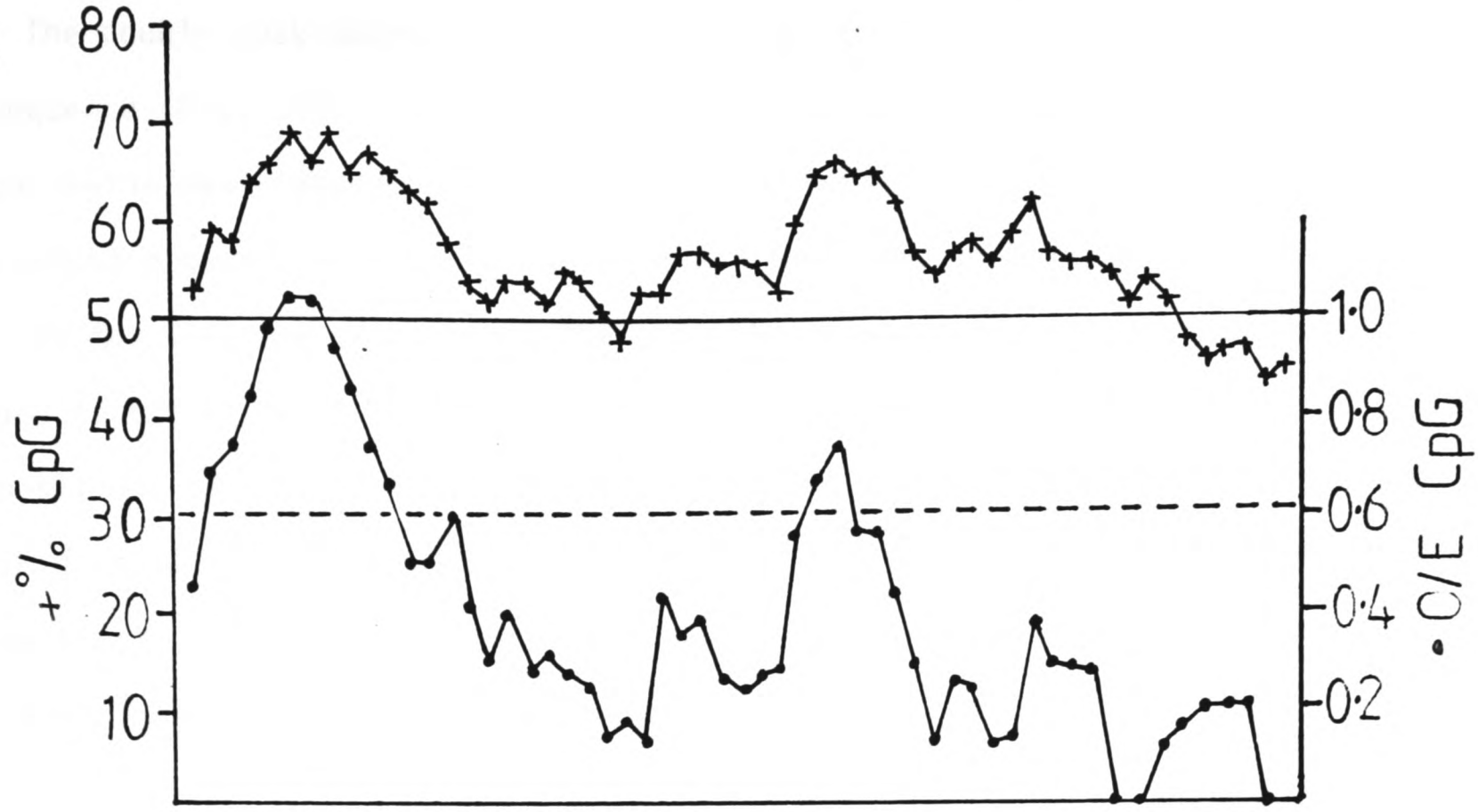

D
CRE
Sp1
CAAT
Sp1
CRE
Sp1
I
II
OCT
AP2
TATA
AP2

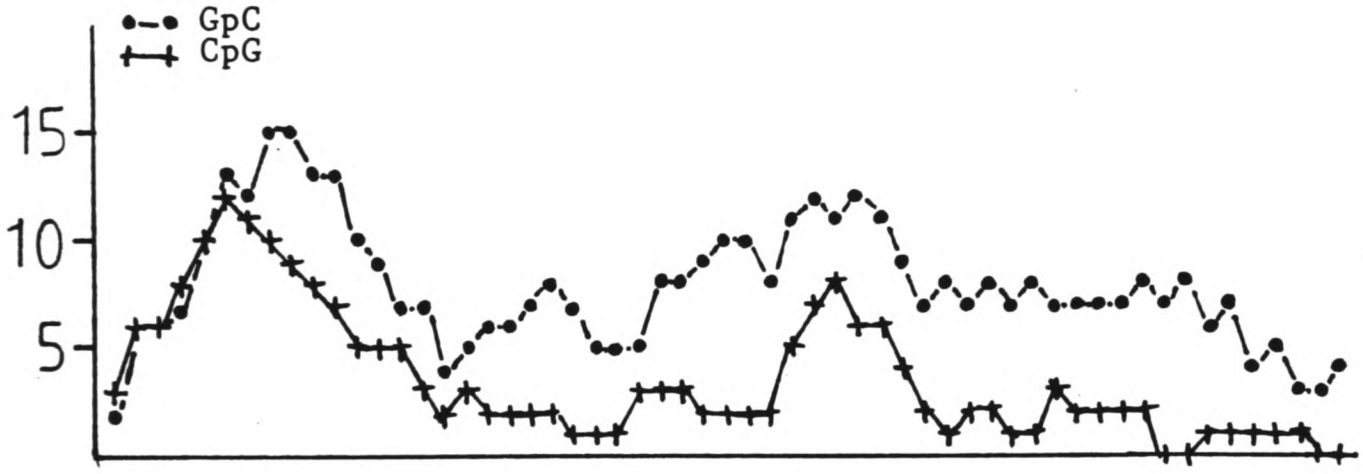
Fig. 5.14

Analysis of the CpG content of the cDNA 2.2 nucleotide sequence. Part A shows the total CpG content and the observed/ expected CpG content (O/E), as calculated with a 100 bp window over 20 bp intervals (see text). The solid horizontal line represents a CpG content of 50%, and the broken horizontal line shows a O/E of 0.6. Part B shows the number of CpG and GpC dinucleotides using the same moving window, and part C shows their relative positions.

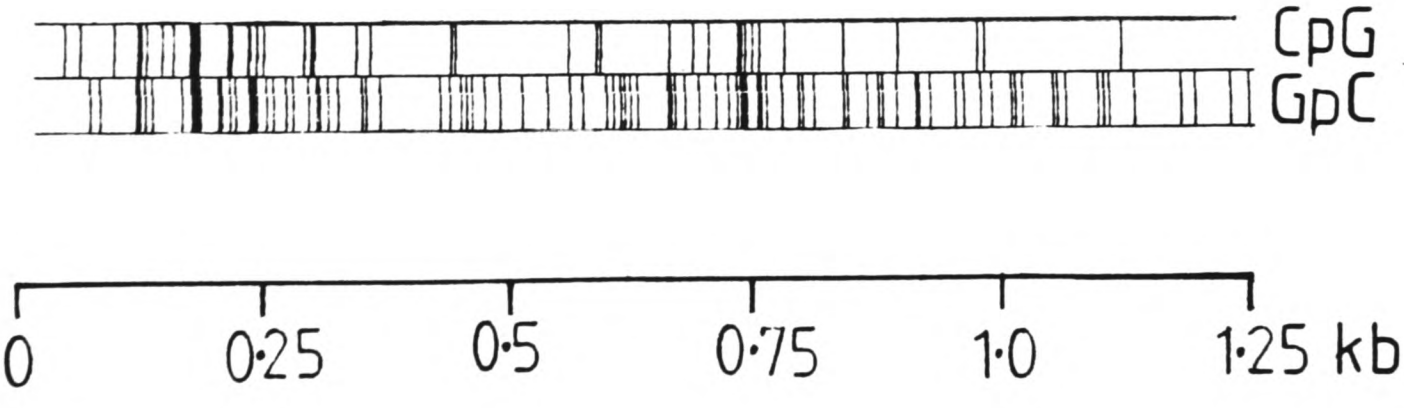
A


80
70
60
50
40
30
20
10
% CpG
1·0
0·8
0·6
0·4
0·2
O/E CpG

B


GpC
CpG
15
10
5

C


CpG
GpC
0
0·25
0·5
0·75
1·0
1·25 kb

The double peak observed in the O/E CpG plot for the 5' genomic sequence (Fig. 5.13) has been seen in the island associated with the nucleolin gene (Bourbon et al., 1988). The local decrease in the CpG content occurs close to the cap site for the mRNA. If this is also the case for G11, the position of the cap site ought to between positions 520 to 620 of the 5' genomic sequence. This is consistent with the positions of the putative CAAT and TATA boxes at positions 496 and 545 (Fig. 5.11), and would imply that the cDNA 2.2 insert is about 100-110 bp short of the 5' end of the transcript. This would predict a full length mRNA species of 1430-1480 bp.

5.2.6 3' Genomic Sequence (Fig. 5.15)

The 3' genomic sequence was found to overlap the ends of the cDNA inserts by ∿180 bp. The polyadenylation signal (5'-AATAAA-3') is at a position which is only 613 bp upstream of the C4A mRNA cap site (Belt, 1985).

5.2.7 Analysis by Southern Blotting

Further analysis at a genomic level was carried out by hybridising the cDNA with a Southern blot of genomic DNA from individuals A and YB (Fig. 5.16). A is equivalent to the cell line used to prepare the cosmid library, and YB is an individual with duplicated C4 and 21OH loci (HLA type: HLA-A2, 3; -B35, 62; C2C; Bf F; C4 A3, 2; C4 BQO; HLA-DR1, 6). The samples were digested with the enzymes Bam HI, Bgl II, Hind III and Taq I. Following washing and autoradiography, the probe was found to hybridise to fragments at 9 and 0.8 kb (Bam HI), >14 kb (Bgl II), >20 and >12 kb (Hind III), and 7 and 0.95 kb (Taq I) in both sets of genomic

Fig. 5.15

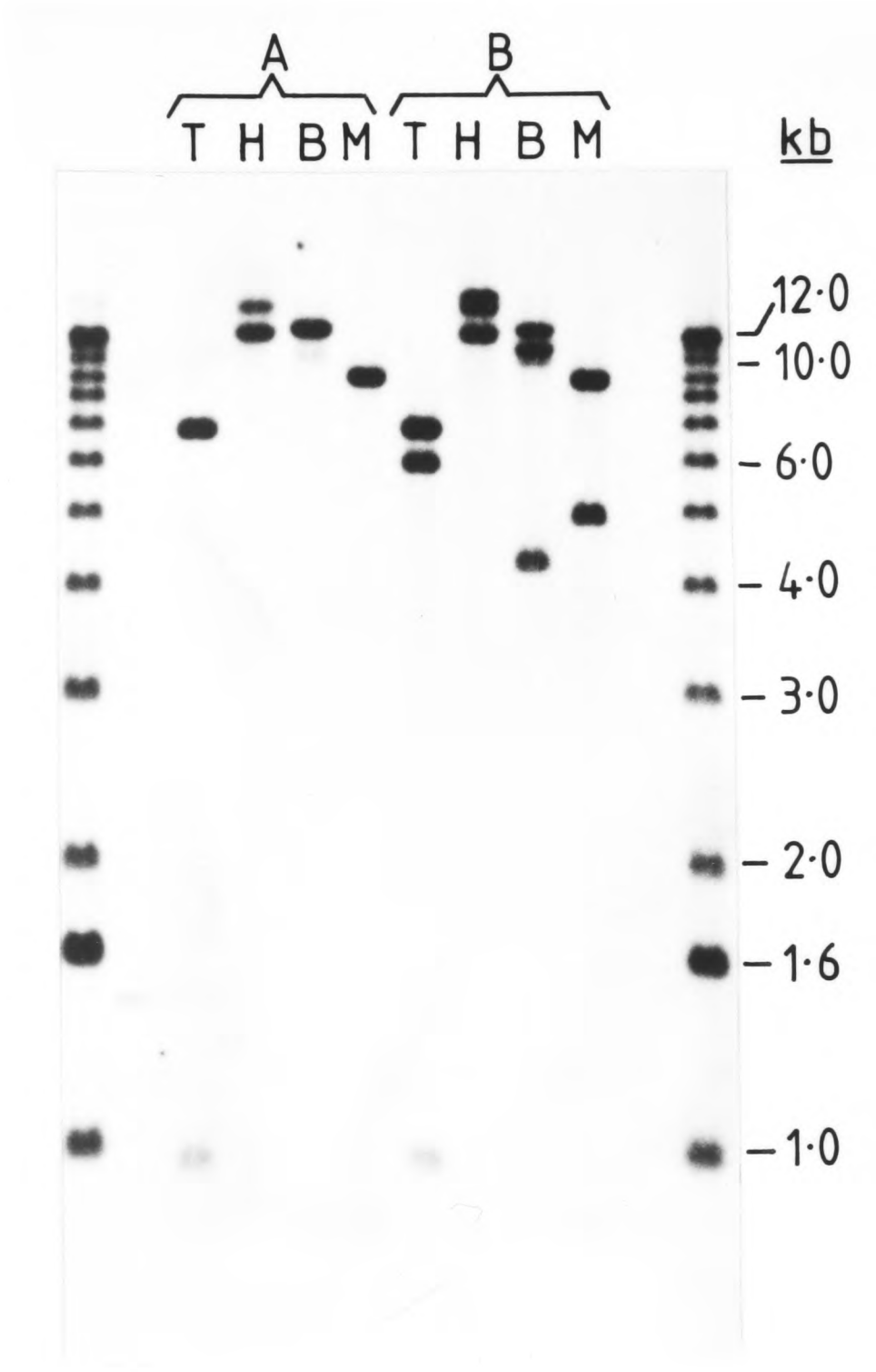
The nucleotide sequence of the 3' genomic clones. The positions of the poly (A) addition site is indicated by the vertical arrow-head. The C4A mRNA cap site is marked with an *.

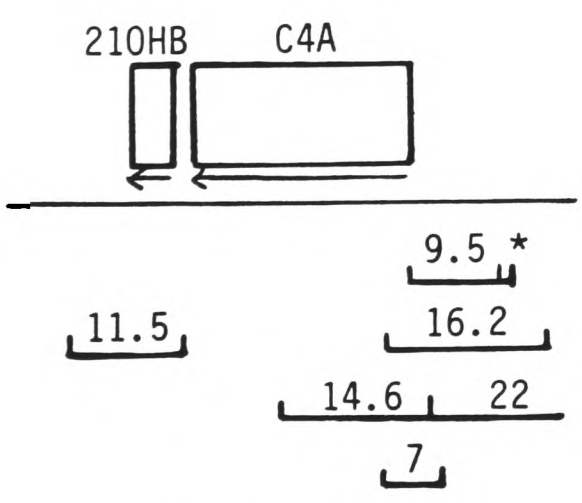
poly (A) addition site

```
GCCTGTGATATTTTCTGGATGTCCTTTATTTACTGTGACGTGTGTTTGGGTGCCTTGTTT -561

ATGGGGTAGAGGTGAAGTCTGAGCTTTGCCTCATTCAGAGAGCAAAGGGGTCAGGGGTTC -501

ACTCTGACGTTCAGGCCATTCTCCCTCGTGGAGTGGTGAGGGTGTACCTAATCTCCTAAA -441

CCACGGAATTTCTGTTAGGGCCTAAAAAAGCAAAAGCCTAGTATAGTTCAATTTGTGTTG -381

GAATGAAAGTAAGAGACAAGTGTCTTAGAAGCCTGTCATTGTTTTGTGAGGGCCTTTAAA -321

TATCCTGTACTCGTGGGCCATGTTGGGCCCTTGTACGCCCAGGTATACATGAGCTTGTGT -261

GCACCTATACCCTGATACAGATATACCTGGTAGGGGGAGGTGCTCAGGCACTGGAATGAG -201

AGGAGTTAACGGGGAAGGACAGGGTTATTTCTGGGCCAAGATTCAGAGTTTCCCATGGAC -141

ACCCAGGTGTCCGGGGTGCCCCCACAACTCTGGGCCTGAGGCCAGTTGCACTTCTTGGCT  -81

GTCACGTGGTTTCCCAGCTTAGCTGGGCTGGGGGAGGAGCAAGGTCCAGAGTCAACTCTG  -21

                    +1
                     **
CCCCGAGGCCTAGCTTGGCCAGAAGGTAGCAGACAGACAGACGGATCTAACCTCTCTTGG   40

           -19
             M  R  L  L  W  G  L  I  W  A  S  S  F  F  T  L
ATCCTCCAGCCATGAGGCTGCTCTGGGGGCTGATCTGGGCATCCAGCTTCTTCACCTTAT  100

          1
S  L  Q  K  P  R
CTCTGCAGAAGCCCAGGTCCTGGAGGCGGGATGCTGGGTGCTTGGATTGGGGCAGGGCTG  160
```

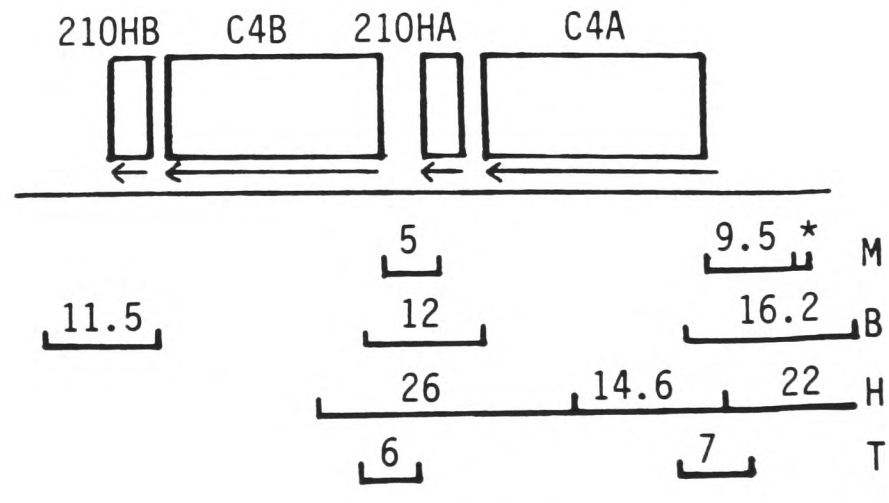
Fig. 5.16

Southern blot analysis with cDNA 2.2. DNA from two individuals was digested with Bam HI (M), Bgl II (B), Hind III (H) and Taq I (T). Individual A is the cell line used to prepare the cosmid libraries. Individual B has both C4A and C4B genes, with their associated 21OH loci. The origins of the fragments detected using the cDNA are shown schematically in the diagram underneath. The precise position of the 0.95 kb Taq I fragment is not shown as it has not been determined.





Individual A



Individual B

*= 0.8 kb Bam HI fragment

digests. These bands are equivalent to the 9.5 and 0.8 kb (Bam HI), 16.2 kb (Bgl II), 22 and 14.6 kb (Hind III), and 0.95 and 7 kb (Taq I) fragments mapped in cloned DNA to the region 5' to the C4A locus. Additional bands at 4.9 kb (Bam HI), 4.4 and 11 kb (Bgl II), >20 kb (Hind III) and 6 kb (Taq I) were visible in YB. These fragments are associated with the 5' end of the C4B locus (Fig. 5.16), and indicate that at least part of the coding sequence of G11 is involved in the duplication event which has given rise to the two C4 and 21OH loci. A faint band at ~10 kb in the Bgl II digest is also present in both haplotypes. The origin of this fragment is uncertain, as there are no equivalent faint bands in the other tracks. However, it may be equivalent to the 11.5 kb Bgl II fragment that is associated with the 21OHB gene.

5.3 DISCUSSION

Two HTF islands, designated 13 and 14, have been mapped within the 30 kb separating the factor B and C4A genes (chapter IV). One of these, island 14, has been characterised for the presence of an associated gene by hybridisation of unique sequence flanking probes to animal Southern blots, and Northern blots. Although both probes detected single copy sequences in Bam HI and Bgl II digests of genomic DNA from other mammalian species, only probe D3 hybridised to a transcript. The 1.4 kb mRNA, G11, was expressed at a relatively low level in comparison with other ubiquitous transcripts from the class III region.

Probe D3 was used to screen a cDNA library prepared from PMA stimulated U937 cells. Two independent recombinants were recovered. These were found to contain inserts of 1.2 and 1.3 kb, which differed by the presence of a Bam HI site in the larger clone. Both cDNAs were

sequenced, and two major variations were discovered: firstly, the inserts had unique 5' ends, and, secondly, they were distinguished by the presence or absence of a GpT dinucleotide within a 240 bp Nco I/ Nhe I fragment.

To resolve the origin of these differences, the distribution of the coding sequence was determined by Southern blot analysis of cos2$_\gamma$. Subclones containing the 5' genomic sequence, including the untranslated region, and the fragments encompassing the GpT dinucleotide were isolated and sequenced. The differences between the cDNA inserts were shown to arise from alternative splicing of a precursor RNA. cDNA 2.1 included part of the first intron defined in the genomic sequence, which accounted for the divergence in the 5' ends of the clones. Two 5' splice junction sites defined within a 152 bp Nco I genomic fragment explained the variation in the number of GpTs. If the first site is used, the contiguous sequence of the mRNA is identical to cDNA 2.1. If the second site is used, an extra GpT is inserted to give a contiguous sequence equivalent to that seen in cDNA 2.2.

The splice junctions obey the GT/AG rule of Breathnach and Chambon (1981). Comparisons with extended consensus sequences (Padgett et al., 1986) show close agreements for both splice junctions of the 5' intron, and the 3' splice junction of the intron with the alternative 5' splice sites. The latter are atypical as they contain a T rather than a G as the last nucleotide before the exon/ intron boundary. Putative branch sites have also been located on the basis of homology to the PyNPyTPuAPy consensus of Reed and Maniatis (1985) (Fig. 5.17).

Analysis of the nucleotide sequence of cDNA 2.2 predicts an open reading frame of 258 residues. The sequence around the initiating methionine does not conform to the consensus of Kozak (1987), nor does any other methionine codon within the genomic or cDNA subclones. The

		5' splice junction	3' splice junction
consensus		AG:GTAAGT	$^{TTTTTTTTTTT}_{CCCCCCCCCCC}$NCAG:G
5' intron		AG:GTAACC.....CCGTGAGCCGATTTATCTGCCCAG:G	
"GpT" intron:	2.1	CT:GTGTGC.....TCCCACTGCCCTCAG:G	
	2.2	GT:GTGCGT.....TCCCACTGCCCTCAG:G	
		: intron	intron :

Fig. 5.17

The splice junctions of the 5' intron, and the alternative internal splice junction responsible for the GpT dinucleotide difference between cDNA 2.1 and cDNA 2.2. The consensus sequences are as described in Padgett et al. (1986). The best match for the consensus branch point sequence PyNPyTPuAPy (Reed and Maniatis, 1985) in the small intron is underlined.

alternative splicing observed in cDNA 2.1 interrupts this putative coding region, and would result in the production of truncated polypeptides sharing regions of homology with the product of cDNA 2.2. Either cDNAs 2.1 and 2.2 encode distinct isoforms of the same polypeptide, or the splicing of the precursor mRNA is functioning as a regulatory mechanism, to control gene expression. Multiple, but related transcripts from a single gene have been observed for a number of gene families in organisms ranging from Drosophila to man (Breitbart et al., 1987). In the case of the "CAAT" box binding proteins, the alternative mRNAs encode functional isoforms (Santaro et al., 1988), but, analysis of cDNA clones from the murine tyrosinase gene show that only the full length transcript encodes an active protein (Ruppert et al., 1988; Muller et al., 1988).

Characterisation of the 5' genomic sequences and cDNA 2.2 shows that the distribution of CpG dinucleotides is consistent with that observed for other HTF island associated genes (Gardiner-Garden and Frommer, 1987). The island structure spans ~0.9 kb, and includes the first exon and part of the second exon of the putative coding sequence. The island contains a cluster of overlapping sites for the enzymes Eag I, Nar I and Sac II in the region immediately 5' to the gene, and two overlapping Bss HII sites in the second exon. Of these, the sites for Sac II, Eag I and Bss HII are known to cut at the chromosomal level (see chapter IV). Exact matches for the consensus sequences of the transcription factors Sp1 and AP2 have been found within the island. One Sp1 site occurs 165 nucleotides 3' to the initiating methionine. Similar G/C boxes have been located within the coding sequences of other island associated genes (Gardiner-Garden and Frommer, 1987).

Additional motifs include the Ig octamer sequence, and three copies of the cAMP response element (CRE). One of the latter occurs within a

palindromic octamer. The significance of these elements is unknown, but they may prove important for understanding the expression of G11, given that the levels of mRNA detected by D3 were much lower than those observed for other ubiquitous transcripts from the class III region.

Mapping of the coding sequences in cos2γ suggests that the G11 gene contains a large intron of at least 4 kb close to the 5' end. The bulk of the cDNA is encoded within a 3 kb region 5' to the C4A locus. Sequencing of genomic fragments has established that the poly A addition site of G11 is only 613 bp upstream of the C4A mRNA cap site (Belt, 1985). Furthermore, genomic Southern blot analysis of the cDNA 2.2 infers that at least part of G11 is involved in the duplication responsible for the two C4 loci in normal haplotypes. A weakly hybridising Bgl II fragment of 10-11 kb is also present in haplotypes with single and duplicated C4 genes. If this is equivalent to the 11.5 kb restriction element containing the 21OHB gene, there may be an extra region of sequence homology to G11. These may be significant in understanding the deletion and duplication events which result in chromosomes carrying different numbers of C4 and 21OH loci. Regions of homology could provide areas where misalignment can occur, resulting in the exchange of DNA in a manner analogous to that involving Alu repeats in the α and β globin loci (Nicholls et al., 1987; Henthorn et al., 1986).

Finally, although the whole of the Sal I probe detected cross-hybridising fragments on Southern blots of genomic DNA from a range of animal species, the 1.7 kb Bam HI fragment failed to detect a transcript on a Northern blot. Either

(a) the region between factor B and C4 is very highly conserved between mammals, and especially between primates as the man and monkey digests appear to contain fragments of a common size,

or,

(b) there is a third gene within the 30 kb gap which is tissue specific, or expressed at levels below the detection limits of the system used.

As the probe hybridises strongly with other mammals, a panel of RNA from different tissues could be prepared from a laboratory animal, such as mouse, for use in further analyses.

CHAPTER VI

IDENTIFICATION OF AN MHC-LINKED STRESS PROTEIN: A DUPLICATED LOCUS FOR HUMAN HSP 70

6.1 INTRODUCTION

Characterisation of the class III region of the human MHC for HTF islands and their associated transcripts revealed a duplicated sequence within the region 92 to 104 kb telomeric to the C2 gene. Two islands containing sites for Eag I, Pvu I and Sac II were positioned 12 kb apart from PFGE data and restriction enzyme mapping of the cloned DNA (chapter IV). Each island was adjacent to a copy of 0.9 kb Bgl II fragment which hybridised with probe H, a 0.8 kb Bgl II/ Sal I genomic fragment isolated from the end of cos7I (Fig. 6.1). As with other class III region probes used for island analysis, probe H was hybridised to Southern blots of genomic DNA isolated from a range of animal species. The results revealed distinctive features which were not observed with any other unique sequence. Firstly, probe H detected a family of restriction fragments with varying intensities; secondly, the probe cross-hybridised with shark, as well as the mammalian samples, under stringent washing conditions.

Few genes are so highly conserved between different classes of the animal kingdom. Of those mapped by cytogenetic techniques to the p21 region of chromosome 6, only HSP 70 has characteristics in common with probe H (Goate et al., 1987; Harrison et al., 1987). A number of human HSP 70 cDNA and bacteriophage λ genomic clones have been published (Hunt and Morimoto, 1985; Voellmy et al., 1985; Wu et al., 1985). Comparison with detailed restriction maps of the cosmid inserts suggested that the genomic clones independently isolated by Hunt and Morimoto (1985) shared identity for the enzymes Bam HI and Hind III. The presence of HSP 70 in

the human MHC was confirmed by sequence analysis of suitable subclones containing the two copies of probe H.

The mapping of two loci for HSP 70 family members within the class III region may be of interest with respect to disease association. Owing to the high degree of homology between analogues from bacteria, protozoa and man (Lindquist, 1986), presentation of HSP 70 peptides during infection could lead to the production of cross-reacting cytotoxic T lymphocytes and antibodies. Furthermore, as heat shock proteins have been attributed with a role in the protection of ribonucleoprotein (RNP) structure (Pelham, 1984, 1986; Welch and Suhan, 1986) they could contribute to the auto-immune aetiology of diseases such as SLE, where anti-RNP immunoglobulins are a common feature.

6.2 RESULTS

6.2.1 Isolation of Cosmid Clones

Probe H was isolated from the end of cos7I using a Bgl II/ Sal I double digest. It represented 0.8 kb of a 0.9 kb Bgl II restriction fragment mapped within the slightly larger insert of cos8I (Fig. 6.1). The genomic fragment was used to rescreen the Mbo I cosmid library, and 7 new recombinants were taken for analysis. The inserts were cleaved with Bam HI, Bgl II and Cla I in single and double digest combinations, and compared with cos7I and cos8I using agarose gel electrophoresis and Southern blotting. The cosmids could be divided into three groups on the basis of the restriction patterns, and hybridisation to probe H:

(a) cos9III overlapped the cos8I and cos7I cluster by 34 kb and extended it by 52 kb. All three inserts contained a single copy of probe H.

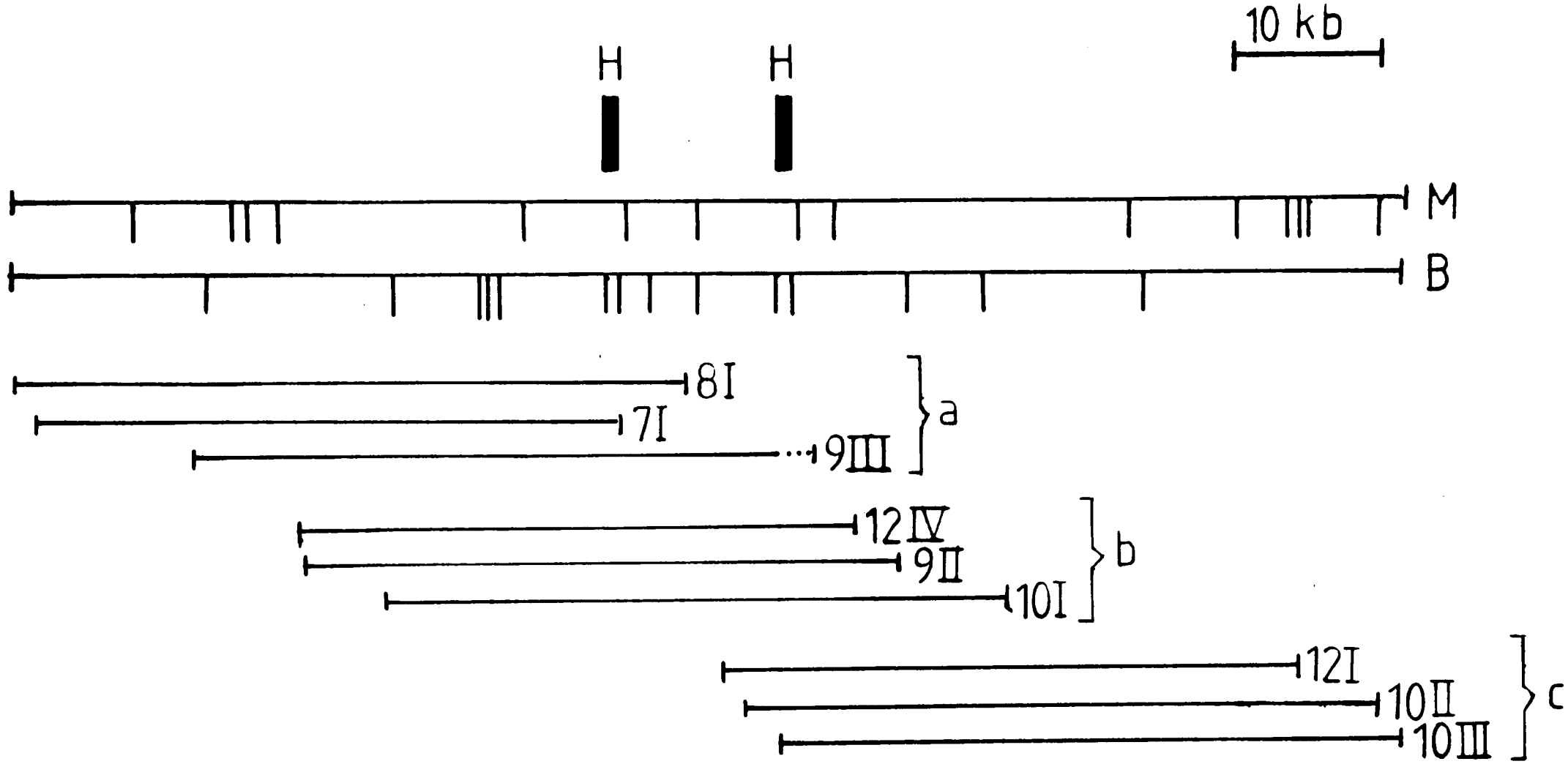


Fig. 6.1

Clones isolated with probe H. Restriction sites for Bam HI (M) and Bgl II (B) are shown by the vertical bars. The non-contiguous portion of cos9III is indicated by the broken line. The positions of the copies of probe H are shown by the solid boxes.

(b) cos10II, cos10III and cos12I also contained a single copy of probe H, plus a region which hybridised weakly. The cluster could not be overlapped with any of the cosmids in group (a).

(c) cos9II, cos12IV and cos10I had restriction fragments common to groups (a) and (b). The molecular map suggested that these cosmids contained duplicated 7 kb Bam HI fragments each of which encompassed a copy of probe H. In addition, the 3 cosmids in group (c) also had the cross-hybridising region seen in group (b).

The 3 cosmid clusters overlapped to give a total extension of 52 kb beyond the end of cos8I (Fig. 6.1). All the cosmids that were isolated from the library could be assigned to chromosome 6.

More detailed mapping of the cosmids with the enzymes Eco RV, Hind III, Kpn I, Sal I and Xho I revealed that cos9III was non-contiguous for the last 2.5 kb of the genomic insert. As the cosmids in group (c) all overlapped with cos8I and cos7I, the final map of this region could still be confirmed using the data from these genomic clones.

6.2.2 Analysis of the Duplication

Cos9II, from group (c), was chosen for further analysis of the duplicated region. The genomic insert was characterised with the standard restriction enzymes Bam HI, Bgl II, Cla I, Eco RV, Hind III, Kpn I, Sal I and Xho I, and with the infrequently cutting enzymes Bss HII, Eag I, Pvu I and Sac II. Comparison of the data showed that the 0.9 kb Bgl II fragment, from which probe H was derived, occurred within a duplicated 7 kb Bam HI fragment. Other restriction sites in common were those for Cla I, Eag I, Pvu I and Sac II. The two 7 kb Bam HI fragments were separated by 5 kb, positioning the Bgl II fragments ~11 kb apart, (Fig. 6.2).

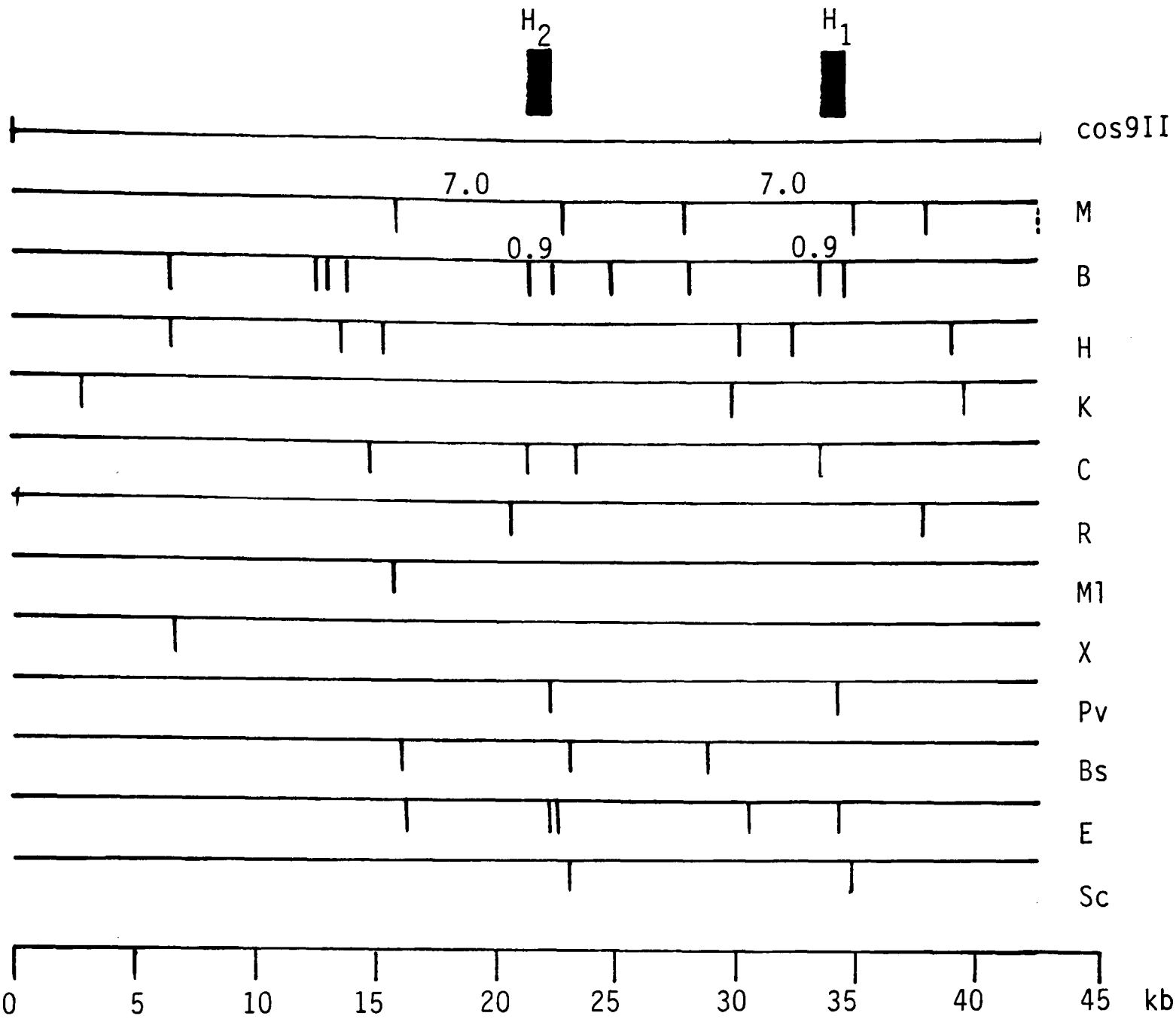


Fig. 6.2

A restriction map of cos9II. The insert was mapped using the enzymes Bam HI (M), Bgl II (B), Hind III (H), Kpn I (K), Cla I (C), Eco RI (R), Mlu I (Ml), Xho I (X), Pvu I (Pv), Bss HII (Bs), Eag I (E) and Sac II (Sc). The positions of the restriction sites are indicated by the vertical bars, and the copies of probe H by the solid boxes. The broken line represents a reformed site at the end of the genomic DNA insert. The sizes of the duplicated fragments are shown in kb.

The copies of probe H, designated H_1 and H_2, could be distinguished in double digests with Hind III, owing to the presence of a Hind III site within the more telomeric of the two Bam HI fragments. In a Hind III digest, the probe hybridised strongly to fragments of 15 kb (H_2) and 6.5 kb (H_1). In a Bam HI/ Hind III double digest, the probe hybridised strongly to fragments of 7 kb (H_2) and 2.4 kb (H_1), and weakly to a 3.7 kb fragment (the homologous region) (Fig 6.3).

6.2.3 Analysis of the Region of Homology

Southern blots of digested cos9II DNA hybridised with probe H revealed a region of homology which hybridised more weakly than the duplicated fragments. In a Bam HI/ Hind III double digest probe H detected a 3.7 kb fragment, equivalent to the end of the genomic insert. In a Bam HI/ Kpn I digest, the probe hybridised to a 1.7 kb fragment. From the restriction map of cos9II, the limits of the homology could be defined to lie within a 0.5 kb region between the Kpn I and Hind III sites (Fig. 6.3). This is located about 5 kb telomeric to the duplication.

As the homologous fragment is smaller than probe H, and does not hybridise as strongly as the duplicated regions, it is not identical in sequence. The homology may represent the remnants of a partially deleted third copy of the probe.

6.2.4 Genomic Southern Blot Analysis using Probe H

Probe H was hybridised to a Southern blot of human genomic DNA cleaved with Bam HI, Bgl II, Eco RI and Hind III in single and double digest combinations. Following autoradiography, not all of the observed fragments could be attributed to the homologous regions defined by

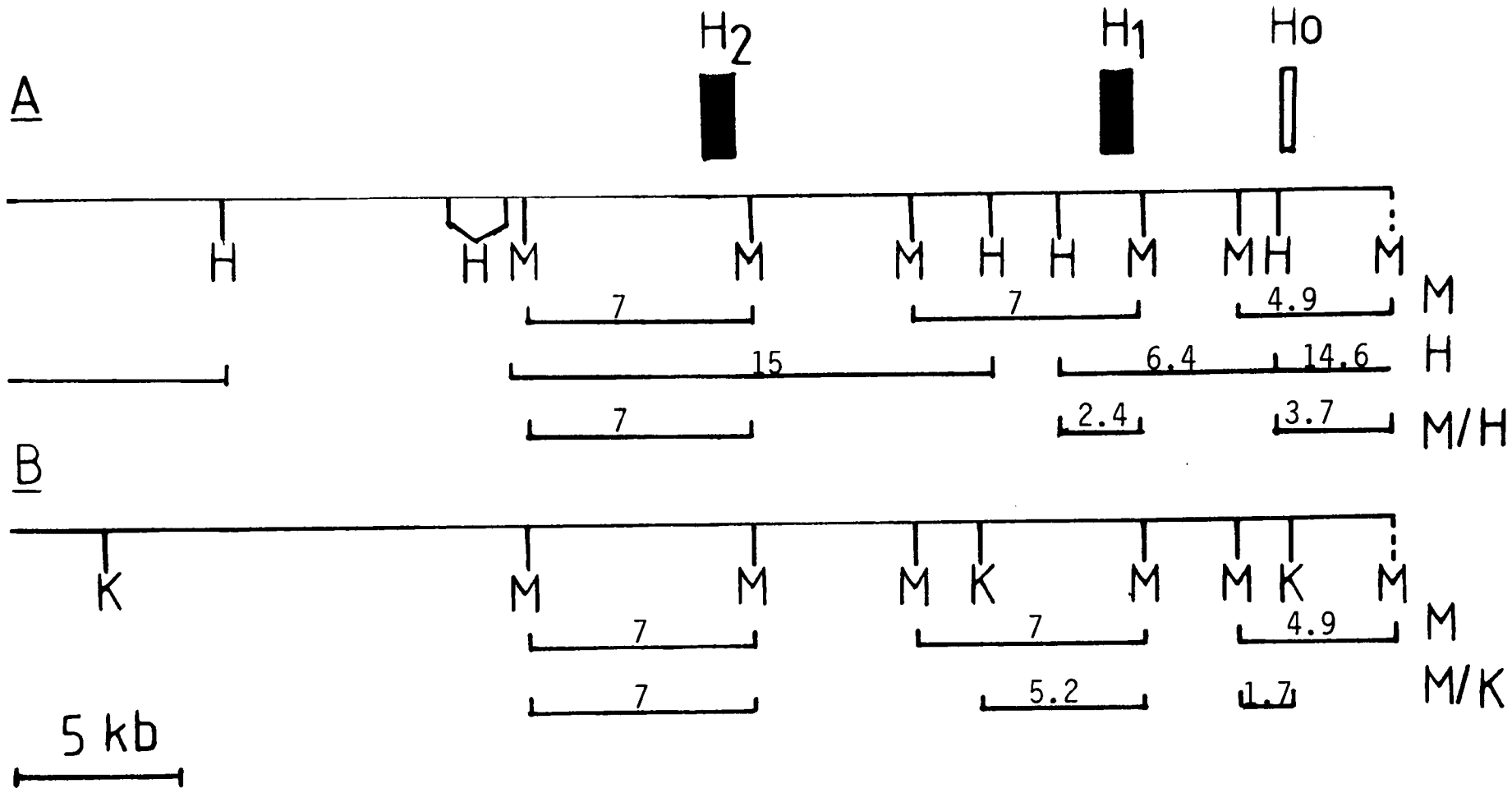


Fig. 6.3

Mapping the region of homology. Part A shows the sites for Bam HI (M) and Hind III (H), and part B shows the sites for Bam HI and Kpn I (K). The fragments hybridising to probe H are indicated by the horizontal bars. All sizes are in kb. The solid boxes indicate the positions of the copies of probe H, and the open box indicates the position of the homology (Ho).

analysis of the cosmids isolated from chromosome 6 (Table 6.1). The strongly hybridising fragments could be equated with the restriction elements mapped in the cosmid inserts. The extra fragments were not as intense as those known to contain the probe (Fig. 6.4c). The result implied that probe H belonged to a family of repetitive elements dispersed throughout the human genome at a relatively low copy number.

6.2.5 PFGE Analysis of Probe H

As with other walking probes, probe H was linked to the class III region of the MHC by PFGE (see 3.2.4). Probe H hybridised to 210 kb Not I and 640 kb Nru I fragments in common with probes I and J, which had been mapped between the C2 gene and probe H. In addition, the Pvu I sites associated with each copy of the probe were found to cut in chromosomal DNA, and probe H detected fragments of 12, 340, and 780 kb. The 340 kb fragment was in common with the complement/ 21OH gene cluster, the 780 kb fragment was in common with the TNF loci, and the 12 kb fragment represented the distance between the duplicated restriction sites.

A search of the cloned fraction of the class III region for HTF islands (see chapter IV) showed that the sites for the enzymes Bss HII, and Sac II were cleaved at a chromosomal level. The two clusters of rare enzyme sites were designated islands 8 and 9 (chapter IV, Fig. 4.9).

6.2.6 Animal Blot Analysis of Probe H

Probe H was now known to lie within a duplicated Bam HI fragment associated with HTF islands containing sites for Bss HII, Pvu I and Sac II. To determine whether the probe represented part of a gene, it was

Fig. 6.4

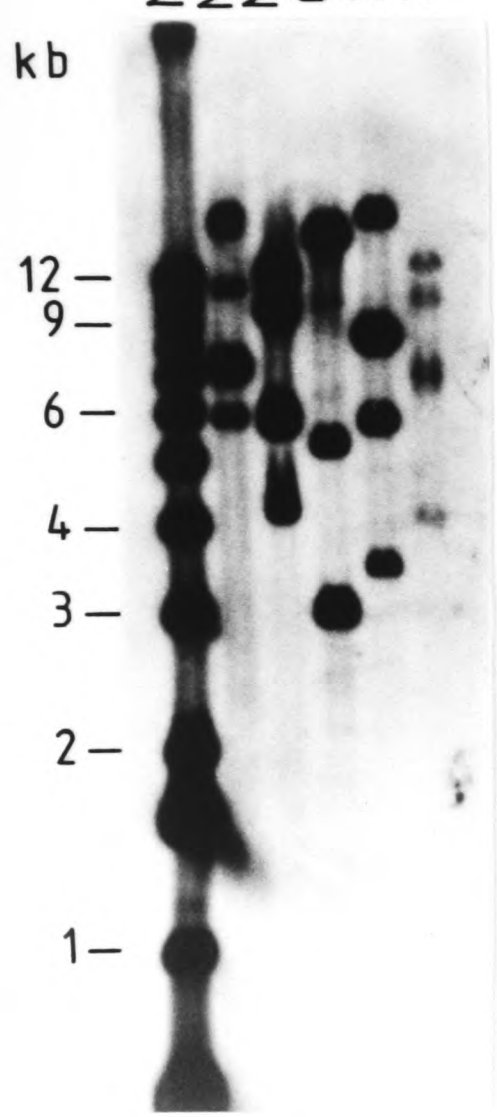
Analysis of probe H, and confirmation of the HSP 70 loci.

Part (a) shows the hybridisation pattern obtained with a Southern blot of Bam HI digested DNA from a variety of animal species. Cross-hybridising bands are evident in all the samples, including that of the shark.

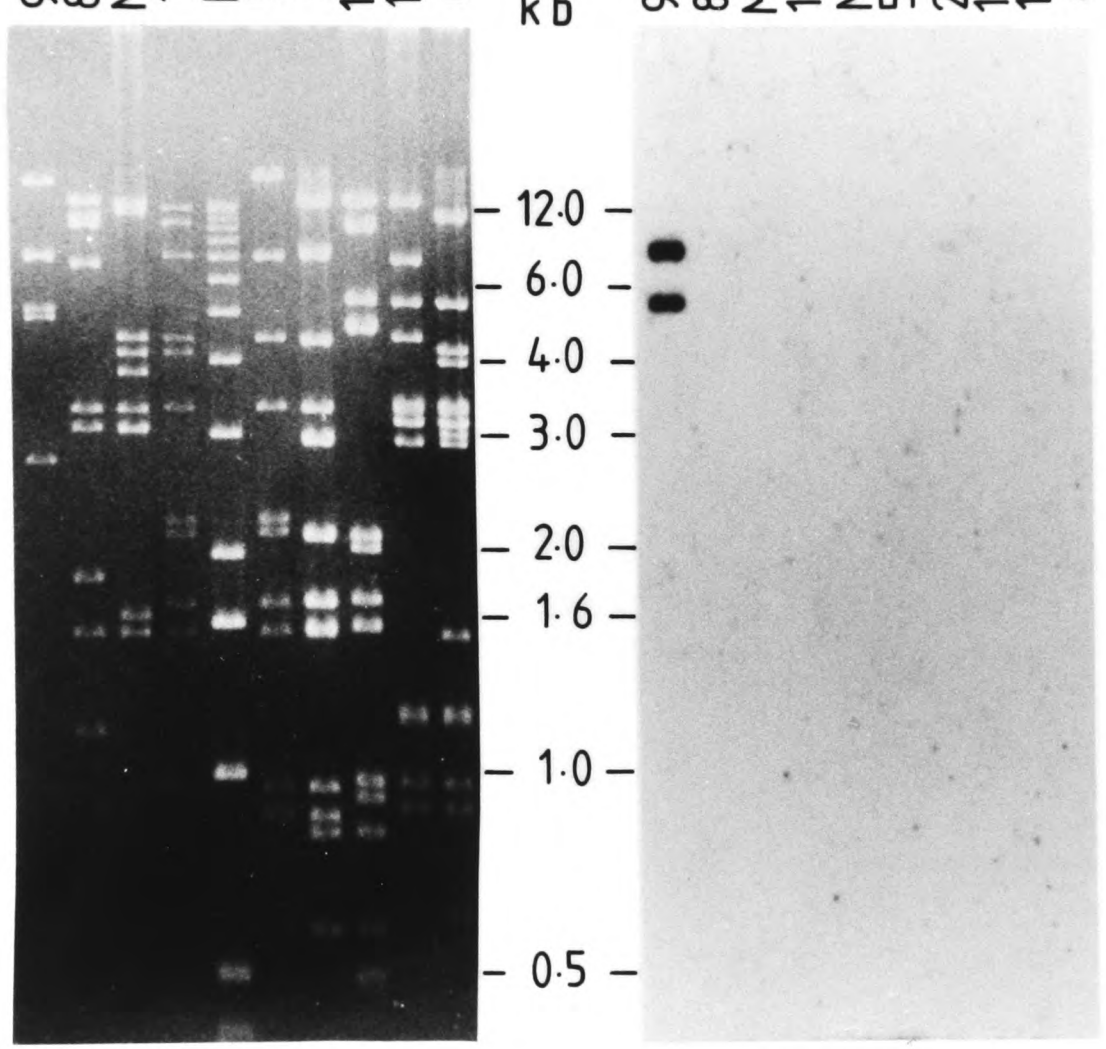
Part (b) shows the result obtained by hybridising an HSP 70 oligonucleotide to Bam HI digested cosmids spanning 300 kb of the cloned portion of the class III region. Only cos9III is detected. The fragments at 7 kb and 5 kb correspond to the Bam HI restriction elements at 7 kb and 4.9 kb which contain the duplicated 0.9 kb Bgl II fragment and the region of homology to probe H.

Part (c) shows the analysis of probe H using a Southern blot of human DNA cleaved with Bam HI (M), Eco RI (R), Bgl II (B) and Hind III (H) in single and double digest combinations. A comparison of the observed fragment sizes against the cloned fragment sizes is found in table 6.1.

a


Marker
Man
Mouse
Cat
Sheep
Shark
kb
12
9
6
4
3
2
1

b


9II
8b
M2
11α
Marker
5β
2α
11.2
12.2
6·8
kb
12·0
6·0
4·0
3·0
2·0
1·6
1·0
0·5

c

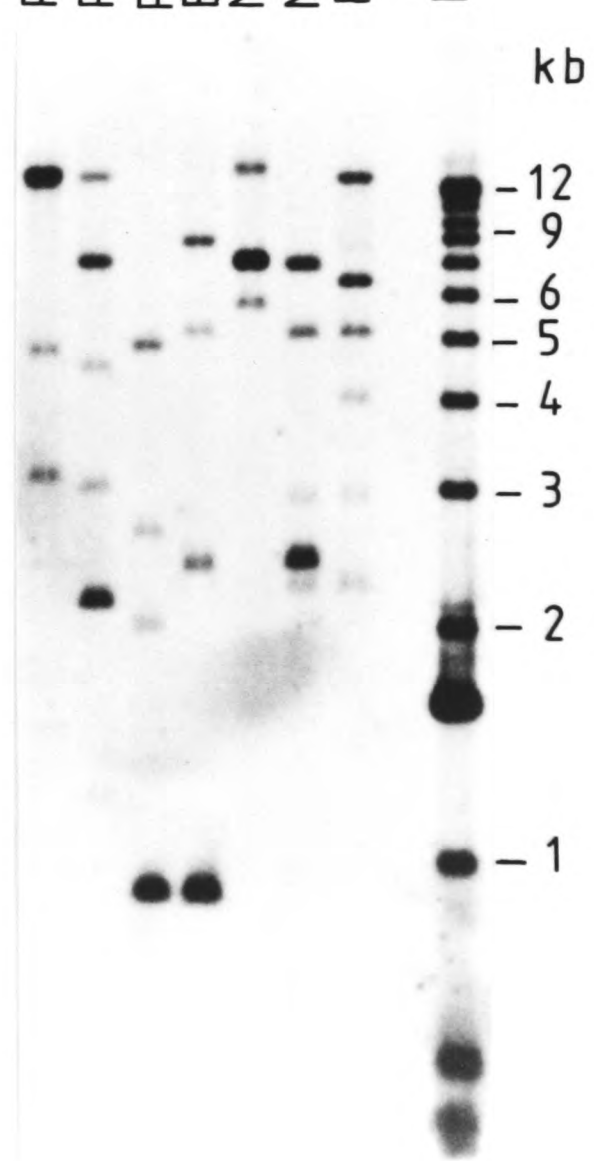

R
R+M
R+B
B
M
M+H
H
Marker
kb
12
9
6
5
4
3
2
1

Table 6.1

Comparison of the fragment sizes observed on Southern blot analysis of probe H with the restriction elements mapped in the cosmids isolated from chromosome 6.

Digest	*Fragments (kb)*	*Fragments in cloned DNA*
Bam HI	5.8	
	*7.0	7.0 (H_1, H_2)
	>15.0	20.4 (Ho)
Bam HI/ Hind III	2.2	
	*2.4	2.4 (H_1)
	3.0	
	5.1	4.9 (Ho)
	*7.0	7.0 (H_2)
Hind III	2.2	
	3.0	
	4.0	
	5.1	
	*6.5	6.4 (H_1)
	*>15.0	15.0 (H_2, Ho)
Bgl II	*0.9	0.9 (H_1, H_2)
	2.4	
	5.1	
	8.0	8.2 (Ho)
Bgl II/ Eco RI	*0.9	0.9 (H_1, H_2)
	2.0	
	2.8	3.0 (Ho)
	4.9	
Bam HI/ Eco RI	*2.1	2.2 (H_2)
	3.1	
	4.5	
	*7.0	7.0 (H_1)
	>15.0	
Eco RI	3.2	
	4.8	
	*>15.0	17.0 (H_1, H_2)

*The fragments marked with an asterisk * are the most strongly hybridising fragments on the Southern blot. H_1, H_2 and Ho define the cloned fragments containing copy 1, 2 or the region of homology to probe H, respectively.*

hybridised to a Southern blot of Bam HI digested genomic DNA samples from man, mouse, cat, sheep and shark. The blot was washed under high stringency conditions (0.2x SSC/ 0.1% SDS, 65^{o}C, 1 h) and autoradiographed. Multiple fragments were observed in all the animal tracks, including that for shark (Fig. 6.4a). The results confirmed that probe H was probably a member of a multigene family, but was unusual in the high degree of nucleotide sequence homology between different species. The intensities of the hybridising fragments in the mammalian tracks were similar to one another, and the gene under investigation had analogues in shark, an elasmobranch. Therefore, the gene family has been highly conserved throughout vertebrate evolution, and may have a fundamental role that selects against a high rate of nucleotide sequence mutation.

6.2.7 Assignment of the HSP 70 Loci

Of the genes assigned to the p21-23 region of chromosome 6 by classical cytogenetic techniques, one of the heat shock proteins, HSP 70, was the most likely candidate for the novel loci (Goate et al., 1987; Harrison et al., 1987). The heat shock proteins in general, and HSP 70 in particular, are extremely highly conserved members of multigene families (Lindquist, 1986). Protein analogues are found in organisms as diverse as bacteria, protozoa, plants and man, with 50% identity at the amino acid level between the E.coli dnaK protein and a human HSP 70 product (Hunt and Morimoto, 1985).

The novel loci were investigated using three approaches: comparison with the published bacteriophage λ clones; oligonucleotide hybridisation to class III region cosmid clones; and sequence data analysis. Each of these will be discussed in turn.

A: comparison with HSP 70 clones

A number of human HSP 70 genes and pseudogenes have been cloned in bacteriophage λ. One of these, published by Wu et al. (1985), revealed striking similarities between the sites for Bam HI and Hind III with the cos9II genomic insert. Both clones had 2.8 and 7 kb Bam HI fragments, and the 7 kb Bam HI fragment contained a 2.3 kb Hind III fragment. Analysis of the phage clone with a Drosophila HSP 70 probe defined a full length gene within a 2.5 kb Bam HI/ Hind III fragment, and a partial copy in a 1.2 kb fragment. These were positioned within the genomic DNA at sites comparible with one of the copies of probe H and its region of homology (Fig. 6.5).

The HSP 70 gene isolated from the phage insert was equivalent to that characterised by H_1. It had been used to define the chromosomal distribution of the HSP 70 family in man, using somatic cell hybrids. Three hybridising fragments of 2.5, 5.7 and 7.8 kb in a Bam HI/ Hind III genomic double digest were attributed to chromosome 6 (Harrison et al., 1987). These are probably equivalent to the 2.4, 5.0 and 7.0 kb fragments mapped within the cosmid cloned DNA (Fig. 6.5). The three regions were subsequently found to lie within an Eco RI fragment of 20 kb, which has also been mapped in the cosmid clones (Fig. 6.2).

B: oligonucleotide analysis

An oligonucleotide corresponding to nucleotides 31-56 of the published HSP 70 gene sequence (Hunt and Morimoto, 1985) was synthesised for use in further hybridisation analysis. A Southern blot of Bam HI digested inserts from cosmids spanning 300 kb cloned from the class III region was hybridised with the 26mer. Only cos9II was detected by the oligonucleotide (Fig. 6.4b). Products at 4.9 and 7 kb agreed with the restriction map constructed using probe H. Furthermore, the 4.9 kb fragment confirmed that the region of homology contained at least part

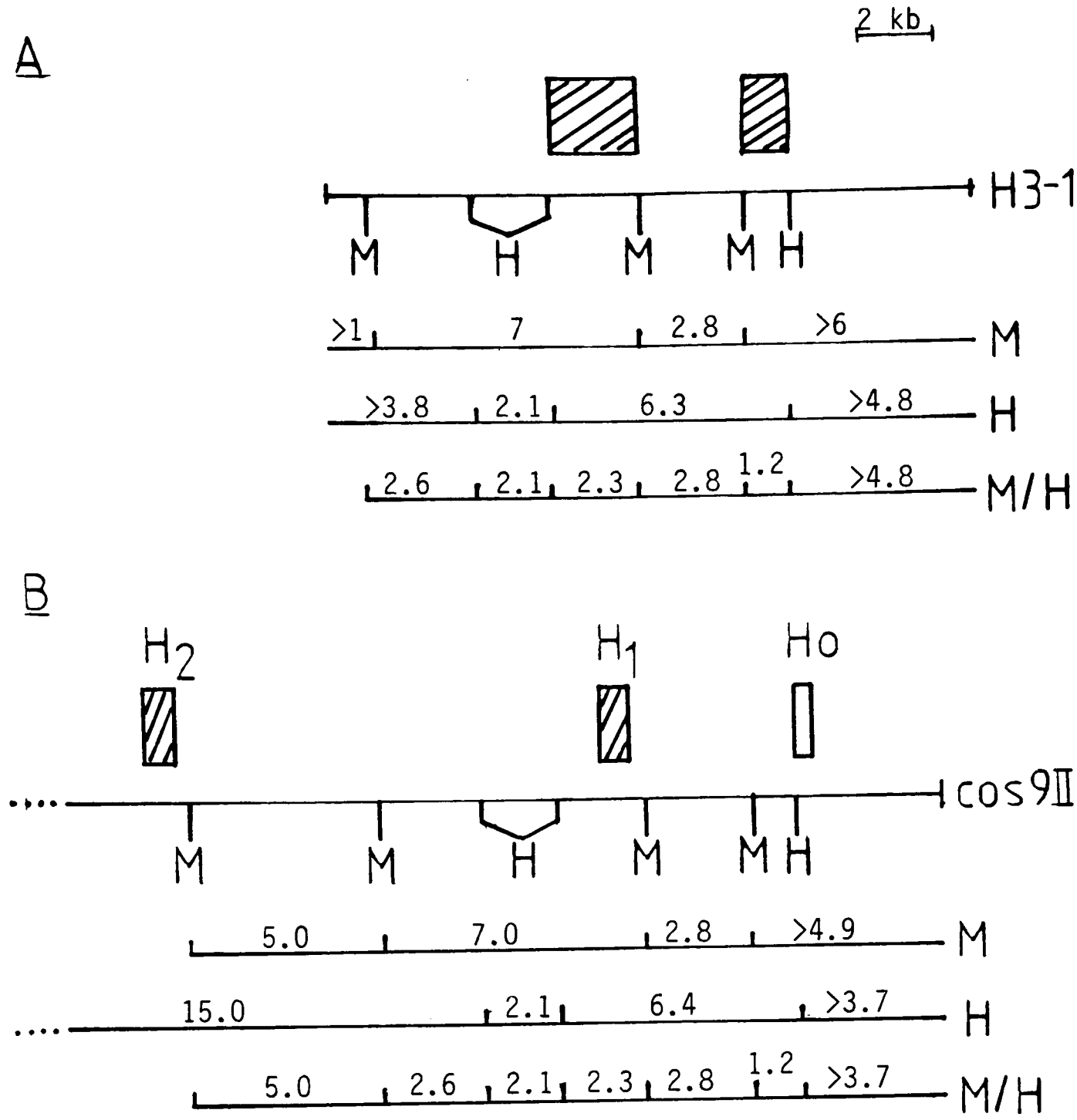


Fig. 6.5

Comparison of the restriction map of cos9II with the published bacteriophage λ clone H3-1 (Wu et al., 1985). Part A shows the genomic insert of the bacteriophage clone, with the positions of the sites for Bam HI (M) and Hind III (H). The limits of the regions hybridising to a Drosophila HSP 70 probe are indicated by the shaded boxes. Part B shows a portion of the insert of cos9II, with the positions of the copies of probe H, and the region of homology (Ho). The sizes of the single and double digest products (shown below, in kb) are remarkably similar over the region in common.

of the 5' end of the gene, including the sequence against which the oligonucleotide was made.

Additional digests using cos9II DNA with the enzymes Bam HI, Bgl II, Eco RI, Hind III, Pst I, Pvu II and Sma I were fractionated on agarose gels and blotted onto nitrocellulose for hybridisation with the HSP 70 oligonucleotide. Both copies of the duplicated sequences and the region of homology were detected. Hybridisation products at 20 kb (Eco RI), 7 kb, 4.7 kb and 2.2 kb (Eco RI/ Bam HI), 7 kb and 4.9 kb (Bam HI), 7 kb, 3.7 kb and 2.3 kb (Bam HI/ Hind III), 15 kb and 6.5 kb (Hind III), 14 kb and 3.4 kb (Pvu II), 8.5 kb, 2.9 kb and 2.7 kb (Pst I) and 3.6 kb and 2.8 kb (Sma I) proved to be consistent with those assigned to the HSP 70 genes located on chromosome 6 by other groups (Goate et al., 1987; Harrison et al., 1987; Hunt and Morimoto, 1985) (Fig. 6.6).

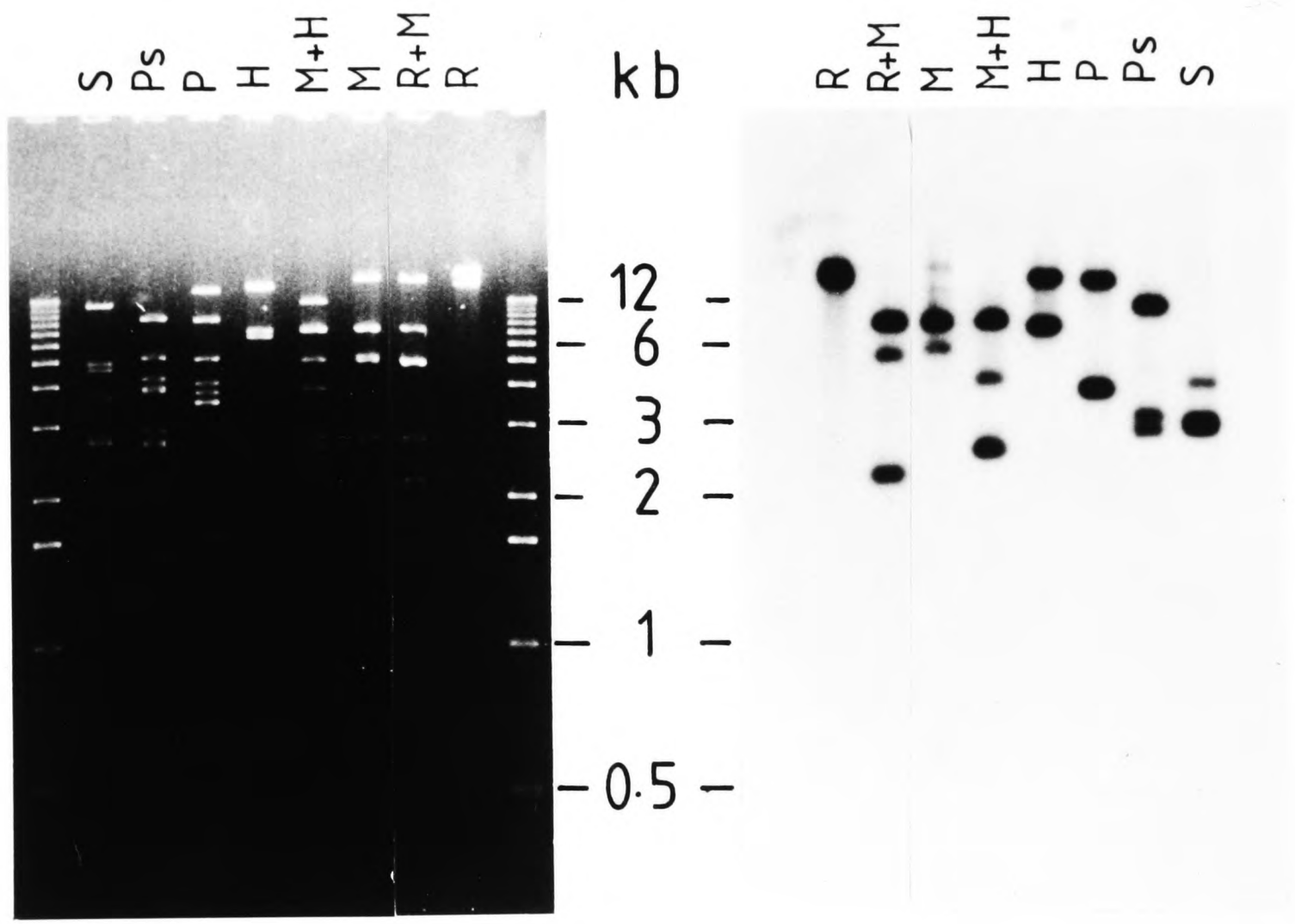
C: sequence analysis of Probe H

Final confirmation of the presence of the HSP 70 loci within the class III region came from sequence data analysis of subclones containing H_1 and H_2 (Fig. 6.7). The 5' sequence of the locus designated HSP 70-1 was determined from the Bam HI site. It was compared over 168 nucleotides with the published genomic sequence (Hunt and Morimoto, 1985), and only 3 differences were detected. Two of these occur in the 5' untranslated region, a replacement of C by G at position -26 and an extra G between positions -1 and -2 of the published data (Fig. 6.8A). The third, a replacement of G by A, lies in the first position of codon 7, and results in a Ile to Val substitution at a protein level.

In addition, both HSP 70-1 and HSP 70-2 were sequenced from the Bgl II site at nucleotide position 1280. Over 160 nucleotides were compared, and only one difference from the published sequence was observed. The replacement of G by A results in a substitution of Gly by Ala at amino acid position 278 (Fig. 6.8B).

Fig. 6.6

Digests of cos9II with the enzymes Bam HI (M), Hind III (H), Eco RI (R), Pvu II (P), Pst I (PS) and Sma I (S) in single and double digest combinations (A). Fragments hybridising with probe H (B) are consistent with the established map, and are in agreement with previously published data.


S
Ps
P
H
M+H
M
R+M
R
kb
12
6
3
2
1
0.5
R
R+M
M
M+H
H
P
Ps
S

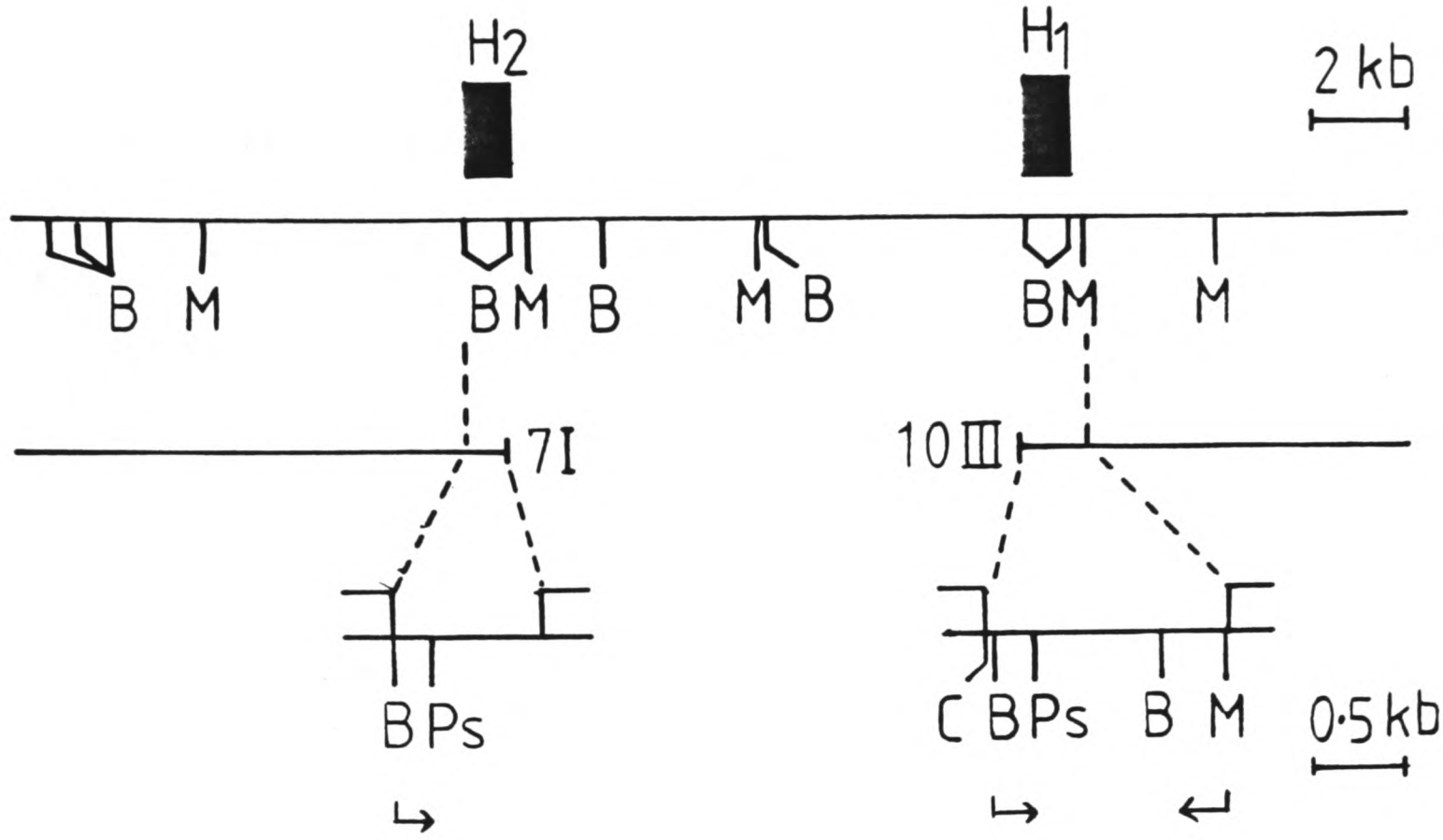
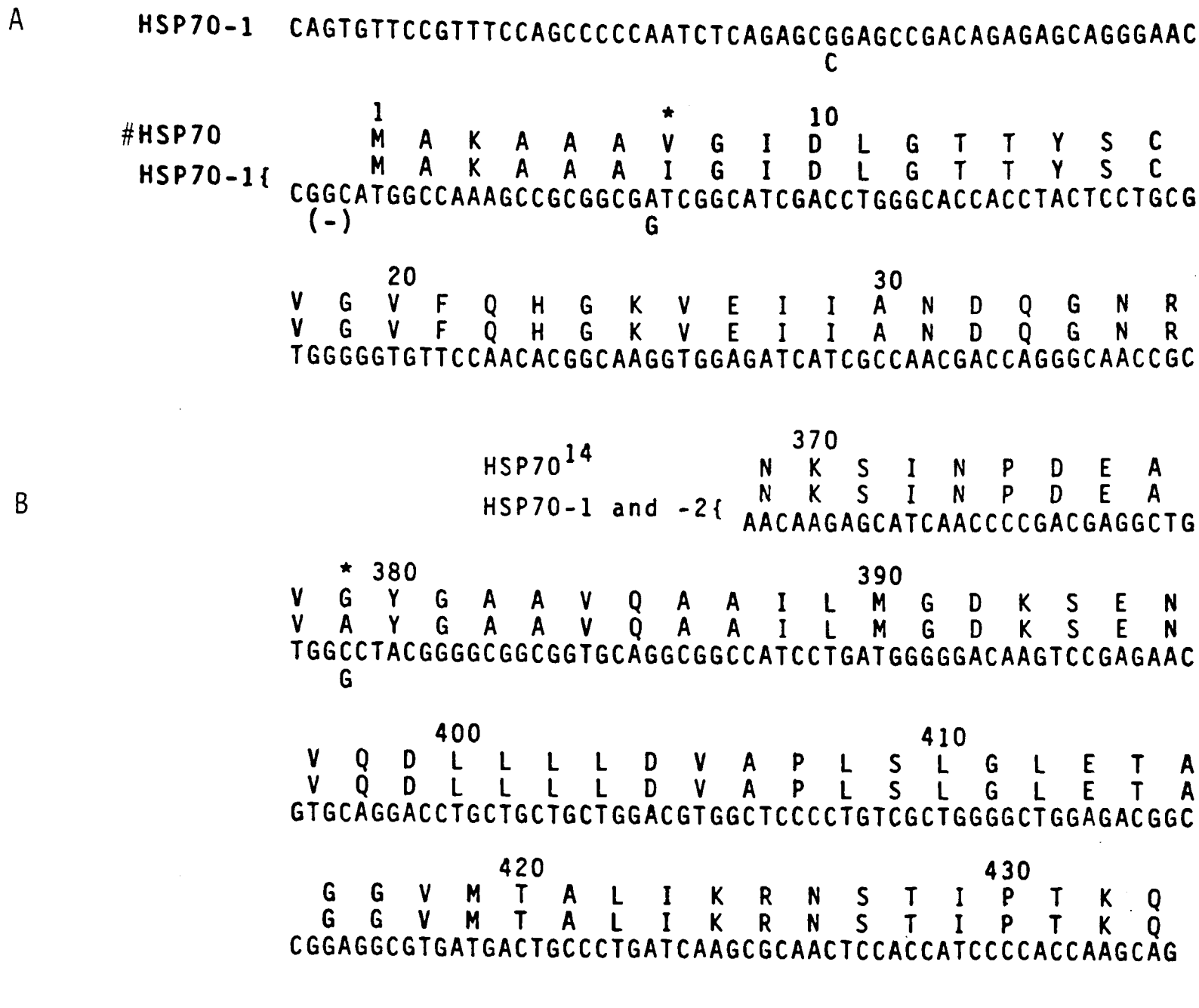


Fig. 6.7

Isolation of subclones containing H_1 or H_2. The appropriate fragments were subcloned from the ends of cosmid inserts containing either H_1 (cos10III) or H_2 (cos7I). The fragments were ligated into pATX, prior to preparing subclones for sequence analysis. Each copy was sequenced between the Bgl II (B) and Pst I (Ps) sites. In addition, the 5' end of the locus containing H_1 was sequenced from the Bam HI (M) site.



is the published sequence of Hunt and Morimoto (1985).

Fig. 6.8

Nucleotide and deduced amino acid sequences of portions of the HSP 70-1 and HSP 70-2 genes. Part A shows the nucleotide sequence of HSP 70-1 from the Bam HI site at the 5' end of the gene. Part B shows HSP 70-1 and HSP 70-2 between nucleotides 1083 and 1272. Differences between the nucleotide sequence here and that published are shown below. Differences in the derived amino acid sequences are indicated with a *. The numbers above the sequence refer to amino acid positions.

6.3 DISCUSSION

Human HSP 70 loci have been mapped to chromosomes 6, 14, 17 and 21 (Goate et al., 1987; Harrison et al., 1987). The data assimilated from the analysis of cos9II were consistent only with those genes assigned to chromosome 6. Formal linkage to the MHC class III region has been confirmed by hybridisation of probe H to PFGE blots. From the overlapping cosmid clones, locus 2 lies 92 kb telomeric to the C2 gene, and locus 1 is ~230 kb centromeric to the TNFα gene. The duplications are separated by 12 kb (Fig. 6.9).

True heat shock proteins are encoded by intronless genes (Lindquist, 1986). This allows rapid accumulation of the protein product under conditions of physiological stress, especially as mRNA processing is often inhibited. The HSP 70 gene cloned by Hunt and Morimoto (1985) is known to be functional from expression studies in human cell lines (Wu et al., 1985; Wu and Morimoto, 1985; Wu et al., 1987). It is equivalent to one of the HSP 70 genes mapped to the class III region of the MHC, defined here as HSP 70-1. This particular member of the human HSP 70 family has also been designated HSX 70 (Pelham, 1986), owing to a number of unique properties conferred by the 5' region regulatory sequences (Xiao and Lis, 1988). Unlike other mammalian HSP 70 genes, HSX 70 is expressed at a relatively high basal level, and appears to be under cell-cycle control. When fresh serum is added back to serum starved cells, levels of HSX 70 protein increase by up to ten times the basal level (Wu and Morimoto, 1985). The serum response element (SRE) shares a core sequence (5'-GGGAAA-3') in common with other growth regulated cellular promoters for the genes c-fos, IL-2 and IFN β (Wu et al., 1987). Other proteins which affect HSX 70 expression support the hypothesis that there is a link between HSX 70 and normal cell growth. These activators include the

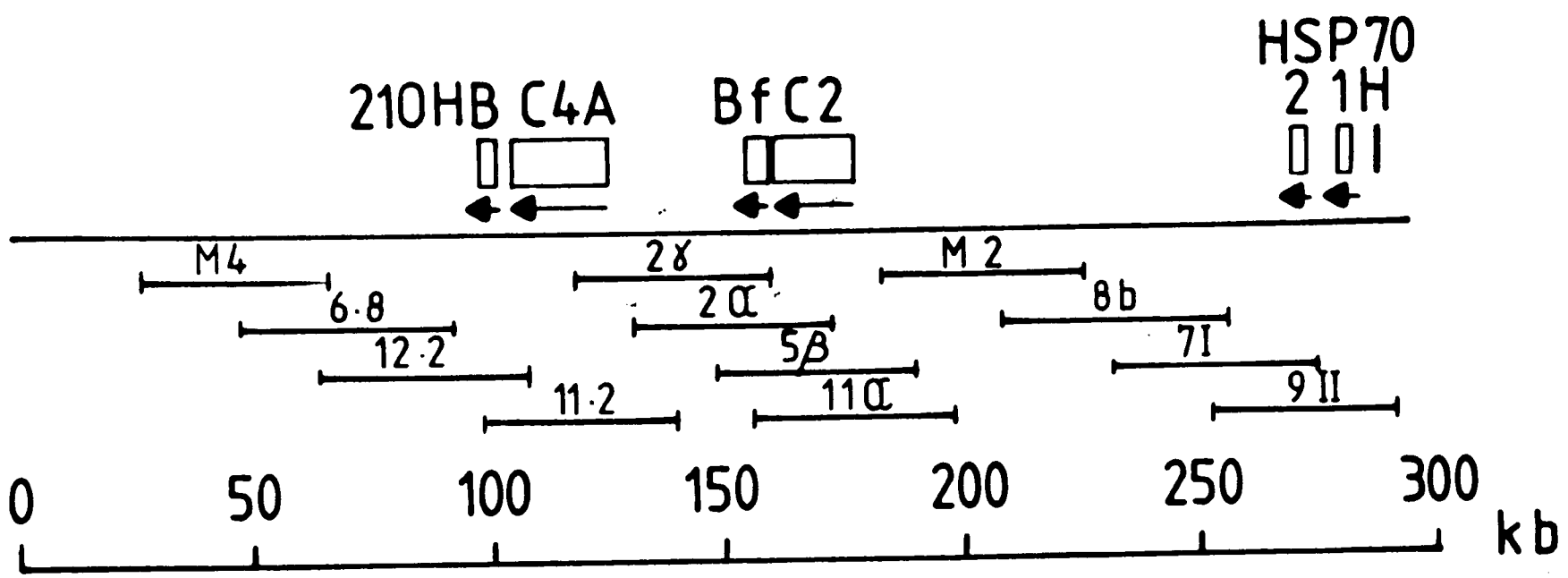


Fig. 6.9

The position of the duplicated HSP 70 loci with respect to the complement/ 21OH gene cluster. The genes are shown as open boxes, and the 5' to3' orientation by the arrows. The cosmid inserts spanning this part of the class III region are defined by the horizontal bars.

cellular proto-oncogene c-myc (Kingston et al., 1984), and, in T lymphocytes, IL-2. Here, stimulation occurs prior to transition from G_1 to S phase (Ferris et al., 1988). In addition, the SV40 large T-antigen (Khandjian and Turler, 1983) and the adenovirus 5 E1A protein (Simon et al., 1988) activate the HSX 70 gene via specific regulatory elements. The temporal pattern of expression is similar to that for histones, and appears to be essential for both DNA and RNA synthesis. It has been proposed that the heat shock proteins may be involved in the maintenance and stabilisation of multimeric complexes, including transcriptional and translational machinery (Pelham, 1986). This role could be interlinked with "chaperoning" and preribosomal assembly, described below.

Recently, members of the HSP 70 family have been attributed with a role in the intracellular translocation of proteins (Pelham, 1988; Deshaies et al., 1988; Chirico et al., 1988). In yeast, HSP 70 proteins are found to accompany secreted and organelle proteins from their site of synthesis, the ribosome, to the appropriate membrane surface. More generally, HSP 70s may act as ATP-dependent catalysts for protein folding (Pelham, 1986). If the conditions of stress which induce HSP 70 are considered; anoxia, heat shock, treatment with heavy metals or amino acid analogues, as a few examples; a common feature is the occurrence of damaged or abnormal proteins within the cell (Ananthan et al., 1986). HSP 70 members may protect these structures from forming aggregates, and promote refolding to the native form. In higher organisms, this may be important for the presentation of viral antigens in the context of class I molecules at the cell surface. A current model for the interaction of the class I antigen with immunogenic peptides suggests that it comes into contact with regions of the Golgi apparatus specialising with improperly folded proteins. Denatured fragments from this compartment bind to the class I molecule and reach the cell surface (Germain, 1986).

During recovery following a period of physiological stress HSP 70 proteins have been found to migrate to the nucleus and nucleolus (Pelham, 1984; Welch and Suhan, 1986). Nucleoli are particularly sensitive to heat shock, and the disruption of ribonucleoprotein (RNP) complexes is very evident. HSP 70 has been shown to associate with pre-ribosomal particles and other RNPs, and may play an important part in aiding reassembly, thus promoting a return to normal nuclear morphology. A common characteristic of many diseases such as systemic lupus erythematosis (SLE), mixed connective tissue disease (MCTD), rheumatoid arthritis and polymyositis, is the presence of autoantibodies directed against RNPs in the patient sera (Tan, 1982). In at least one study of SLE a high proportion of patients were found to have antibodies to HSP 70 itself (Minota and Winfield, 1988). It is possible that amino acid substitutions within a functionally important region of the HSP 70 molecule could reduce the efficiency of the RNP reassembly process. Thus, cells would fail to recover from periods of physiological stress, and cell death and lysis would liberate partially folded proteins or peptide fragments into the patient's serum. Although evidence for allelic variation at the HSP 70 loci is tentative, it would be worth studying the differences found between the sequence data obtained here and those previously published, with respect to those haplotypes frequently associated with these diseases. As there is no formal proof that HSP 70-2 encodes a functional protein, complete nucleotide sequences for both genes are required. The HSP 70 loci display a feature shared with other class III region genes, that of duplication, and there is a high degree of homology between the two loci. Therefore, there may be individuals where the number of MHC-linked HSP 70 genes differ. Although all the HSP 70 family members probably share overlapping activities, the unique interaction of HSX 70 with the cell-cycle could

have important consequences if unequal crossing over were to result in the deletion of an active gene.

Finally, as HSP 70 analogues are highly conserved within and between phyla, they may provide a particular challenge to the immune system in identifying self and non-self, leading to autoimmunity as a byproduct of infection. Immunisation of rats with Mycobacterium tuberculosis (MT) causes adjuvant arthritis, an animal model of rheumatoid arthritis in man. Cloning of one of the major epitopes recognised by T lymphocytes has shown that it is derived from a 71 kD heat shock protein (van Eden et al., 1988). Similarly, 20% of CD4+ T cell clones in mice inoculated with MT are directed against the same protein (Young et al., 1988). In certain individuals, exposure to an HSP 70 analogue could lead to long term immunity to related proteins from other infective agents. In others, it could cause tolerance, depending upon the peptide antigen presented to the immune system. Furthermore, molecular mimicry of this kind could exacerbate damage to tissues long after the initial exposure, owing to persisting cross-reactivity, for example with a local increase in the synthesis of HSP 70 during inflammation in rheumatoid arthritis (Polla, 1988).

CHAPTER VII

CONCLUSIONS

7.1 THE ORIENTATION OF THE CLASS III REGION

The order of the complement and 21OH genes has been established in genomic cosmid clones from the class III region (Carroll et al., 1984, 1985a). The orientation with respect to HLA-DR and HLA-B, however, has remained unknown, owing to the inconclusive nature of haplotype analysis and family recombination studies (Wilton and Charlton, 1986; Abbal et al., 1987). During chromosome walking from the complement/ 21OH gene cluster, new class III region probes have been characterised. These have allowed the precise orientation of the loci to be determined using PFGE (Dunham et al., 1987). The 21OHB gene is centromeric to the C2 gene, and is separated from the HLA-D region by 350 kb. The distance between the C2 gene and HLA-B is estimated at approximately 650 kb.

7.2 MAPPING THE TNF α AND TNF β LOCI TO THE CLASS III REGION

Clones for tumour necrosis factors α (cachectin) and β (lymphotoxin) have been shown to be arranged tandemly in man and mouse (Nedospasov et al., 1985, 1986; Nedwin et al., 1985; Gardner et al., 1987). They have been linked to the MHC in man by somatic cell hybridisation (Nedwin et al., 1985), analysis of chromosome 6 deletion mutants (Spies et al., 1986) and PFGE (Dunham et al., 1987; Inoko and Trowsdale, 1987; Carroll et al., 1987; Ragoussis et al., 1988), and, in the mouse by inbred strain analysis (Gardner et al., 1987; Muller et al., 1987a) and chromosome walking (Muller et al., 1987b).

Probes isolated from a cosmid cluster encompassing the TNF loci have been used to show linkage to the complement genes (Dunham et al., 1987).

By comparison with the published restriction maps of TNF genomic clones (Nedospasov et al., 1986) the human TNF α gene is shown to be centromeric to TNF β. The direction of transcription is, therefore, identical to that of the genes in the complement/ 21OH cluster.

7.3 A COMPARISON BETWEEN THE MURINE AND HUMAN MHC

The H-2 complex of the mouse is the most extensively mapped of other mammalian MHCs. A substantial proportion has been cloned in overlapping cosmids, and the general organisation has been determined using PFGE (Muller et al., 1987b). The arrangement of the genes is very similar to that established for the human HLA region (Fig. 7.1). The distances between the class II and 21OHB genes in man and mouse is 350 and 430 kb, respectively. Similarly, the C2 gene is separated from the TNF α locus by 390 and 420 kb, respectively.

The two major differences between the molecular maps occur within the class III region complement gene cluster, and in the organisation of the class I and TNF genes.

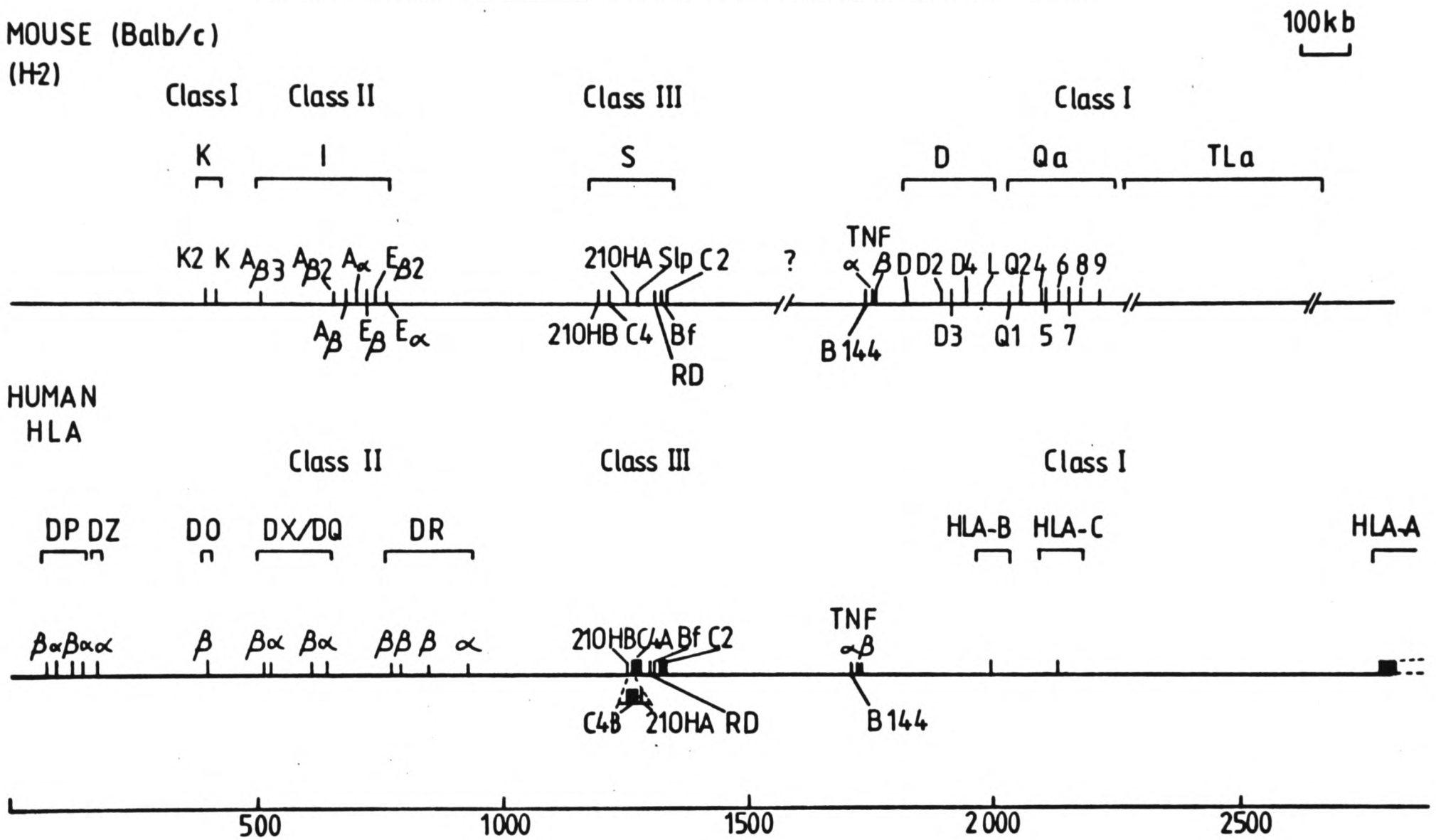
Within the S (class III) region of the H-2 complex, the first of the two C4 loci encodes a protein called the sex-limited protein (Slp). Although this shares structural homology with C4, it is not haemolytically active in complement assays, and has no known function (Chaplin, 1985). The distance between the two murine C4-like genes is 70-80 kb, indicating a much larger region of duplication than in the human MHC. Lastly, as in the HLA region, the C4 analogues are associated with genes for 21OH. However, in the murine system, the 21OHA gene is functionally active, and the 21OHB gene is the pseudogene .

In mouse, the TNF genes have been physically linked to the class I region in overlapping genomic clones (Muller et al., 1987b). They are

Fig. 7.1

A comparison between the human HLA region and the mouse H-2 complex. The genes for RD and B144 are included in this map, showing their comparable positions in the two species. It is, as yet, not known if the other loci mapped to the cloned part of the class III region in man have analogues in the mouse. (Adapted from Campbell et al., 1988.)

COMPARISON OF THE MOLECULAR MAPS OF THE HUMAN AND MURINE MHC



only 70 kb from the H-2 D locus, whereas in man the distance between the analogous genes is much greater, being in the proximity of 300 kb (Dunham et al., 1987). The apparent decrease in the DNA content of this portion of the H-2 complex may be a consequence of the chromosomal rearrangement which has positioned one of the class I loci (H-2 K) centromeric to the immune response (class II) genes (Hood et al., 1983; Dunham et al., 1987).

Telomeric to the murine H-2 complex there is a number of class I related genes called the T1 and Qa antigens. Molecular mapping has shown that the number of genes within this region varies according to the strain of mouse used in the analysis. In man, there are also between 15 and 20 class I related genes telomeric to the HLA-A locus. The function of these is unknown.

In other respects, the gene organisation of the two MHCs is highly conserved. The C2, factor B, C4 and 21OH genes are comparably arranged, and although there are more genes and pseudogenes in the human class II region, the chromosomal order of the structural analogues is the same in both species (Hardy et al., 1986). One consequence of the large number of class II genes in man is the increase in the size of the MHC with respect to the H-2 complex, 3-4 Mbp as opposed to 2 Mbp. In the mouse, a recombinational hot-spot has been located within the Eβ gene (Kobori et al., 1986; Steinmetz et al., 1986) of the I region. A similar hot-spot within the human class II region has been proposed to explain the strong linkage disequilibrium between DQ and DR loci, but not between the DP and DR subregions (Bodmer and Bodmer, 1984; Cohen et al., 1985).

7.4 A MOLECULAR MAP OF THE CLASS III REGION OF THE MHC GENERATED BY COSMID CHROMOSOME WALKING

Two cosmid clusters were isolated from the class III region of the human MHC by chromosome walking. The walks were initiated using probes from the complement/ 21OH gene cluster, and a genomic probe for TNF α. Together, they encompass 541 kb of genomic DNA. Although no linking clone could be obtained from the cosmid libraries, the gap between the clusters was estimated at 22-23 kb by PFGE analysis with flanking probes. A complete map showing all 61 cosmids, potential unique sequences and sites for the characterising restriction endonucleases can be found in Appendix 1.

Probes generated during chromosome walking in the class III region have been used to define rare enzyme sites which are also cleaved at a chromosomal level. From the most centromeric to the most telomeric endpoint of the cloned DNA, 25 single sites and clusters of sites, have been mapped. These probably represent the positions of CpG rich, unmethylated sequences called HTF islands, which frequently occur at the transcriptional start sites of housekeeping genes (Brown and Bird, 1986; Bird et al., 1985; Bird, 1986; Lindsay and Bird, 1987; Gardiner-Garden and Frommer, 1987). Putative islands between the TNF α and the 21OHB loci have been characterised for associated genes using a two step approach. Firstly, flanking probes were tested on Southern blots of genomic DNA from a variety of animal species to look for phylogenetic conservation of nucleotide sequence. Secondly, the probes were used to screen a panel of total RNA species to search for possible transcripts. Thirteen novel products were assigned to the class III region on these criteria. With the exception of the HSP 70 loci, these appear to represent transcripts from unique, single copy genes. In addition,

confirmation of the locations of the human RD and B144 genes suggests that they are situated in positions which are identical to the murine analogues (Levi-Strauss et al., 1988; Tsuge et al., 1987). The observed gene density, inclusive of 21OHB and TNF α, in this haplotype is 1/ 20 kb. The true gene density may be higher, as many tissue specific loci (including C2, Bf, C4, 21OH, TNF α and β) are not island associated.

Two of the loci mapped close to HTF islands have been identified as members of the human HSP 70 gene family by comparison with published genomic and cDNA clones (Hunt and Morimoto, 1985; Wu et al., 1985). One of the duplicated genes (HSP 70-1) is known to be functional from expression studies (Wu et al., 1985; Wu and Morimoto, 1985), but the transcriptional status of the second locus (HSP 70-2) is unknown.

Another gene (G11), situated between the factor B and C4A loci, produces alternatively spliced transcripts in PMA stimulated U937 cells. Both have been sequenced. Neither the nucleotide data, nor the polypeptides encoded by putative reading frames, have been found to match previously described gene products. G11 may also be of interest in understanding deletion and insertion events involving certain haplotypes, as a region of homology has been detected between the C4A and C4B genes by Southern blot analysis. This may provide an area where misalignment of sister chromatids can occur during meiosis, analogous to the role suggested for Alu repeats in the α and β globin clusters (Henthorn et al., 1986; Nicholls et al., 1987).

7.5 GENE DUPLICATION

Like the class I and II regions of the MHC, the class III region contains duplications which give rise to both transcriptionally active loci and to highly homologous pseudogenes. Within the complement/ 21OH

gene cluster, the C2 and factor B genes are believed to be derived from a common ancestral locus (Morley and Campbell, 1984). Duplication, followed by divergence, has produced two proteins with structural and functional homologies that perform analogous roles in the classical and alternative pathways of complement activation (Reid, 1986). Similarly, the C4/ 21OH unit is duplicated to give the C4A-21OHA-C4B-21OHB arrangement in haplotypes with the "normal" gene composition. However, the haemolytic activities of the C4 isotypes are distinct (Law et al., 1984; Isenman and Young, 1984), and the 21OHA gene is non-functional, although highly homologous to the transcriptionally active 21OHB gene at the nucleotide level (White, et al., 1986; Higashi et al., 1986; Rodrigues et al., 1987).

Four of the more recently mapped loci also reflect this theme of gene duplication: the TNF loci, and the HSP 70 loci. In addition, a partial copy of the HSP 70 gene, which includes a portion of the 5' end, has been positioned 5 kb telomeric to locus 1.

However, the remaining transcripts mapped to the cloned portion of the class III region of MHC appear to represent single-copy gene products. As high stringency conditions were used for all hybridisation experiments, homology of a similar level to that seen between the C2 and factor B genes would have been overlooked. Therefore, until the full nucleotide sequences are available, the extent of the theme ofgene duplication is unknown.

7.6 THE CLASS III REGION AS A CONSTANT BACKGROUND DOMAIN OF THE MHC

Computer analysis of the class I and class II gene sequences shows that they, too, have C+G rich domains at their 5' ends (Tykocinski and Max, 1984). It is not known how many of these sequences are methylation

free, and constitute true HTF islands. In the class I region, at least, only some of the rare enzyme sites are cleaved at a chromosomal level (Chimini et al., 1988). However, in conjunction with the results of the class III region characterisation, these findings suggest that the MHC constitutes a C+G rich chromosomal domain. Other properties of such segments of the genome have been determined. They include a high Alu content, early DNA replication in the synthetic phase of the cell cycle, and a high proportion of housekeeping to tissue specific genes (Comings, 1978; Bernardi et al., 1985; Goldman et al., 1984; Korenberg and Ryowski, 1988). The latter is consistent with the large number of HTF islands and ubiquitously expressed transcripts mapped between 21OH and TNF α.

The high gene density of the portion of the class III region cloned and described here, and in particular the large number of HTF island associated genes, has several implications for the structure of the MHC. Firstly, as most island associated loci are housekeeping genes, deletions within the class III region are likely to be deleterious. This is especially important if the loci represent single-copy genes. Therefore, the total DNA content of the class III region ought to be constant in different individuals. With the exception of the C4/ 21OH gene duplications, this appears to be confirmed from PFGE analysis of a variety of cell lines with different extended haplotypes (Ian Dunham, personal communication). Secondly, the separation between putative genes within clusters, for example those mapped to islands 1-4 centromeric to TNF α and the region between C2 and 21OHB, is very small. Thus, if any of these loci prove to be polymorphic, it is more likely that the alleles will display strong linkage disequilibrium. This is of relevance to the study of HLA disease associations, as discussed below.

7.7 THE CLASS III REGION AND DISEASE ASSOCIATION

With so many novel loci mapped to the cloned portion of the class III region of the MHC, how can this information be used to further our understanding of the HLA and disease association? In particular, can the characterisation of these newly defined transcripts help to explain the aetiologies of diseases for which the predisposing genetic loci are unknown? Obviously there are a number of other questions to be answered before the relationships with disease can be studied. First of all, the transcripts must be analysed to define the putative products, then, the functions of these polypeptides have to be determined. At this point, we have to search for further characteristics which could be important for interpreting the results, and identifying whether the class III loci are significant in disease analyses. For example, are any of the products polymorphic at the protein or nucleotide levels? If so, can specific alleles be equated with extended haplotypes which are shown to be increased in the patient population, or, do they subdivide the patients from haplotype matched controls? Alternatively, are there tissue specific variations in the levels of expression of the gene products? Some of the class III products could be structural proteins which are present in higher concentrations in some cell types than others. Levels may also be increased in response to the developmental status of the tissue, or, as a result of physiological stress. The latter could interact with the immune system during periods of infection or inflammation in such a way as to amplify the extent of tissue damage, or initiate an autoimmune response. Are there cases where inappropriate under- or over-expression of any of these products lead to a diseased state? Perhaps germline methylation of island structures could cause novel mutations which prevent transcription of the gene. Finally, are

there other tissue specific genes within the class III region? Those which exhibit a limited tissue distribution, such as B144 and G1, are expressed in cell lines of tissues known to be involved in the immune response. These may help to understand the cellular interactions which characterise normal reactions.

If any of these genes are found to be the primary genetic loci for some of the HLA related diseases, the next question to ask is, what form does the association take? For diseases of a non-immune aetiology, this may be answered already by characterising the expression and function of the transcript and its product. Defects in metabolic enzymes, receptors, and structural proteins mapped to the HLA region of chromosome 6 would be sufficient to explain linkage disequilibrium with the class I, II, and III alleles which have been used as disease markers. With autoimmune diseases, we must look for the nature of the interaction with the immune system. Is there a direct role, in the form of a control protein, lymphocyte growth factor, or, receptor molecules? Alternatively, are other mechanisms, such as molecular mimicry, involved? Precedents for the first case are the genes for TNF α and β, which have been mapped to the class III region. TNF a has been shown to enhance class I mRNA stability and cell surface expression in the presence of interferons (Collins, et al., 1986). Furthermore, it can stimulate inappropriate class II gene expression when administered in conjunction with interferon γ (Pujol-Borrell, 1987), which may suggest a role for the TNF gene products in the development of autoimmunity following virally induced infections. Indeed, an RFLP in an animal model for SLE has been shown to correlate with the TNF genes (Jacob and McDevitt, 1988). In addition, if the expression of a specific protein is stimulated during infection by viral or bacterial pathogens, cell lysis, followed by the liberation of a normally cytoplasmic protein into the patient serum

might trigger the production of an abnormal humoral response.

An alternative mechanism for the development of auto-reactive lymphocytes is that of molecular mimicry. Immunogenic peptides from an invading organism may share sufficient homology with stretches of protein sequence from a host protein to provoke the clonal expansion of cross-reactive T or B cells. Thus, presentation of the host peptides elicits an autoimmune response. This has been observed in the case of the heat shock proteins, as analogues from different species share a high degree of homology (Young et al., 1988).

At present, very few of these questions have been answered. Of the transcripts defined in this study, only two show a limited tissue distribution (B144 and G1), although others appear to display cell line specific variations in the levels of expression. Apart from the duplicated HSP 70 loci, no gene product has been identified. However, future analyses of cDNA and genomic clones should help to elucidate their functions, and may yet yield further clues to the molecular basis of some of the HLA associated diseases.

REFERENCES

Abbal, M., Thomsen, M., Cambon-Thomsen, A., Archambeau, J., Calot, M., Fathallah, D. (1985) Hum. Genet. 69, 181-183.

Abbal, M., Moennarid, C., Cambon-Thomsen, A., Tkaczuk, J., Ohayon, E., Mauff, G. (1987) Immunogenetics 26, 320-322.

Acha-Orbea, H., McDevitt, H. O. (1987) Proc. Natl. Acad. Sci. USA 84, 2435-2439.

Allen, F. H., Jr. (1974) Vox. Sang. 27, 382-384.

Alper, C. A. (1976) J. Exp. Med. 144, 1111-1115.

Alper, C. A., Awdeh, Z., Raum, D., Yunis, E. J. (1986) Biochem. Soc. Symp. 51, 19-28.

Alper, C. A., Boenisch, T., Watson, L. (1972) J. Exp. Med. 135, 68-80.

Anand, R. (1986) Trends Genet. 2. 278-283.

Ananthan, J., Goldberg, A. L., Voellmy, R. (1986) Science 232, 522-524.

Anson, D. S., Choo, K. H., Rees, D. J. G., Giannelli, F., Huddleston, J. A., Brownlee, G. G. (1984) EMBO J. 3, 1053-1060.

Auffray, C., Ben-Nun, A., Roux-Dossato, M., Germain, R. N., Seidman, J. G., Strominger, J. L. (1983) EMBO J. 2, 121-124.

Awdeh, Z. L., Alper, C. A. (1980) Proc. Natl. Acad. Sci. 77, 3576-3580.

Awdeh, Z. L., Raum, D., Yunis, E. J., Alper, C. A. (1983) Proc. Natl. Acad. Sci. USA 80, 259-263.

Batchelor. J. R., McMichael, A. J. (1987) Br. Med. Bull. 43, 156-183.

Beck, J. C., Hansen, T. H., Cullen, S. E., Lee, D. R. (1986) J. Immunol. 137, 916-923.

Bell, J., Smoot, S., Newby, C., Toyka, K., Rassenti, K., Hohfeld, R., McDevitt, H., Steinman, L. (1986) Lancet i 1058-1060.

Bell G. I., Karam, J. H., Rutter, W. J. (1981) Proc. Natl. Acad. Sci. USA 5759-5763.

Belt, K. T., Carroll, M. C., Porter, R. R. (1984) Cell 36, 907-914.

Belt, K. T. (1985) D. Phil. Thesis.

Benacerraf, B. (1981) Science 212, 1229-1238.

Bentley, D. R. (1986) Biochem. J. 239, 339-345.

Bentley, D. R., Campbell, R. D., Cross, S. J. (1985) Immunogenetics 22, 377-390.

Bentley, D. R., Campbell, R. D. (1986) Biochem. Soc. Symp. 51, 7-18.

Bentley, D. R., Porter, R. R. (1984) Proc. Natl. Acad. Sci. USA 81, 1212-1215.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny,

G., Meunier-Rotival, M., Rodier, F. (1985) Science 228, 953-957.
Bernards, R., Flavell, R. A. (1980) Nuc. Acids Res. 8, 1521-1534.
Beutler, B., Cerami, A. (1986) Nature 320, 584-588.
Biggin, M. D., Gibson, T. J., Hong, G. F. (1980) Proc. Natl. Acad. Sci. USA 80, 3963-3965.
Bird, A. P. (1986) Nature 321, 209-213.
Bird, A., Taggart, M., Frommer, M., Miller, O. J., MacLeod, D. (1985) Cell 40, 91-99.
Bird, A. P., Taggart, M. H., Nicholls, R. D., Higgs, D. R. (1987) EMBO J. 6, 999-1004.
Birnboim, H. C., Doly, J. (1979) Nuc. Acids Res. 7, 1513-1523.
Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennet, W. S., Strominger, J. L., Wiley, D. C. (1987a) Nature, 329, 506-512.
Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennet, W. S., Strominger, J. L., Wiley, D. C. (1987b) Nature, 329, 512-518.
Blue, M-L., Craig, K. A., Anderson, P., Branton. K. R. Jr., Schlossman, S. F. (1988) Cell 54, 413-421.
Blum, J. S., Cresswell, P. (1988) Proc. Natl. Acad. Sci. USA 85, 3975-3979.
Bodmer, J., Bodmer, W. (1984) Immunol. Today 5, 251-254.
Bodmer, W. F. (1980) J. Exp. Med. 152, 3535-3575.
Bodmer, W. F., Albert, E., Bodmer, J. G., Dupont, B., Mach, B., Mayr, W., Sasazuki, T., Schreuder, G. M. T., Svejgaard, A., Terasaki, P. I. Bulletin of the W.H.O.(1988) Nomenclature for factors of the HLA system, 1987.
Bodmer, W. F., Bodmer, J. G. (1978) Br. Med. Bull. 34, 309-316.
Bohme, J., Andersson, M., Andersson, G., Moller, E., Peterson, P. A., Rask, L. (1985) J. Immunol. 135, 2149-2155.
Bohme, J., Carlsson, b., Wallin, J., Moller, E., Persson, B., Peterson, P. A., Rask, L. (1986) J. Immunol. 137, 941-947.
Bourbon, H-M., Prudhomme, M., Amalric, F. (1988) Gene 68, 73-84.
Breathnach, R., Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
Breitbart, R. E., Andreadis, A., Nadal-Ginard, B. (1987) Ann. Rev. Biochem. 56, 467-495.
Brown, J. H., Jardetzky, T., Saper, M. A., Samraoui, B., Bjorkman, P. J., Wiley, D. C. (1988) Nature 332, 845-850.
Brown, W. R. A., Bird, A. P. (1986) Nature 322, 477-481.
Bushkin, Y., Demaria, S., Le, J., Schwab, R. (1988) Proc. Natl. Acad. Sci. USA 3985-3989.

Cami, B., Kourilsky, P. (1978) Nuc. Acids Res. 5, 2381-2390.

Campbell, R. D., Dodds, A. W., Porter, R. R. (1980) Biochem. J. 189, 67-80.

Campbell, R. D., Porter, R. R. (1983) Proc. Natl. Acad. Sci. USA 80, 4464-4468.

Campbell, R. D., Bentley, D. R., Morley, B. J. (1984) Phil. Trans. R. Soc. Lond. B306, 367-378.

Campbell, R. D., Carroll, M. C., Porter, R. R. (1986) Adv. Immunol. 38, 203-244.

Campbell, R. D., Law, S-K. A., Reid, K. B. M., Sim, R. B. (1988) Ann. Rev. Immunol. 6, 161-195.

Carle, G. F., Olson, M. V. (1984) Nuc. Acids Res. 12, 5647-5664.

Carroll, M. C., Campbell, R. D., Bentley, D. R., Porter, R. R. (1984) Nature 307, 237-241.

Carroll, M. C., Campbell, R. D., Porter, R. R. (1985a) Proc. Natl. Acad. Sci. USA 82, 521-525.

Carroll, M. C., Belt, K. T., Palsdottir, A., Yu, Y. (1985b) Imm. Rev. 87, 39-60.

Carroll, M. C., Palsdottir, A., Belt, K. T., Porter, R. R., (1985c) EMBO J. 4, 2847-2552.

Carroll, M. C., Alper, C. A. (1987) Br. Med. Bull. 43, 50-65.

Carroll, M. C., Katzman, P., Alicot, E. M., Koller, B. H., Geraghty, D. E., Orr, H. T., Strominger, J. L., Spies, T. (1987) Proc Natl. Acad. Sci. USA 84, 8535-8539.

Casadaban, M. J., Cohen, S. N. (1980) J. Mol. Biol. 138, 179-207.

Chaplin, D. D. (1985) Imm. Rev. 87, 61-80.

Chimini, G., Pontarotti, P., Nguyen, C., Toubert, A., Boretto, J., Jordan, B. R. (1988) EMBO J. 7, 395-400.

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., Rutter, W. J. (1979) Biochem. 18, 5294-5299.

Clarke, L., Carbon, J. (1976) Cell 9, 91-99.

Chirico, W. J., Waters, M. G., Blobel, G. (1988) Nature 332, 805-810.

Cohen, D., Le Gall, I., Marcadet, A., Font, M-P., Lalouel, J-M., Dausset, J. (1985a) Proc. Natl. Acad. Sci. USA 81, 7870-7874.

Cohen, D., Paul, P., Le Gall, I., Marcadet, A., Font, M-P., Cohen-Haguenauer, O., Sayagh, B., Cann, H., Lalouel, J-M., Dausset, J. (1985b) Immunol. Rev. 85, 98-105.

Collins, J., Hohn, B. (1978) Proc. Natl. Acad. Sci. USA 4242-4246.

Collins, T., Lapierre, L. A., Fiers, W., Strominger, J. L., Pober, J.

S. (1986) Proc. Natl. Acad. Sci. USA 83, 446-450.
Comings, D. E. (1978) Ann. Rev. Genet. 12, 25-46.
Cooper, D. N., Taggart, M. H., Bird, A. P. (1983) Nuc. Acids Res. 11, 647-658.
Cooper, D. N., Youssoufian, H. (1988) Hum. Genet. 78, 151-155.
Coulondre, C., Miller, J. H., Farabaugh, P. J., Gilbert, W. (1978) Nature 274, 775-777.
Cragg, S. J., Drysdale, J., Worwood, M. (1985) Hum. Genet. 71, 108-112.
Cresswell, P. (1987) Br. Med. Bull. 43, 66-80.
Cresswell, P., Blum, J. S., Kelner, D. N., Marks, M. S. (1987) Crit. Rev. Immunol. 7, 31-53.
Cross, S. J., Edwards, J. H., Bentley, D. R., Campbell, R. D. (1985) Immunogentics 21, 39-48.
Cuypers, H. T., Selten, G., Berns, A., van Kessel, G. A. H. M. (1986) Hum. Genet. 72, 262-265.
DeLisi, C., Berzofsky, J. A. (1985) Proc. Natl. Acad, Sci. USA 82, 7048-7052.
Deshaies, R., Koch, B. D., Werner-Washburne, M., Craig, E. A., Schekman, R. (1988) Nature 332, 800-805.
Drmanac, R., Petrovic, N., Glisin, V., Crkvenjokov, R. (1986) Nuc. Acids Res. 14, 4691-4692.
Duckworth, M. L., Gait, M. J., Goelet, P., Hong, G. F., Singh, Timas, R. C. (1981) Nuc. Acids Res. 9, 1691-1706.
Dunham, I., Sargent, C. A., Trowsdale, J., Campbell, R. D. (1987) Proc. Natl. Acad. Sci. USA 84, 7237-7241.
Duquesnoy, R. J., Trucco, M. (1988) Crit. Rev. Immunol. 8, 103-145.
Dynan, W. S., Tjian, R. (1983) Cell 32, 669-680.
Dynan, W. S., Tjian, R. (1985) Nature, 316, 774-778.
Eisenbarth, G. S. (1986) New Eng. J. Med. 314, 1360-1368.
Feinberg, A. P., Vogelstein, B. (1984) Analyt. Biochem. 137, 266-267.
Ferris, D. K., Harel-Bellan, A., Morimoto, R. I., Welch, W. J., Farrar, W. L. (1988) Proc. Natl. Acad. Sci. USA 85, 3850-3854.
Festenstein, H., Ollier, B. (1987) Br. Med. Bull. 43, 122-155.
Fielder, A. H. L., Walport, M. J., Batchelor, J. R., Rynes, R. I., Black, C. M., Dodi, I. A., Hughes, G. R. V. (1983) Br. Med. J. 186, 425-428.
Figueroa, F., Klein, D., Tewarson, S., Klein, D. (1982) J. Immunol. 129, 2089-2093.
Flaherty, L. (1981) The Role of the Major Histocompatibility Complex in

Immunobiology, pp33-57. M. Dorf, ed. (Garland STPM, N. Y.).

Flavell, R. A., Allen, H., Burkly, L. C., Sherman, D. H., Waneck, G. L., Widera, G. (1986) Science 233, 437-443.

Fleischnick, E., Raum, D., Alosco, S. M., Gerald, P. S., Yunis, E. J., Awdeh, Z. L., Granados, J., Crigler, J. F. Jr., Giles, C. M., Alper, C. A. (1983) The Lancet i 152-156.

Fourney, R. M., Miyakoshi, J., Day, R. S. III, Paterson, M. C. (1988) Focus 10, 5-7 (Bethesda Research Laboratories).

Francke, U., Pellegrino, M. A. (1977) Proc. Natl. Acad. Sci. 74, 1147

Fu, S. M., Kunkel, H. G., Brusman, H. P., Allen, F. H., Fotino, M. (1974) J. Exp. Med. 140, 1108-1111.

Gardiner-Garden, M., Frommer, M. (1987) J. Mol. Biol. 196, 261-282.

Gardner, S. M., Mock, B. A., Hilgers, J., Huppi, K. E., Roeder, W. D. (1987) 139, 476-483.

Germain, R. N., (1986) Nature 322, 687-689.

Gibson, T. J., Coulson, A. R., Sulston, J. E., Little, P. F. R. (1987) Gene 53 275-281.

Glass, D., Raum, D.,Gibson, D., Stillman, J. S., Schur, P. H. (1976) J. Clin. Invest. 58, 853-861.

Goate, A. M., Cooper, D. N., Hall, C., Leung, T. K. C., Solomon, E., Lim, L. (1987) Hum. Genet. 75, 123-128.

Goldman, M. A., Holmquist, G. P., Gray, M. C., Caston, L. A., Abhijit, N. (1984) Science 223, 686-692.

Gorer, P. A. (1937) J. Pathol. Bacteriol. 44, 691-697.

Gotze, D. ed. (1977) The Major Histocompatibility System in Man and Animals (Springer-Verlag, Berlin).

Grosveld, F. G., Dahl, H. H. M., de Boer, E., Flavell, R. A. (1981) Gene 13, 227-237,

Grosveld, F. G., Lund, T., Murray, E. J., Mellor, A. L., Dahl, H. H. M., Flavell, R. A. (1982) Nuc. Acids Res. 10, 6715-6732.

Hanahan, D., Meselson, M. (1983) Meth. Enzymol. 100, 333-342.

Handy, D. E., McClusky, J., Lew, A. M., Coligan, J. E., Margulies, D. H. (1988) Immunogenetics 28, 81-90.

Harada, F., Nishimura, Y., Suzuki, K., Matsumoto, H., Oohira, T., Matsuda, I., Sasazuki, T. (1987) Hum. Genet. 75, 91-92.

Hardy, D. A., Bell, J. I., Long, E. O., Lindsten, T., McDevitt, H. (1986) Nature 323, 435-455.

Harrison, G. S., Drabkin, H. A., Kao, F-T., Hartz, J., Hart, I. M., Chu, E. H. Y., Wu, B. J., Morimoto, R. I. (1987) Som. Cell Mol. Genet. 13,

119-130.
Henthorn, P. S., Mager, D. L., Huisman, T. H. J., Smithies, O. (1986) Proc. Natl. Acad. Sci. USA 83, 5194-5198.
Higashi, Y., Yoshioka, H., Yamane, M., Gotoh, O., Fujii-Kuriyami, Y. (1986) Proc. Natl. Acad. Sci. USA 83, 2841-2845.
Hilkens, J., Cuypers, H. T., Selten, G., Kroezen, V., Hilgers, J., Berns, A. (1986) Som. Cell Mol. Genet. 12, 81-88.
Hohn, B., Murray, K. (1977) Proc. Natl. Acad. Sci. USA 74, 3259-3263.
Hood, L., Steinmetz, M., Malissen, B. (1983) Ann. Rev, Immunol. 1, 529-568.
Hunt, C., Morimoto, R. I. (1985) Proc. Natl. Acad. Sci. USA 82, 6455-6459.
Inoko, H., Ando, A., Kimura, M., Tsuji, K. (1985) J. Immunol. 135, 2156-2159.
Inoko, H., Trowsdale, J. (1987) Nuc. Acids Res. 15, 8957-8962.
Isenman, D. E., Young, J. R. (1984) J. Immunol. 132, 3019-3027.
Ish-Horowicz, D., Burke, J. F. (1981) Nuc. Acids Res. 9, 2989-2998.
Israel, A., Kimura, A., Fournier, A., Fellous, M., Kourilsky, P. (1986) Nature 322, 743-746.
IUPAC-IUB Commission on Biochemical Nomenclature (1969) Biochem. J. 113, 1-4.
Jackson, D. A., Cook, P. R. (1985) EMBO J. 4, 913--918.
Jacob, C. O., McDevitt, H. O. (1988) Nature 331, 356-358.
Jeffreys, A. J., Flavell, R. A. (1977) Cell 12, 429-439.
Jones, N. C., Rigby, P. W. J., Ziff, E. B. (1988) Genes and Development 2, 267-281.
Josse, J., Kaiser, A. D., Kornberg, A. (1961) J. Biol. Chem. 236, 864-875.
Kappes, D. J., Strominger, J. L. (1986) Immunogenetics 24, 1-7.
Katz, D. H., Benacerraff, B. (eds.) (1976) The Role of the Histocompatibility Gene Complex in the Immune Response (Academic Press, N. Y.).
Kaufman, J. F., Auffray, C., Korman, A. J., Shakelford, D. A., Strominger, J. (1984) Cell 36, 1-13.
Kavathas, P., DeMars, R., Bach, F. H., Shaw, S. (1981) Nature 293, 747-749.
Khandjian, E. W., Turler, H. (1983) Mol. Cell. Biol. 3, 1-8.
Kingston, R., Baldwin, A. Jr., Sharp. P. (1984) Nature 312, 280-282.
Klein, J. (1975) Biology of the Mouse Histocompatibility-2 Complex

(Springer-Verlag, N. Y.)

Klein, J., Figueroa, F., Nagy, Z. A. (1983) Ann. Rev. Immunol. 1, 119-142.

Klein, J., Figueroa, F., (1986) Crit. Rev. Immunol. 6, 295-375.

Kobori, J. A., Strauss, E., Minard, K., Hood, L. (1986) Science 234, 173-179.

Koller, B. H., Geraghty, D. E., Shimizu, Y., DeMars, R., Orr, H. T. (1988) J. Immunol. 141, 897-904.

Kolsto, A-B, Kollias, G., Giguere, V., Isobe K-I, Prydz, H., Grosveld, F. (1986) Nuc. Acids Res. 14, 9667-9678.

Korenberg, J. R., Ryowski, M. C. (1988) Cell 53, 391-400.

Kozak, M. (1987) Nuc. Acids Res. 15, 8125-8148.

Krangel, M. S., Taketani, S., Biddison, W. E., Strong, D. M., Strominger, J. L. (1982) Biochemistry 21, 6313-6321.

Krangel, M. S., Biddison, W. E., Strominger, J. L. (1983) J. Immunol. 130, 1856-1862.

Lamb, J. R., Rees, A. D. M. (1988) Br. Med. Bull. 44, 600-610.

Lamm, L. U., Friedrich, U., Peterson, G. B., Jorgensen, J., Nielsen, J., Therkelsen, A. J., Kissmeyer-Nielsen, F. (1974) Hum. Hered. 24, 273-274.

Lamm, L. U., Olaisen, B. (1985) Cytogenet. Cell Genet. 40, 128-155.

Laskey, R. A., Mills, A. D. (1977) FEBS Letts. 82, 314-316.

Law, S-K. A., Dodds, A. W., Porter, R. R. (1984) EMBO J. 3, 1819-1823.

Lawrance, S. K., Smith, C. L., Srivastava, R. Cantor, C. R., Weissman, S. M. (1987) Science 2, 1387-1390.

Lee, S. J., Trowsdale, J., Travers, P. J., Carey, J., Grosveld, F., Jenkins, J., Bodmer, W. F. (1982) Nature 299, 750-752.

Lehrach, H., Diamond, D., Wozney, J. M., Boedtker, H. (1977) Biochemistry 16, 4743-4751.

Levi-Strauss, M., Carroll, M. C., Steinmetz, M., Meo, T. (1988) Science 240, 201-204.

Lindquist, S. (1986) Ann. Rev. Biochem. 55, 1151-1191.

Lindsay, S., Bird, A. P. (1987) Nature 327, 336-338.

Long, E. O., Wake, C. T., Gorski, J., Mach, B. (1983) EMBO J. 2, 389-394.

MacDonald, H. R., Schneider, R., Lees, R. K., Howe, R. C., Acha-Orbea, H., Festenstein, H., Zinkernagel, R. M., Hengartner, H. (1988) Nature 332, 40-45.

Malissen, M., Damotte, M., Birnbaum, D., Trucy, J., Jordan, B. R. (1982)

Gene 20, 485-489.

Maniatis, T., Fritsch, E. F., Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual (Cold Spring Harbour Press).

Marrack, P., Kappler, J. (1986) Ann. Rev. Immunol. 38, 1-30.

Marrack, P., Kappler, J. (1988) Nature 332, 840-843.

Marshall, W. H., Neugebauer, M., Baur, M. P., Albert, E. D. (1984) Histocompatibility Testing 1984, 313-316 (Springer-Verlag, Berlin).

Mauff, G., Hauptmann, G., Hitzeroth, H. W., Gauchel, F., Scherz, R. (1978) Z. Immunitaetsforsch. 154, 115-119.

Mauff, G., Alper, C. A., Awdeh, Z., Batchelor, J. R., Bertrams, T., Bruun-Peterson , G., Dawkins, R. L., Demant, P., Edwards, J., Grosse-Wilde, H., Hauptmann, G., Klouda, P., Lamm, L., Mollenhaure, E., Nerl, C., Olaisen, B., O'Neill, G., Rittner, C., Roos, M. H., Skanes, B., Teisberg, P., Wells, L., (1983) Immunobiology 164, 184-191.

McGill, J. R., Moore, C. M., Sakaguchi, A. Y., Boyd, D., Jain, S. K., Drysdale, J. W., Naylor, S. L. (1985) Cytogenet. Cell Genet. 40, 696-697.

Messing, J., Crea, R., Seeburg, P. H. (1981) Nuc. Acids Res. 9, 309-321.

Messing, J., Vieira, J. (1982) Gene 19, 268-275.

Meunier, H. F., Carson, S., Bodmer, W. F., Trowsdale (1986) Immunogenetics 23, 172-180.

Minota, S., Winfield, J. B. (1988) Arthrit. Rheum. 31, S13.

Monaco, A. P., Neve, R. L., Colletti-Feener, C., Bertelson, C. J., Kurnit, D. M., Kunkel, L. M. (1986) Nature 323, 646-650.

Morley, B. J., Campbell, R. D. (1984) EMBO J. 3, 153-157.

Morton, C. C., Kirsch, I. R., Nance, W. E., Evans, G. A., Korman, A. J., Strominger, J. L. (1984) Proc. Natl. Acad. Sci. USA 81, 2816-2820.

Muller, G., Ruppert, S., Schmid, E., Schutz, G. (1988) EMBO J. 7, 2723-2730.

Muller, U., Jongeneel, C. V., Nedospasov, S. A., Lindahl, K. F., Steimetz, M. (1987a) Nature 325, 265-267.

Muller, U., Stephan, D., Philippsen, P., Steinmetz, M. (1987b) EMBO J. 6, 369-373.

Nagarajan, L., Louie, E., Tsujimoto, Y., Ar-Ruhdi, A., Huebner, K., Croce, C. M. (1986) Proc. Natl. Acad. Sci. USA 83, 2556-2560.

Nedospasov, S. A., Shakov, A. N., Turetskaya, R. L., Mett, V. A., Georgiev, G. P., Dobrynin, V. N., Korobko, V. G. (1985) Dokl. Acad. Nauk. SSSR 285, 1487-1490.

Nedospasov, S. A., Shakov, A. N., Turetskaya, R. L., Mett, V. A., Azizov, M. M., Georgiev, G. P., Korobko, V. G. Dobrynin, V. N., Filipov, S. A., Bystrov, N. S., Boldyreva, E. F., Chuvpilo, S. A., Chumakov, A. M., Shingarova, L. N., Ovchinnikov, Y. A. (1986) Cold Spring Harbour Symposia on Quantitative Biology LI, 611-624.

Nedwin, G. E., Naylor, S. L., Sakaguchi, A. Y., Smith, D., Jarret-Nedwin, J., Pennica, D., Goeddel, D. V., Gray, P. W. (1985) Nuc. Acids Res. 23, 6361-6373.

Nerbert, D. N., Gonzalez, F. J. (1987) Ann. Rev. Biochem. 56, 945-993.

Nicholls, R. D., Fischel-Ghodsian, N., Higgs, D. R. (1987) Cell 49, 369-378.

Novotny, J., Auffray, C. (1984) Nuc. Acids Res. 12, 243-255.

Ogasawara, K., Maloy, W. L., Schwartz, R. H. (1987) Nature 325, 450-452.

Okada, K., Prentice, H. L., Boss, J. M., Levy, D. J., Kappes, D., Spies, T., Raghupathy, R., Mengler, R. A., Auffray, C., Strominger, J. L. (1985a) EMBO J. 4, 739-748.

Okada, K., Boss, J. M., Prentice, H., Spies, T., Mengler, R. A., Auffray, C., Lillie, J., Grossberger, D., Strominger, J. L. (1985b) Proc. Natl. Acad. Sci. USA 82, 3410-3414.

Olaisen, B., Sayaguchi, A. Y., Naylor, S. L. (1987) Cytogenet. Cell Genet. 46, 147-169.

Olaisen, B., Teisberg, P., Nordhagen, R., Michaelson, T., Gedde-Dahl, T. (1979) Nature 279, 736-737.

Olaisen, B., Teisberg, P., Jonassen, R., Gedde-Dahl, T., (1983) Ann. Hum. Genet. 47, 285-292.

Old, L. J. (1985) Science 230, 630-632.

Old, L. J. (1987) Nature 330, 602-603.

Oldstone, M. B. A. (1987) Cell 50, 819-820.

O'Neill, G. J., Yang, S. Y., Dupont, B. (1978) Proc. Natl. Acad. Sci. USA 75, 5165-5169.

Oohira, T., Nagata, N., Akabashi, I., Matsuda, I., Naito, S. (1985) Hum. Genet. 70, 341-343.

Orr, H. T., DeMars, R.(1983) Nature 302, 534-536.

Owerbach, D., Rich, C., Taneja, K. (1986) Immunogenetics 24, 41-46.

Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., Sharp, P. A. (1986) Ann. Rev. Biochem. 55, 1119-1150.

Palsdottir, A., Fossdal, R., Arnason, A., Edwards, J. H., Jensson, O. (1987a) Immunogenetics 25, 299-304.

Palsdottir, A., Arnason, A., Fossdal, R., Jenson, O. (1987b) Hum. Genet.

76, 220-224.
Parham, P., Lomen, C. E., Lawlor, D. A., Ways, J. P., Holmes, N., Coppin, H. L., Salter, R. D., Wan, A. M., Ennis, P. D. (1988) Proc. Natl. Acad. Sci. USA 85, 4005-4009.
Paul, P., Fauchet, R., Boscher, M. Y., Sayagh, B., Masset, M., Medrignac, G., Dausset, J., Cohen, D. (1987) Proc. Natl. Acad. Sci. USA 84, 2872-2876.
Pelham, H. R. B. (1984) EMBO J. 3, 3095-3100.
Pelham, H. R. B. (1986) Cell 46, 959-961.
Pelham, H. R. B. (1988) Nature 332, 776-777.
Pleogh, H. L., Orr, H. T., Strominger, J. L. (1981) Cell 24, 287-299.
Polla, B. S. (1988) Immunol. Today 9, 134-137.
Pontarotti, P. A., Mashimo, H., Zeff, R. A., Fischer, D. A., Hood, L., Mellor, A., Flavell, R. A., Nathanson, S. G. (1986) Proc. Natl. Acad. Sci. USA 83, 1782-1786.
Porter, R. R. (1983) Mol. Biol. Med. 1, 161-168.
Prochazka,M., Leiter, E. H., Serreze, D. V., Coleman, D. L. (1987) Science 237, 286-289.
Pujol-Borrell, R., Todd, I., Doshi, M., Bottazzo, G. F., Sutton, R., Gray, D., Adolf, G. R., Feldman, M. (1987) Nature 326, 304-306.
Pytela. R. (1988) EMBO J. 7, 1371-1378.
Radloff, R., Bauer, W., Vinograd, J. (1967) Biochem. 57, 1514-1521.
Ragoussis, J., Bloemer, K., Weiss, E. H., Ziegler, A. (1988) Immunogenetics 27, 66-69.
Raleigh, E. A., Wilson, G. (1986) Proc. Natl. Acad. Sci. USA 83, 9070-9074.
Reckelhoff, J. F., Bond, J. S., Beynon, R. J., Savarirayan, S., David, C. S. (1985) Immunogenetics 22, 617-623.
Reed, R., Maniatis, T. (1985) Cell 41, 95-105.
Reid, K. B. M. (1986) Essays in Biochem. 22, 27-67.
Reid, K. B. M., Bentley, D. R., Campbell, R. D., Chung, L. P., Sim, R. B., Kristensen, T., Tack, B. F. (1986) Immunol. Today 7, 230-234.
Rigby, P. W. J., Diekman, M., Rhodes, C., Berg, P. (1977) J. Mol. Biol. 113, 237-251.
Robertson, M. (1988) Nature 332, 18-19.
Robson, E. B., Lamm, L. U. (1984) Cytogenet. Cell Genet. 37, 47-70.
Rodrigues, N. R., Dunham, I., Yu, C-Y., Carroll, M. C., Porter, R. R., Campbell, R. D. (1987) EMBO J. 6, 1653-1661.
Rosenberg, S. M. (1985) Gene 39, 313-315.

Rosenberg, S. M. (1987) Meth. Enzymol. 153, 95-103.

Rosenberg, S. M., Stahl, M. M., Kobayashi, I., Stahl, F. W. (1985) Gene 38, 165-175.

Ruppert, S., Muller, G., Kwon, B., Shutz, G. (1988) EMBO J. 7, 2715-2722.

Samollow, P. B., Ford, A. L., Kunz, H. W., Gill, T. J. III (1987) Immunogenetics 26, 188-189.

Sanger, F., Nicklen, S., Coulson, A. R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.

Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H., Roe, B. A. (1980) J. Mol. Biol. 143, 161-178.

Santoro, C., Mermod, N., Andrews, P. C., Tjian, R. (1988) Nature 334, 218-224.

Scalenghe, F., Turco, E., Edstrom, J. E., Pirrotta, V., Melli, M. (1981) Chromosoma 32, 205-216.

Schenning, L., Larhammar, D., Bill, P., Wiman, K., Jonsson, A K., Rask, L., Peterson, P. A. (1984) EMBO J. 3, 447-452.

Schneider, P. M., Carrol, M. C. , Alper, C. A., Rittner, C., Whitehead, A. S., Yunis, E. J., Colten, H. R. (1986) J. Clin. Invest. 78, 650-657.

Schwartz, D. C., Cantor, C. R. (1984) Cell 37, 67-75.

Sealey, P. G., Whittaker, P. A., Southern, E. M., (1985) Nuc, Acids Res. 13, 1905-1922.

Seed, B., Parker, R. C., Davidson, N. (1982) Gene 19, 201-209.

Serfling, E., Jasin, M., Schaffner, W. (1985) Tr. Genet. 1, 224-231.

Shaw, S., Kavathas, P., Pollack, M. S., Charmot, D., Mawas, C. (1981) Nature 293, 745-747.

Shevach, E. M., Rosenthal, A. S. (1973) J. Exp. Med. 138, 1213-1229.

Sim, E., Gill, E. W., Sim, R. B. (1984) Lancet ii, 422-424.

Simon, M. C., Fisch, T. M., Benecke, B. J., Nevins, J. R., Heintz, N. (1988) Cell 52, 723-729.

Smith, D. I., Golembieski, W., Gilbert, J. D., Kizyma, L., Miller, O. J. (1987) Nuc, Acids Res. 15, 1178-1184.

Smith, H. O., Nathans, D. (1973) J. Mol. Biol. 81, 419-423.

Snyder, D. S., Beller, D. I., Unanue, E. R. (1982) Nature 299, 163-165.

Sorrentino, R., Lillie, J., Strominger, J. L. (1985) Proc. Natl. Acad. Sci. USA 82, 3794-3798.

Southern, E. M. (1975) J. Mol. Biol. 98, 503-517.

Southern, E., M., Anand, R., Brown, W. R. A., Fletcher, D. S. (1987)

Nuc. Acids Res. 15, 5925-5943.

Spielman, R. S., Lee, J., Bodmer, W. F., Bodmer, J. G., Trowsdale (1984) Proc. Natl. Acad. Sci. USA 81, 3461-3465.

Spies, T., Sorrentino, R., Boss, J. M., Okada, K., Strominger, J. L. (1985) Proc. Natl. Acad. Sci. USA 82, 5165-5169.

Spies, T., Morton, C. C., Nedospasov, S. A., Fiers, W., Pious, D., Strominger, J. L. (1986) Proc. Natl. Acad. Sci. USA 83, 8699-8702.

Srivastava, R., Chorney, M. J., Lawrance, S. K., Pan, J., Smith, Z., Smith, C. L., Weissman S. M. (1987) Proc. Natl. Acad. Sci. USA 84, 4224-4228.

Srivastava, R., Duceman, B. W., Biro, P. A., Ashwani, K. S., Weissman, S. M. (1985) Immunol. Rev. 84, 93-119.

Staden, R. (1982a) Nuc. Acids Res. 10, 4731-4751.

Staden, R. (1982b) Nuc. Acids Res. 10, 2951-2961.

Staden, R. (1984) Nuc. Acids Res. 12, 521-538.

Steinmetz, M., Hood, L. (1983) Science 222, 727-733.

Steinmetz, M., Stephan, D., Dastoornikoo, G. R., Gibb, E. Romaniuk, R. (1986) Immunological Methods vol. 3, 1-19 (Academic Press, N. Y.) (Lefkovitz, I., Pernis, B., eds.).

Steinmetz, M., Stephan, D., Lindahl, K. F. (1986a) Cell 44, 895-904.

Strachan, T. (1987) Brit. Med. Bull. 43, 1-14.

Strominger, J. L. (1987) Brit. Med. Bull. 43, 81-93.

Stroynowski, I., Clark, S., Henderson, L. E., Hood, L., McMillan, M., Forman, J. (1985) J. Immunol 135, 2160-2166.

Sugarman, B. J., Aggarwal. B. B., Hass, P. E., Figari, I. S., Palladino, M. A. Jr., Shepard, H. M. (1985) Science 230, 943-945.

Tada, N., Tanigaki, N., Pressman, D. (1978) J. Immunol. 120, 513-519.

Takata, Y., Kinoshita, T., Kozono, H., Takeda, J., Tanaka, E., Hong, K., Inoue, K. (1987) J. Exp. Med. 165, 1494-1507.

Tan E. M. (1982) Adv. Immunol. 33, 167-240.

Teng, Y. S., Tan, S. G. (1982) Hum. Hered. 32, 362-366.

Thomas, P. S. (1980) Proc. Natl. Acad. Sci. USA 77, 5201-5205.

Tiwari, J. L., Terasaki, P. I. (1985) HLA and Disease Associations (Springer-Verlag, Berlin).

Todd, J. A., Acha-Orbea, H., Bell, J. I., Chao, N., Fronek, Z., Jacob, C. O., McDermott, M., Sinha, A. A., Timmerman, L., Steiman, L., McDevitt, H. O. (1988a) Science 240, 1003-1009.

Todd, J. A., Bell, J. I., McDevitt, H. O. (1988b) Tr. Genet. 4, 129-134.

Toniolo, D., D'Urso, M., Martini, G., Persico, M., Tufano, V.,

Battistuzzi, G., Luzzatto, L. (1984) EMBO J. 3, 1987-1995.

Tonnelle, C., DeMars, R., Long, E. O. (1985) EMBO J. 4, 2839-2847.

Townsend, A. R. M., Rothbard, J., Gotch, F. M., Bahadur, G., Wraith, D., McMichael, A. J. (1986) Cell 44, 959-968.

Trowsdale, J., Campbell, R. D. (1988) Immunol. Today 9, 34-35.

Trowsdale, J., Lee, J., Carey, J., Grosveld, F., Bodmer, J., Bodmer, W. F. (1983) Proc. Natl. Acad. Sci. USA 80, 1972-1976.

Trowsdale, J., Kelly, A., Lee, J., Carson, S., Austin, P., Travers, P. (1984) Cell 38, 241-249.

Trowsdale, J., Kelly, A. (1985) EMBO J. 4, 2231-2237.

Trowsdale, J., Young, J. A. T., Kelly, A. P., Austin, P. J., Carson, S., Meunier, H., So, A., Erlich, H. A., Spielman, R. S., Bodmer, J., Bodmer, W. F. (1985) Immunol. Rev. 85, 5-43.

Trowsdale, J. (1987) Brit. Med. Bull. 15-36.

Tsuge, I., Shen, F-W., Steinmetz, M., Boyse, E. A. (1987) Immunogenetics 26, 378-380.

Tykocinski, M. L., Max, E. C. (1984) Nuc. Acids Res. 12, 4385-4396.

Unanue, E. R., Allen, P. M. (1987) Science 236, 551-557.

van Eden, W., Thole, J. E. R., van der Zee, R., Noordzij, A., van Embden, J. D. A., Hensen, E. J., Cohen, I. R. (1988) Nature 331, 171-173.

van Leeuwen, A., Festenstein, H., van Rood, J. J. (1980) J. Exp. Med. 152, 23S.

van Ommen, G. J. B., Verkerk, J. M. H. (1986) Human Genetic Disease, A Practical Approach (K. E. Davies, ed.) (IRL, Oxford).

Vega, M. A., Bragado, R., Ezquerra, A., Lopez de Castro, J. A. (1984) Biochemistry 23, 823-831.

Vieira, J., Messing, J. (1982) Gene 19, 259-268.

Voellmy, R., Ahmed, A., Schiller, P., Bromley, P., Rungger, D. (1985) Proc. Natl. Acad. Sci. USA 82, 4949-4953.

Watts, T. H., Gariepy, J., Schoolnik, G. K., McConnell, H. M. (1985) Proc. Natl. Acad. Sci. USA 82, 5480-5484.

Weiss, E. H., Golden, L., Zakut, R., Mellor, A., Fahrner, E., Kvist, S., Flavell, R. A. (1983) EMBO J. 2, 453-462.

Weiss, E. H., Golden, L., Fahrner, K., Mellor, A. L., Devlin, J. J., Bullman, H., Tiddens, H., Bud, H., Flavell, R. A. (1984) Nature 310, 650-655.

Weitkamp, L. R., Lamm, L. U. (1982) Cytogenet. Cell Genet. 32, 130-143.

Welch, W. J., Suhan, J. P. (1986) J. Cell Biol. 103, 2035-2052.

White, P. C., Grossberger, D., Onufer, B. J., Chaplin, D. D., New, M. I., Dupont, B., Strominger, J. L. (1985) Proc. Natl. Acad. Sci. USA 82, 1089-1093.

White, P. C., New, M. I., Dupont, B. (1986) Proc. Natl. Acad. Sci. USA 83, 5111-5115.

Widera, G., Flavell, R. A. (1985) Proc. Natl. Acad. Sci. USA 82, 5500-5504.

Williams, T., Yon, J., Huxley, C., Fried, M. (1988) Proc. Natl. Acad. Sci. USA 85, 3527-3530.

Wilton, A. N., Charlton,B. (1986) Immunogenetics 24, 79-83.

Winota, A., Steinmetz, M., Hood, L. (1983) Proc. Natl. Acad. Sci. USA 80, 3425-3429.

Wolf, S. F., Dintzis, S., Toniolo, D., Persico, G., Lunnen, K. D., Axelman, J., Migeon, B. R. (1984a) Nuc. Acids Res. 12 9333-9348.

Wolf. S. F., Jolly, D. J., Lunmen, K. D., Friedman, T., Migeon, B. R. (1984b) Proc. Natl. Acad. Sci. USA 81, 2806-2810.

Woods, D. E., Markham, A. F., Ricker, A. T., Goldberger, G., Colten, H. (1982) Proc. Natl. Acad. Sci. USA 79, 5661-5665.

World Health Organisation (1968) W.H.O. Bull. 39, 935-938.

World Health Organisation (1981) W.H.O. Bull. 59, 331-342.

Wu, B., Hunt, C., Morimoto, R. (1985) Mol. Cell Biol. 5, 330-341.

Wu, B. J., Morimoto, R. I. (1985) Proc. Natl. Acad. Sci. USA 82, 6070-6074.

Wu, B. J., Williams, G. T., Morimoto, R. I. (1987) Proc. Natl. Acad. Sci. USA 84, 2203-2207.

Wu, L-C., Morley, B. J., Campbell, R. D. (1987) Cell 48, 331-342.

Wu, R., Taylor, E. (1971) J. Mol. Biol. 57, 491-511.

Wyman, A. R., Wertman, K. F. (1987) Meth. Enzymol. 152, 173-180.

Xiao, H., Lis, J. T. (1988) Science 239, 1139-1142.

Young, D., Lathigra, R., Hendrix, R., Sweetser, D., Young, R. A. (1988) Proc. Natl. Acad. Sci. USA 85, 4267-4270.

Young, J. A. T., Trowsdale, J. (1985) Nuc. Acids Res. 13, 8883-8891.

Yu, C. Y., Belt, K. T., Giles, G. M., Campbell, R. D., Porter, R. R. (1986) EMBO J. 5, 2873-2881.

Yu, C. Y., Campbell, R. D. (1987) Immunogenetics 25, 383-390.

Yu, C. Y., Campbell, R. D., Porter, R. R. (1988) Immunogenetics 27, 399-405.

Yunis, E. J., Awdeh, Z., Johnson, A., Suciu-Foca, N., Robinson, M. A., Hartzman, R., Raum, D., Fleischnick, E., Alper, C. A. (1985)

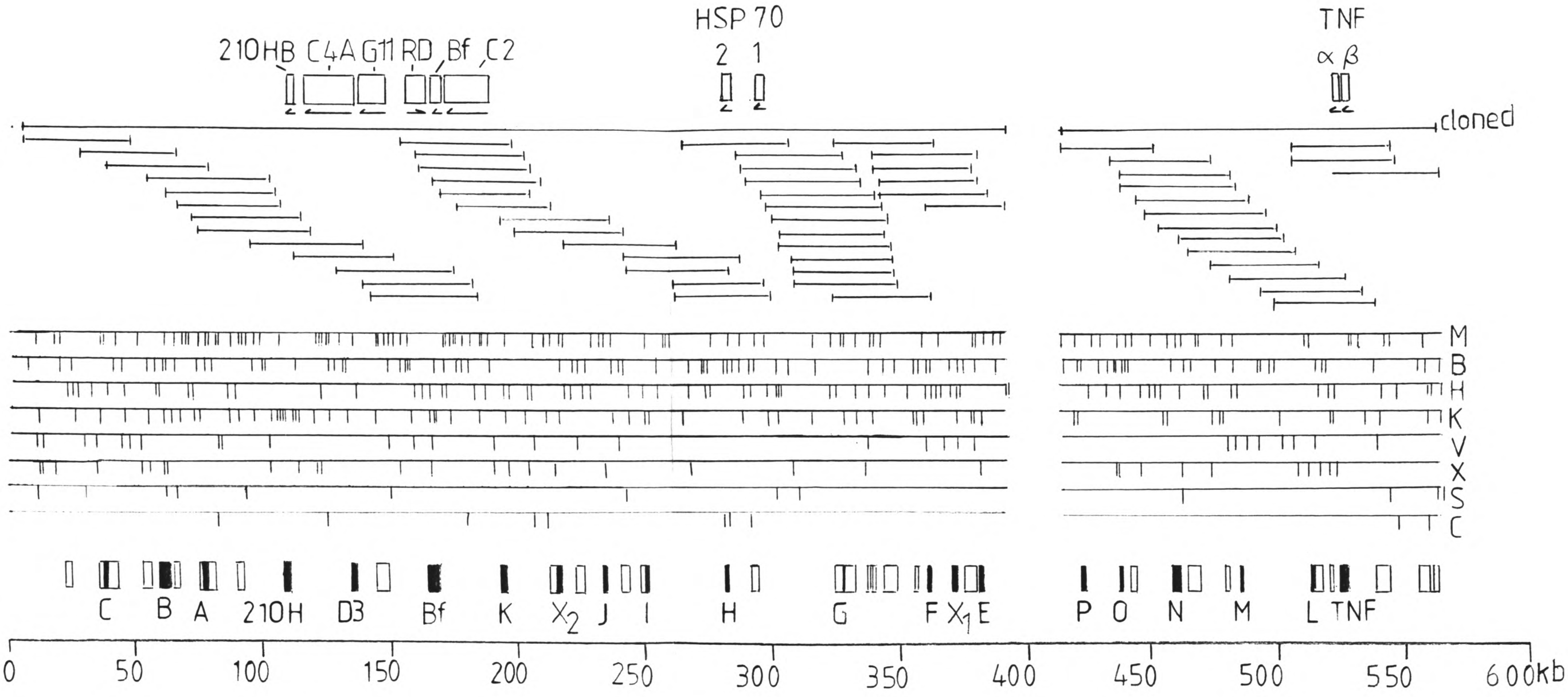
Immunogenetics 21, 25-31.
Zeff, R. A., Gopas, J., Steinhauer, E., Rajan, T. V., Nathenson, S. G. (1986) J. Immunol. 137, 897-903.
Zinkernagel, R. M., Doherty, P. C. (1974) Nature 248, 701-704.
Zinkernagel, R. M., Doherty, P. C. (1979) Adv. Immunol. 27, 51-177.

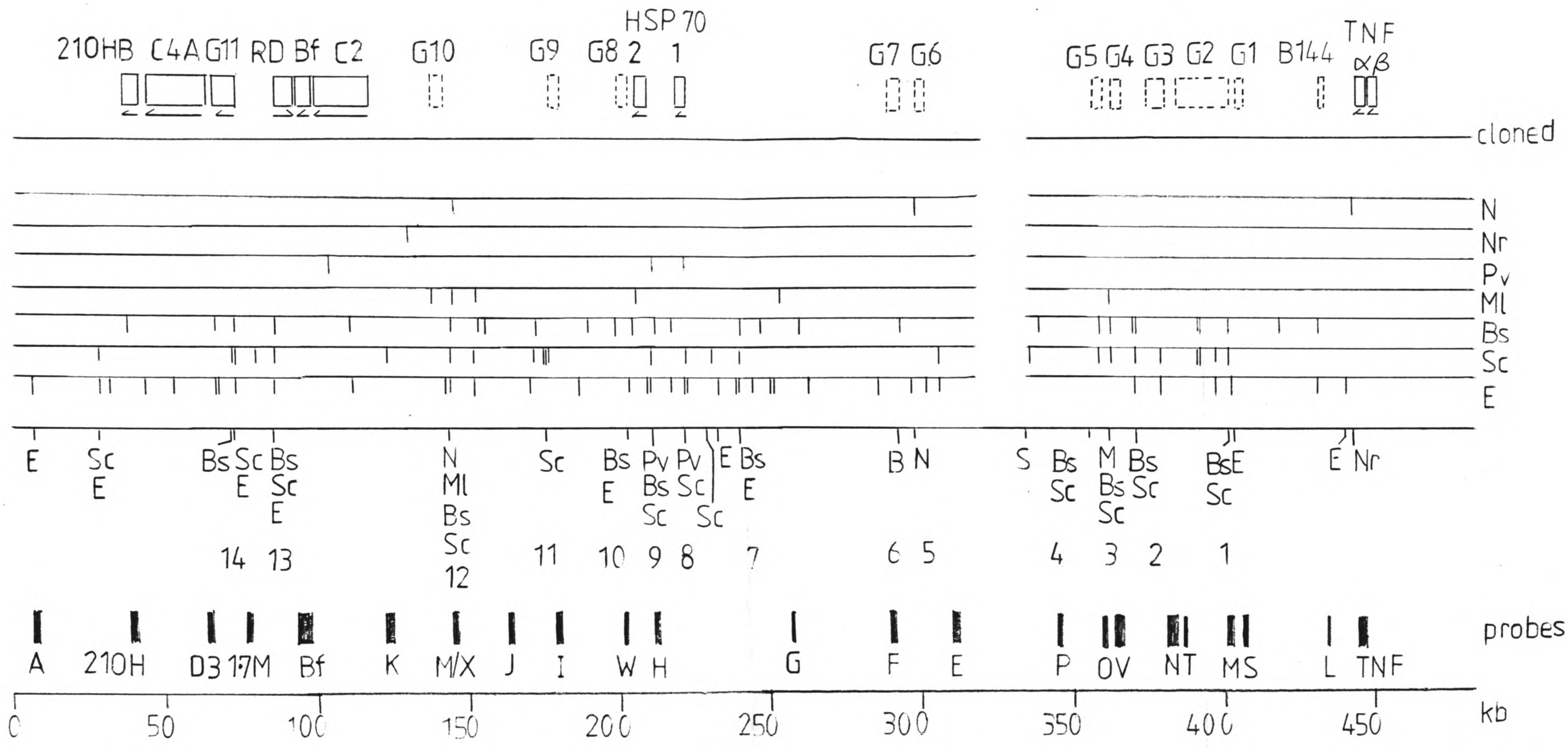
APPENDIX

Appendix

(I) A detailed restriction map of the cloned part of the class III region of the human MHC. The positions of the genes for 21OHB, C4A, RD, factor B, C2, HSP 70-1, HSP 70-2, TNFα and TNFβ are shown by the open boxes. The 5' to 3' orientations are indicated by the horizontal arrows. Sites for all the endonucleases used in restriction mapping are shown as vertical bars. Bam HI (M), Bgl II (B), Hind III (H), Kpn I (or its isoschizimer Asp 718) (K), Eco RV (V), Xho I (X), Sal I (S) and Cla I (C). The limits of the cosmid inserts are defined by the horizontal bars, and the positions of unique regions of DNA are defined by the boxes below the cosmids.

(II) A map showing the positions of the additional loci mappped to the cloned portion of the class III region. Where the precise position of a gene is known, it is shown as an open box. Where the limits of the genomic sequence have not been defined, the probable position of the locus is shown by the broken lines. The sites for the infrequently cutting endonucleases Not I (N), Nru I (Nr), Pvu I (Pv), Mlu I (Ml), Eag I (E), Bss HII (Bs) and Sac II (Sc) are shown as vertical bars. Sites which are known to cleave at a chromosomal level are given below. The probes used in the analysis are defined by the closed boxes.


210HB C4A G11 RD Bf C2
HSP 70
2 1
TNF
α β
cloned
M
B
H
K
V
X
S
C
C B A 210H D3 Bf K X2 J I H G F X1 E P O N M L TNF
0 50 100 150 200 250 300 350 400 450 500 550 600kb


HSP 70
210HB C4A G11 RD Bf C2 G10 G9 G8 2 1 G7 G6
G5 G4 G3 G2 G1 B144 TNF αβ
cloned
N
Nr
Pv
Ml
Bs
Sc
E
E Sc E Bs Sc E Bs Sc E N Ml Bs Sc 12 Sc Bs E Pv Bs Sc Pv Sc Sc E Bs E B N S Bs Sc M Bs Sc Bs Sc Bs E Sc E Nr
14 13 11 10 9 8 7 6 5 4 3 2 1
probes
A 210H D3 17M Bf K M/X J I W H G F E P OV NT MS L TNF
0 50 100 150 200 250 300 350 400 450
kb

PUBLICATIONS

*DUNHAM, I., SARGENT, C., TROWSDALE, J. & CAMPBELL, R.D. (1987) Molecular mapping of the human major histocompatibility complex by pulsed-field gel electrophoresis. PROC.NATL.ACAD.SCI.USA 84, 7237-7241.

*CAMPBELL, R.D., DUNHAM, I. & SARGENT, C.A. (1988) "Molecular Mapping of the HLA-linked Complement Genes and the RCA Linkage Group." EXP. & CLIN. IMMUNOGENET. 5, 81-98.

DUNHAM, I., SARGENT, C.A., TROWSDALE, J. & CAMPBELL, R.D. "Mapping of the Human Major Histocompatibility Complex by Pulsed-Field Gel Electrophoresis" in Immunobiology of HLA Vol.2 Immunogenetics and Histocompatibility, in press.

ABSTRACTS

DUNHAM, I., SARGENT, C.A.,TROWSDALE, J. & CAMPBELL, R.D. "Orientation and Location of the Complement Genes in the Human MHC by Pulsed-Field Gel Electrophoresis" in Complement 4, 124-244.

* Copies of these publications are included in Appendix II.

APPENDIX II

Experimental and Clinical Immunogenetics
Editor-in-Chief: K. Bauer, Heidelberg

Reprint
Publisher: S. Karger AG, Basel
Printed in Switzerland

0254-9670/88/0053-0081$2.75/0

Genetics of Complement
Expl clin. Immunogenet. *5:* 81–98 (1988)

Molecular Mapping of the HLA-Linked Complement Genes and the RCA Linkage Group

R.D. Campbell, I. Dunham, C.A. Sargent

MRC Immunochemistry Unit, Department of Biochemistry, University of Oxford, UK

Key Words. HLA-linked complement genes · RCA linkage group · Complement · Molecular mapping · Genes

Abstract. Phenotypic genetics have established linkage of the genes encoding proteins involved in the activation of the complement component C3. C2, factor B and C4, three of the structural components of the classical and alternative pathway C3 convertases, are encoded by genes which have been mapped to the class III region of the major histocompatibility complex (MHC) on human chromosome 6. The regulatory proteins factor H, C4BP, CR1, CR2 and DAF, which are involved in the control of C3 convertase activity, are encoded by closely linked genes, termed the regulators of complement activation (RCA) linkage group, that have been mapped to human chromosome 1. cDNA clones for all these proteins have been isolated, and this has made it possible to investigate the organization and structure of the MHC class III genes and the genes in the RCA linkage group. This short review summarizes some of the main features which have emerged from recent cloning work.

Introduction

The application of recombinant DNA techniques to the study of components of the complement system over the last few years has resulted in the cloning of cDNA and in most cases genomic DNA encoding these proteins [Campbell et al., 1988]. In many cases the chromosomal assignment has also been established [Reid, 1985]. These cloning studies have provided information which emphasizes that the members of the complement system can be divided into families of structurally and functionally related proteins, many of which are encoded by closely linked genes on the same chromosome.

Activation of the complement system occurs via two pathways, the classical and alternative, with the major component C3 playing a central role in each [Reid and Porter, 1981; Müller-Eberhard and Miescher, 1985; Reid, 1986]. In the classical pathway, activation is triggered primarily by the binding of C1 to IgG, or IgM, in immune complexes [Schumaker et al., 1987]. This results in the generation of proteolytic ac-

tivity in $C\overline{1}$ which cleaves and activates C4 to C4a and C4b. C2 associates with C4b and is cleaved by $C\overline{1}$ to C2a and C2b, thus yielding the C3 convertase $C\overline{4}b\overline{2}a$ which cleaves and activates C3. Activation of the alternative pathway can be mediated by a wide range of substances such as high molecular weight polysaccharides found in microorganisms, and also by complexes of IgG, IgA and IgE. Efficient activators are considered to possess sites for C3b where it is protected from control by factor I and its cofactors. Factor B associates with C3b, or $C3(H_2O)$ (a C3b-like form of C3), and is cleaved by factor $\overline{D}$ to Ba and Bb, thus yielding the C3 convertase $C\overline{3}b\overline{B}b$. This complex enzyme is stabilized by properdin and this causes potentiation of alternative pathway activation. The binding of C3b to $C\overline{4}b\overline{2}a$, or additional molecules of C3b to $C\overline{3}b\overline{B}b$, results in the generation of C5 convertase which activates C5 and initiates the self-assembly of the terminal components C5b to C9 involved in membrane lysis [Müller-Eberhard, 1986].

Control of the complement system is mediated in a number of different ways to prevent damage to the animal's own tissue. The C3/C5 convertases have a short half-life due to dissociation of C2a and Bb. In addition, the activity of the C3/C5 convertases is regulated by a number of proteins found in plasma, and by membrane-bound proteins and receptors [Holers et al., 1985; Sim et al., 1986; Reid et al., 1986; Kristensen et al., 1987]. These regulatory proteins include the plasma proteins factor H and C4b-binding protein (C4BP), and the membrane bound

Table I. Proteins involved in the activation of C3 and C5 [adapted from Law and Reid, 1988]

Generation of C3/C5 convertase

	MW kd	Plasma concentration		Chromosome location of gene	Enzymic site in activated form (+) (and natural substrate split)
		μg/ml	μ*M*		
Classical pathway					
C1-C1q[a]	462	80	0.17	1	–
C1r	83	50	0.30	12	+ (C1r, C1s)
C1s	83	50	0.30	12	+ (C4, C2)
C4	205	600	3.00	6	–
C2	102	20	0.20	6	+ (C3, C5)
C3	185	1,300	7.02	19	–
Alternative pathway					
Factor D	24	1	0.04	n.k.	+ (B)
Factor B	92	210	2.20	6	+ (C3, C5)
C3	185	1,300	7.02	19	–

[a] C1q is composed of 18 chains (6A+6B+6C). The genes for the A and B chains are on chromosome 1. The gene for the C chain has not yet been mapped.

Table I (continued)

Plasma proteins involved in control of C3/C5 convertase

Protein	MW kd	Plasma concentration		Specificity	Chromosome location of gene	Role
		µg/ml	µ*M*			
C4-binding protein	500	250	0.45	C4b	1	accelerates decay of $C\overline{4}b\overline{2}a$ and acts as cofactor in the cleavage of C4b by factor I
Factor H	150	480	3.20	C3b	1	accelerates decay of $C\overline{3}b\overline{B}b$ and acts as cofactor in the cleavage of C3b by factor I
Factor I	88	35	0.39	C4b, C3b	4	protease which inactivates C4b and C3b with the aid of cofactors C4BP, H, CR1 and MCP
Properdin	220	20	0.09	C3bBb	X	positive regulator of the alternative pathway which stabilizes the C3/C5 convertases

Membrane proteins involved in control of C3/C5 convertase

Membrane molecule	MW kb	Fragment specificity	Chromosome location of gene	Principal roles	Major human cell types positive[b]
Complement receptor type 1	type D 250 type B 220 type A 190 type C 160 (four structural allotypes)	C3b, C4b	1	regulation of C3b breakdown, binding of immune complexes to erythrocytes, phagocytosis, accelerates decay of C3/C5 convertases	E, B, G, M
Complement receptor type 2	145	C3d, C3dg, iC3b	1	regulation of B cell functions, Epstein-Barr virus receptor	B
Membrane cofactor protein	45–70	C3b, C4b	n.k.	regulation of C3b breakdown	B, T, N, M
Decay-accelerating factor	70	C4b2a, C3bBb	1	accelerates decay of C3/C5 convertases	E, L, P

[b] Human cell types: B, B lymphocytes; E, erythrocytes; G, granulocytes; L, leucocytes; M, monocytes; N, neutrophils; P, platelets.

n.k. = not known.

molecules complement receptor type 1 (CR1), complement receptor type 2 (CR2), decay accelerating factor (DAF) and probably also membrane cofactor protein (MCP). A summary of some of the features of the proteins involved in the generation of the C3/C5 convertases and in the control of C3/C5 convertase activity can be found in table I.

Investigation of the phenotypic genetics of complement components [Hobart, 1984] and of inherited deficiencies of the components [Lachmann, 1984] have established linkage of the genes encoding proteins involved in the activation of C3. Three of the structural components of the C3 convertases, C2, factor B and C4, are encoded by genes which have been mapped to the class III region of the major histocompatibility complex (MHC) on the short arm of human chromosome 6 [Alper, 1981; Campbell et al., 1986]. Phenotypic genetics have also established linkage of the genes encoding regulatory proteins involved in the control of C3 convertase activity [Rodriguez de Cordoba et al., 1985]. This linkage group, termed the regulators of complement activation (RCA) linkage group, maps to human chromosome 1 and includes the genes for factor H, C4BP, CR1, CR2 and DAF.

Several comprehensive reviews dealing with the complement system [Müller-Eberhard and Miescher, 1985; Reid, 1986], the genetics of components of complement [Reid, 1985; Campbell et al., 1986, 1988], and the polymorphism of C2, factor B [Campbell, 1987] and C4 [Carroll and Alper, 1987] have been published recently. This short review will summarize some of the main features which have emerged from recent cloning work. It covers the organization and structure of the MHC class III genes and the genes in the RCA linkage group.

MHC-Linked Complement Genes

The human MHC is located on the short arm of chromosome 6 in the distal portion of the 6p21.3 band [Lamm and Olaisen, 1985]. It consists of three major linked gene clusters. The class I loci (HLA-A, -B, -C) encode cell surface glycoproteins found on almost all nucleated cells [Strachan, 1987], whereas the class II loci (HLA-DP, -DQ, -DR) encode cell surface glycoproteins found principally on B lymphocytes, macrophages and activated T cells [Trowsdale, 1987]. Both class I and class II antigens are highly polymorphic and act as restriction elements in the recognition and interaction of regulatory and effector T lymphocytes with their target cells. Analysis of recombinant MHC haplotypes in family studies has established that the class I loci are telomeric to the class II loci. Within the class II region the DP subregion maps on the centromeric side of the DQ and DR subregions, while the order of the loci within the class I region is HLA-B, HLA-C, HLA-A, telomere.

The first evidence for the existence of complement genes within the MHC was provided by Allen [1974] who demonstrated in family studies that the electrophoretic polymorphism of factor B [Alper et al., 1972] segregated with the MHC. The linkage of C2 and of C4 to the MHC followed from the description of deficiencies of these proteins in plasma [Fu et al., 1974; Hauptmann et al., 1974; Rittner et al., 1975], and from the demonstration of electrophoretic variants [Alper, 1976; Hobart and Lachmann, 1976; Meo et al., 1977; Teisberg et al., 1976, 1977]. However, in order to explain the electrophoretic patterns observed for C4, O'Neill et al. [1978a, b] proposed a two locus model for C4, now referred to as C4A and C4B. Studies

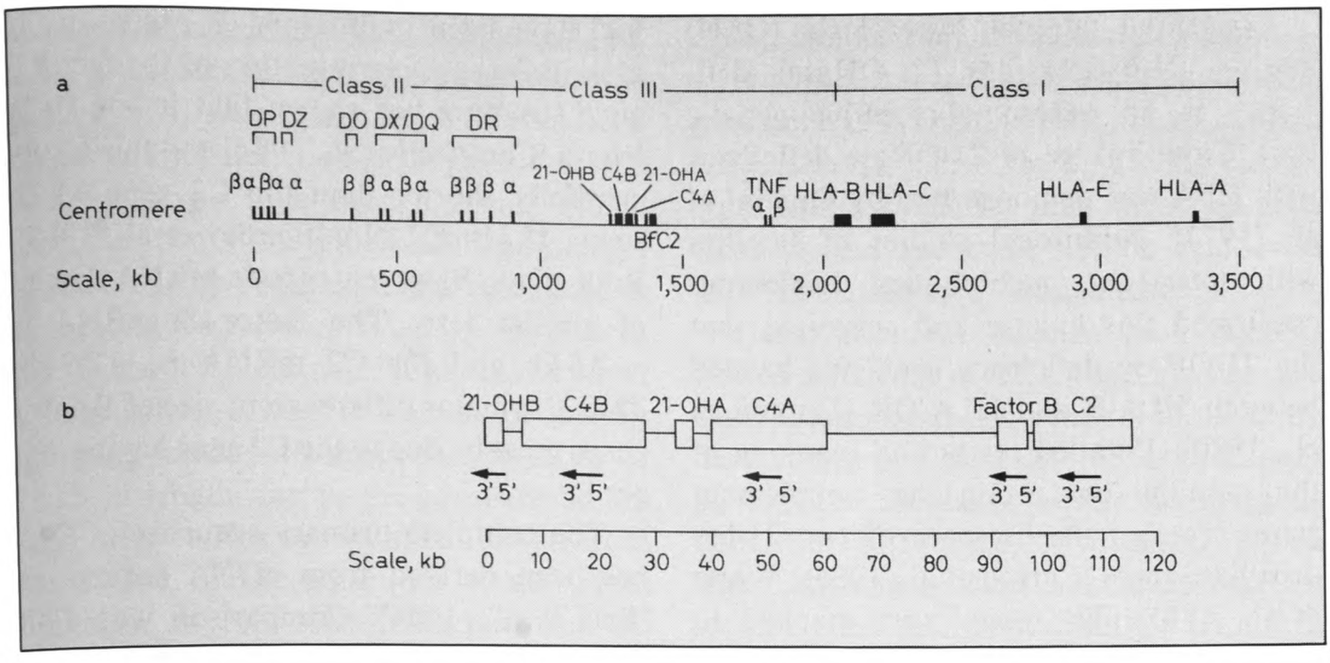


Fig. 1. Molecular map of the human MHC. **a** The molecular map has been derived from PFGE and cosmid cloning experiments which can be found in Hardy et al. [1986], Carroll et al. [1987b] and Dunham et al. [1987]. **b** Expanded map of the complement and 21-OHase (21-OH) loci in the class III region. The solid arrows denote the direction of transcription of the genes.

of the inheritance of the polymorphic variants of C2, factor B and C4, and various recombinant MHC haplotypes have placed the structural genes for these proteins (termed the class III genes) between the HLA-B and HLA-DR loci [Olaisen et al., 1983; Lamm and Olaisen, 1985]. However, despite a large number of studies [Alper et al., 1983] no recombination events have been observed between the genes, and this was taken to suggest that they were closely linked. Further, it was not possible to define the order of the genes relative to one another. The cloning of cDNA for C2, factor B and C4 has made it possible to study the genetic organization of the corresponding genes.

Molecular Cloning of C2, Factor B and C4

The isolation of cDNA clones for C2, factor B and C4 has been achieved through the screening of liver cDNA libraries with oligonucleotides synthesized on the basis of the known protein sequence. These cDNA clones have been used to screen cosmid libraries of human genomic DNA to isolate clusters of overlapping cosmid clones containing the corresponding genes [Carroll et al., 1984b]. Characterization of the cosmids revealed that the C2 and factor B genes were closely linked, and it has since been shown that the transcription start-point of the factor B gene lies only 421 bp from the polyadenylation site of the C2 gene [Wu et al., 1987]. Approximately 30 kb from the factor B gene are the C4 loci which are separated by ~ 10 kb of DNA [Carroll et al., 1984b]. Analysis of the cosmid clones using C4A and C4B class-specific synthetic oligonucleotides showed that the gene encoding C4A lay closer to the factor B gene (fig. 1) [Carroll et al., 1984a].

Congenital adrenal hyperplasia (CAH) due to 21-hydroxylase (21-OHase) deficiency is an autosomal-recessive genetic trait. Close linkage of 21-OHase deficiency with HLA was demonstrated by Dupont et al. [1977]. Additional studies of families with intra-HLA recombinant haplotypes confirmed this linkage and suggested that the 21-OHase deficiency gene was located between HLA-B and HLA-DR [Dupont et al., 1980]. Detailed restriction mapping of the cosmids containing the complement genes revealed the presence of two 21-hydroxylase genes [Carroll et al., 1985a; White et al., 1985]. The genes were mapped to within 3 kb of the 3′ end of each C4 gene (fig. 1). However, only the gene lying 3′ to the C4B locus, termed the 21-OHase B gene, appears to be important in steroid biogenesis in the adrenal gland. Homozygous deletion of the 21-OHase B gene has been described in some individuals suffering from CAH [White et al., 1984]. In addition, DNA sequencing of both genes has established that, although 97% homologous, the 21-OHase A gene is a pseudogene due to deleterious mutations in the coding sequence [Higashi et al., 1986; White et al., 1986; Rodrigues et al., 1987].

The complete primary sequences of C2 [Bentley, 1986] and factor B [Morley and Campbell, 1984; Mole et al., 1984] have been derived from cDNA sequencing. Comparison of the cDNA and amino acid sequences indicates that they share 42% nucleotide and 39% amino acid identity. C2 and factor B perform analogous functions in the classical and alternative pathway C3 convertases. The overall similarity in structure and function of the two proteins indicates that the C2 and factor B genes arose by duplication of an ancestral gene, and the two loci have been maintained in exceptionally close linkage. Determination of the factor B gene structure has shown that it is 6 kb in length [Campbell et al., 1984] and this is substantially shorter than the C2 gene which spans 18 kb of DNA [Bentley et al., 1985]. Both genes, however, encode mRNA species of similar size. The factor B mRNA is ~2.6 kb and the C2 mRNA is ~2.9 kb. Thus the major difference in size of the two genes must be due to the C2 gene having longer introns.

The complete primary sequence of C4A has been derived from cDNA sequencing [Belt et al., 1984]. Comparison with C4B cDNA has revealed that the two C4 isotypes are >99% homologous. Of the 14 differences which were observed, 12 are clustered in a region of the molecule derived from proteolytic cleavage of C4b by factor I and called C4d. These differences cause 9 amino acid substitutions. Comparison of the sequences of 9 different C4A and C4B cDNA and genomic clones in the C4d region has established the pattern of polymorphism in the C4d fragment [Belt et al., 1985; Yu et al., 1986] and has provided a structural basis for the observed serological [see Giles, this volume] and functional differences [Law et al., 1984; Isenman and Young, 1984] between the two C4 isotypes.

The two C4 loci probably arose through a single duplication event which encompassed ~28 kb of DNA. They are often referred to as locus I and locus II. In general, locus I encodes C4A and locus II encodes C4B, though exceptions to this are rather common. Although both are transcribed into mRNA species of ~5.5 kb, a difference in the size of the loci has been observed. Long C4 genes are 22 kb in length while short C4 genes are 16 kb in length. Of the C4A genes

analysed to date all are ~ 22 kb in length [Yu et al., 1986; Schneider et al., 1986; Palsdottir et al., 1987b]. One exception to this has been suggested by Giles et al. [1987] who found a rare haplotype in a French family which expresses two C4A allotypes. Examination of the C4 genes revealed that a short gene at the second C4 locus encoded a C4A protein [see Giles, this volume]. However, no short C4A gene has yet been defined at locus I. On the other hand, C4B genes can be 22 kb or 16 kb in length due to the presence or absence of a 6–7 kb intron about 2.5 kb from the 5′ end of the gene [Yu et al., 1986; Schneider et al., 1986; Palsdottir et al., 1987b]. Long C4B genes included C4B1 genes on most C4A3 C4B1 haplotypes and also C4BQO alleles. An estimate of the frequency of C4B loci with the 6.5 kb intron, based on haplotype frequencies, suggests they are twice as frequent as C4B loci without the intron. The nature of the 6–7 kb intron is not known though it has been suggested that it could be a member of the long interspersed sequences, LINE or LI [Yu et al., 1986]. These sequences are repeated about 10^4 times in the haploid genome and have homology to retroviral reverse transcriptase (retroposons) [Singer and Skowronski, 1985]. Although 80% of haplotypes carry two C4 loci, differences have also been observed in the number of C4 genes present. This was originally demonstrated at the protein level where duplication of C4B as C4B1, C4B2 was found on the extended haplotype B14 DR1 [Raum et al., 1984; Rittner et al., 1984; Uring-Lambert et al., 1984]. The gene frequency has been estimated at 1–2%, though this may be an underestimate as three new kinds of C4 gene duplication were identified in family studies by combined protein and RFLP analysis [Schneider et al., 1986]. Duplication of C4 has also been observed from cosmid cloning where three C4 genes, one C4A and two C4B, were found on one haplotype [Carroll et al., 1984a]. Haplotypes with a single C4 gene have also been characterized and these will be dealt with later in relation to C4 null alleles.

Genetic studies in man have demonstrated deficiency of C2, C4 and in a few rare cases heterozygous deficiency of factor B. By far the most prevalent disorder is deficiency of C2. The incidence of the homozygous deficiency is 1 in 10,000 [Agnello, 1978]. The C2 deficiency allele is usually found in a specific HLA extended haplotype HLA B18 C4A4 C4B2 BfS C2QO DR2 [Awdeh et al., 1981], suggesting that most C2-deficient patients will have the same mutation. No major gene deletion or rearrangement of the C2 gene has been observed by genomic Southern blot analysis in individuals deficient in C2 [Cole et al., 1985]. A defect at the level of transcription of the gene or post-transcriptional processing of C2 mRNA is the most likely reason as no detectable C2 mRNA could be found by Northern blot analysis of peripheral blood monocyte RNA preparations from C2-deficient individuals.

Complete deficiency of C4 in man is a rare disorder, and only 16 cases have been reported [Hauptmann et al., 1986]. Homozygous C4 deficiency is associated with different HLA haplotypes suggesting that different mutations may have occurred. Three complete C4-deficient patients have been analysed by Southern blot analysis [Hauptmann et al., 1987]. All 3 patients had C4A genes present, 2 showed deletion of C4B genes, while the third appeared normal at the DNA level with two C4 loci on each haplotype. This clearly demonstrates that the complete absence of C4 in plasma is due to abnormal-

ities in transcription or translation of the genes, in addition to single gene deletions.

Occurrence of null alleles in the population at either locus defined by the absence of C4A or C4B in plasma is much more common. Gene frequencies of 5–15% for C4AQO alleles and 10–20% for C4BQO alleles [Schendel et al., 1984; Partanen and Koskimies, 1986] have been estimated. About half of the null alleles are due to deletion of the gene usually together with the flanking 21-OHase gene [Carroll et al., 1985b; Schneider et al., 1986]. These single C4 gene haplotypes can be of three types: long, expressing C4A protein; long, expressing C4B protein; or short, expressing C4B protein. In other situations where the null allele is present, defects in transcription or translation as discussed above will be the cause of the absence of C4 protein. However, another possibility is that some of the genes are transcriptionally active, but encode products similar to the adjacent locus. This may be the case for the C4BQO allele on the HLA haplotype B44 C2C BfS C4A3 C4BQO DR6. Yu and Campbell [1987] were able to define an RFLP using NlaIV which distinguishes alleles encoding the C4A or C4B isotypes. It was found that the C4BQO allele had an RFLP pattern consistent with it encoding a C4A isotype. It was suggested on the basis of the phenotypes expressed by the individuals concerned that the C4BQO allele probably encoded another C4A3 allotype.

Haplotypes expressing two different C4A allotypes, e.g. C4A3 C4A2 or C4A5 C4A2 with a null allele at the C4B locus, have been found in several studies [Raum et al., 1984; Uring-Lambert et al., 1984]. It was suggested that they may be due to unequal crossover between homologous chromosomes, resulting in duplication of the C4A locus and deletion of the C4B locus [Rittner et al., 1984]. In a detailed Southern blot analysis of such haplotypes by Palsdottir et al. [1987a] it was found that both loci are present as normal, but that the second (or C4B) locus encodes a C4A protein. However, in another example of a C4A3 A2 BQO haplotype studied by Giles et al. [1987] the presence of three C4 loci has been postulated to account for the intensities of fragments seen in Southern blot analysis.

Molecular Mapping of the HLA Class III Region

A number of studies have been arried out to try and define the position and orientation of the complement genes in the class III region, but the results have been contradictory [see Lamm and Olaisen, 1985]. Whitehead et al. [1985] using deletion mutant cell lines and a C4 probe were able to physically map the C4, and thus the C2 and factor B genes between HLA-B and HLA-DR, and suggested on the basis of the number of cell lines which lost or retained the C4 loci that the complement cluster may be closer to HLA-B than HLA-DR. However, in two recent studies in families with recombinant HLA haplotypes the complement loci segregated more frequently with HLA-DR suggesting that they may be closer to the class II region than the class I region [Robinson et al., 1985; Yunis et al., 1985]. Although clusters of overlapping clones have been isolated from the subregions of the class II region [Trowsdale, 1987] and from the class III [Carroll et al., 1984b; Dunham et al., 1987] and class I [Strachan, 1987] regions, they have not yet been linked by chromosomal walking procedures. Recent advances, however, in the separation of large DNA frag-

ments by gel electrophoresis have made it possible to establish the physical linkage of loci separated by hundreds of kilobases. Pulsed field gel electrophoresis (PFGE), originally developed by Schwartz and Cantor [1984], can be used to separate DNA fragments up to 9,000 kb [Anand, 1986]. Such large DNA fragments can be generated from mammalian genomic DNA by restriction enzymes that cut rarely in the genome. Suitable rare-cutting enzymes are those with 6- or 8-base pair recognition sites that contain one or more CpG dinucleotides since CpG is known to be underrepresented in bulk mammalian DNA [Bird, 1986]. In addition, many enzymes with CpG in their recognition sequence appear to be sensitive to cytosine methylation [Brown and Bird, 1986], and this decreases the frequency at which these enzymes cut in the genome. Thus large DNA fragments generated with these enzymes can be separated by PFGE, and Southern blot hybridization analysis can be performed with probes from the area of interest to construct long-range genomic restriction maps.

Recently these procedures have been applied to the molecular mapping of the human MHC [Ragoussis et al., 1986; Hardy et al., 1986; Lawrance et al., 1987; Carroll et al., 1987b, Dunham et al., 1987]. Physical linkage of the class II and class III loci was established in a NotI fragment of ~980 kb [Dunham, et al., 1987]. This fragment is digested with NruI to yield a fragment of 700 kb that hybridizes with the DQα and DRα probes, and a fragment of 280 kb which contains the complement loci. It should be noted that the size of the NotI fragment has been estimated by others to be 920–1,000 kb. The difference in the estimation of the size of this fragment probably reflects the use of different apparatus in the pulsed field work. However, three different HLA haplotypes have also been studied and it remains possible that actual differences in fragment size exist due to RFLPs and/or insertions or deletions of DNA.

In order to orientate the complement loci it was necessary to isolate further overlapping cosmid clones as none of the restriction enzymes used in the PFGE analysis were found to cut between these genes. A series of overlapping cosmids extending from the C2 gene were mapped for NotI, MluI, NruI, and ClaI and this revealed a cluster of sites (three MluI, two ClaI, one NruI and one NotI) in a 25 kb region ~25 kb from the transcriptional start site of the C2 gene [Dunham et al., 1987]. A new single-copy hybridization probe distal to this cluster of sites and the C2 gene was isolated. This probe hybridizes to a NotI fragment of ~210 kb and not the 980 kb NotI fragment containing the complement and DRα loci. Since the exact position of the NotI site that is cleaved in genomic DNA is known from the cosmid map, this established that the C2 gene lies telomeric to the 21-OHase B gene [Dunham et al., 1987]. The results of further single and double digests established that the class III region spans ~1,100 kb and that the C2 gene lies ~650 kb from the class I region, while the 21-OHase B gene lies ~300–360 kb from the class II region (fig. 1). The physical size of the class III region is similar to that estimated from genetic recombination data (1 cM) assuming 1 cM = 1,000 kb.

Analysis of deletion mutant cell lines has recently established linkage of the genes encoding the cytokines tumour necrosis factors α and β to the MHC [Spies et al., 1986], but it was not possible to determine their exact location. A number of studies using PFGE have established that the TNFα and β genes

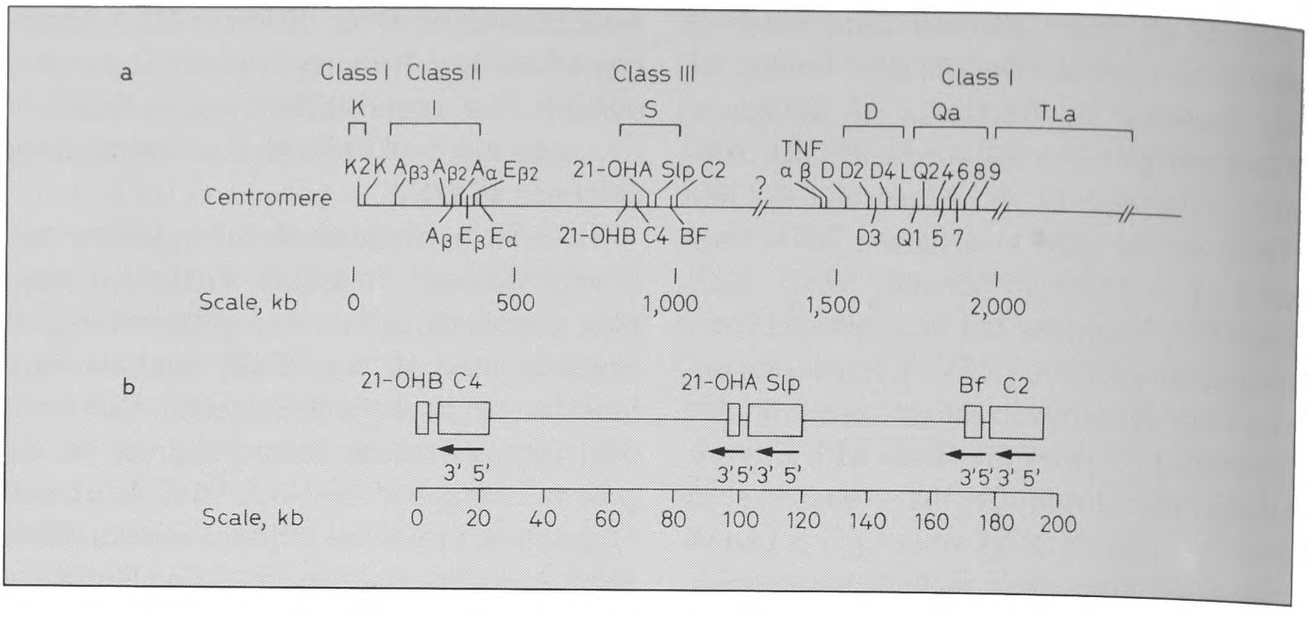


Fig. 2. Molecular map of the mouse H-2 region. **a** The molecular map has been derived from PFGE and cosmid cloning experiments which can be found in Chaplin [1985] and Müller et al. [1987b]. **b** Expanded map of the complement and 21-OHase (21-OH) loci in the S region. The solid arrows denote the direction of transcription of the genes.

are linked to the HLA-B locus [Carroll et al., 1987b; Dunham et al., 1987; Inoko and Trowsdale, 1987; Ragoussis et al., 1988], and lie between HLA-B and the C2 gene (fig. 1). Estimates of ~250 kb have been suggested for the distance between the TNF genes and HLA-B, and these are in reasonable agreement with the 210 kb which has been established through cosmid walking [cited in Carroll et al., 1987b].

Further analysis by PFGE using a number of class I and class II probes has established a long-range map of the MHC (fig. 1). This indicates that the MHC could span 3–4 Mbp of DNA [Carroll et al., 1987b; Dunham et al., 1987], which represents about 1/750 of the human genome. This is in good agreement with the overall size of the MHC of 3–4 cM which has been suggested from studies of recombination events in families and population data on linkage disequilibrium [Lamm and Olaisen, 1985].

The linkage of the C2, factor B and C4 genes to the human MHC appears not to be fortuitous, as they have been mapped to the MHC of a number of other animal species including frog. Of these, the best characterized is the mouse H-2 complex where an analysis by cosmid cloning [Chaplin, 1985] and PFGE [Müller et al., 1987b] has revealed that the organization and orientation of the complement loci is the same as that in man (fig. 2). The single C2 and factor B genes were found to be closely linked [Chaplin, 1985] and to lie ~50 kb from two C4-like genes separated by ~80 kb. It was later established that the gene closest to factor B encoded Slp [Chaplin et al., 1986]. Two genes encoding 21-OHase were also placed immediately 3′ of the Slp and C4 genes, respectively; in contrast to the situation in humans, however, it is the 21-OHase A gene that is important in steroid biogenesis [Chaplin et al., 1986].

Estimates based on PFGE placed the C4 gene ~ 420 kb from the Eα gene, and the C2 gene at least 470 kb from H-2D [Müller et al., 1987b] (fig. 2). Although genes for TNFα and β were shown to lie between the C2 and H-2D genes [Müller et al., 1987a], they are only 70 kb from the mouse class I gene compared with ~ 250 kb for the analogous genes in the human MHC.

One of the interesting features of the genes in the class III region is that they are organized in pairs of related genes and have the same 5′ to 3′ orientation. The C2 and factor B genes probably arose through duplication of an ancestral gene. However, the marked divergence of the intron lengths between the two genes and the fact that the two gene products share only 39% sequence identity suggests that the duplication occurred a considerable time ago. The TNFα and β genes also probably arose through duplication. The two genes are only 1.1 kb apart, and they are of similar size. However, the two genes differ significantly in the 5′-untranslated sequence and in the region encoding the signal peptide. In addition, they share only 35% sequence homology which again suggests that the duplication occurred a considerable time ago.

The finding of two 21-OHase genes closely linked to the two C4 genes suggests that the ancestral 21-OHase gene became linked to the ancestral C4 gene prior to duplication. Comparison of the C4A and C4B coding sequence has shown that they differ by less than 1%. The 21-OHase genes are also highly homologous and differ by only about 3%. The strong homology of these two sets of genes suggests that they have only recently duplicated. The duplication may have occurred after mammalian speciation, as only single C4 genes, and presumably also single 21-OHase genes, have been observed in the guinea pig and the Syrian hamster.

Based on the density of genes which have been identified in the cloned portions of the MHC it is highly likely that other genes will be defined in the large regions of DNA separating the class I, II and III loci. One such gene may be neuraminidase which has been mapped to the S region of the mouse H-2 complex, and may also be linked to the human MHC. Given that duplication is a general theme of the evolution of the genes in class III region it is likely that other duplicated loci will be located in this region.

RCA Linkage Group

Control of complement activation is exerted at a number of levels to ensure that overactivation of the system does not occur. One of the levels of control is the presence in plasma and on cell membranes of a number of control 'cofactor proteins' [Holers et al., 1985]. These cofactors form noncovalent complexes with the activation fragments of C3 or C4 (C3b or C4b) such that C3b or C4b in the complex can be cleaved by the serine protease factor I. Once C3b or C4b have been proteolytically cleaved by factor I they are unable to interact with factor B or C2, respectively, to form functional C3 convertase. The proteins which interact with C3b or C4b either to cause dissociation of the C3 convertase and/or to mediate breakdown of C3b or C4b include factor H, C4BP, CR1, CR2, MCP and DAF (table I).

The linkage of the genes encoding factor H, C4BP and CR1 was established by the definition of genetic polymorphism in these proteins, which has been demonstrated by

electrophoretic techniques. Three forms of C4BP have been identified by IEF of neuraminidase-treated EDTA plasma and family studies have indicated the Mendelian inheritance of the forms [Rodriguez de Cordoba et al., 1983; 1984; Rodriguez de Cordoba and Rubinstein, 1987]. The results were taken to suggest the presence of three alleles at a single autosomal locus. The same technique has also demonstrated five factor H variants [Rodriguez de Cordoba and Rubinstein, 1984, 1987], which were shown to be inherited in Mendelian fashion and to be alleles at a single locus. The CR1 locus is also polymorphic and determines at least four alleles [Dykman et al., 1983, 1985; Wong et al., 1983]. However, CR1 exhibits an unusual form of polymorphism in which allotypic variants that vary in molecular weight have been identified on human erythrocytes and leukocytes by SDS-PAGE. Analysis of the genetic variants of C4BP and CR1 in three pedigrees informative for the segregation indicated no recombination suggesting close linkage of the loci [Rodriguez de Cordoba et al., 1984]. Family segregation data of the genetic variants indicated that the factor H gene was linked to the C4BP and CR1 genes [Rodriguez de Cordoba et al., 1985]. Subsequently the gene encoding DAF was also mapped close to the factor H, CR1 and C4BP genes [Rey-Campos et al., 1987], and studies using cDNA probes have shown that the genes lie on human chromosome 1.

Molecular Cloning of Factor H, C4BP, CR1, CR2 and DAF

cDNA clones for factor H, C4BP, CR1, CR2 and DAF have been reported. In most cases they were isolated from the appropriate cDNA libraries using mixed synthetic oligonucleotides based on the available protein sequence. In one instance, a CR2 cDNA was isolated using the CR1 cDNA under less stringent hybridization conditions [Weis et al., 1986], illustrating the high degree of homology between these proteins. Complete derived amino acid sequences are now available for all these proteins. Although they differ markedly in size, their structures are highly homologous and contain contiguous units each of 60 amino acids long [Reid et al., 1986]. These repeating units, termed short consensus repeats (SCR), are based on a framework of 4 cysteine residues, together with highly conserved tryptophan, glycine and proline residues. Further, these repeat units are homologous to repeat units found in C2 and factor B, and also in a number of noncomplement proteins, illustrating their widespread distribution. A fuller description of the SCR and the proteins which contain it can be found in the article by Bentley [this volume].

The cDNA probes for C4BP, factor H, CR1 and DAF have been used in genomic Southern blot analysis to estimate the size of the genes, and have also been used to isolate the C4BP and CR1 genes from lambda genomic libraries. In man a single C4BP gene of ~30 kb has been reported [Lintin and Reid, 1986; Lintin et al., 1987], and also a single DAF gene of ~35 kb [Stafford et al., 1987]. Preliminary analysis of the factor H gene suggests that it is likely to be 80–110 kb in length [McAleer et al., 1987], and a related gene or pseudogene may also be present. The CR1 gene spans about 140 kb of DNA [Wong et al., 1987], and restriction enzyme mapping and Southern blot analysis of overlapping genomic clones have determined the basis of the size polymorphism observed between the CR1-A and CR1-B allelic variants. The CR1 gene is composed of homologous

genomic segments which encode long homologous repeats (LHR). The LHR is composed of 7 SCRs and large portions of the LHRs show up to 99% identity [Klickstein et al., 1987a, b]. The CR1-B (or S) allele contains 5 distinct segments of 15–25 kb, each encoding a LHR, while the CR1-A (or F) allele contains 4 segments [Wong et al., 1987]. It appears that one of the LHR genomic segments has duplicated in the CR1-A allele to generate the CR1-B allele. The other allelic size variants of CR1 are also probably due to the loss or gain of coding sequence of 1 LHR.

Molecular Mapping of the RCA Locus

Linkage analysis of allotypes of C4BP, factor H and CR1 indicated that they are closely clustered. Analysis of human-mouse somatic cell hybrids using cDNA probes for C4BP and factor H indicate that the genes are located on chromosome 1 [Reid et al., 1986]. In situ hybridization studies have placed the CR1 and CR2 genes on the long arm of chromosome 1 in the q32 region [Weiss et al., 1987].

Recently the technique of PFGE has been used to determine the organization of these genes in man [Carroll et al., 1987a; Rey-Campos et al., 1988]. Four of the 5 genes were found to be clustered in a common NotI/NruI fragment of ~950 kb. Further analysis using NotI/MluI double digestion indicated that the genes could be split into two clusters, one containing the CR1 and CR2 genes in a 450-kb fragment, while the second contains the DAF and C4BP genes in a 500-kb fragment. It was also suggested that the CR1 gene is at least 150 kb in length and lies within 50 kb of the CR2 gene. No information is yet available on the position of the factor H gene relative to the CR1, CR2, DAF and C4BP genes. Although on chromosome 1 the genetic data of Rodriguez de Cordoba et al. [1985, 1986] suggest that it could lie some 5–10 cM from this cluster of genes.

References

Agnello, V.: Complement deficiency states. Medicine *57:* 1–23 (1978).

Allen, F.H.: Linkage of HL-A and GBG. Vox Sang. *27:* 382–384 (1974).

Alper, C.A.: Inherited structural polymorphism in human C2: evidence for genetic linkage between C2 and Bf. J. exp. Med. *144:* 1111–1115 (1976).

Alper, C.A.: Complement and the MHC; in Dorf, The role of the major histocompatibility complex in immunobiology, pp. 173–220 (Garland, New York 1981).

Alper, C.A.; Boenisch, T.; Watson, L.: Genetic polymorphism in human glycine-rich beta-glycoprotein. J. exp. Med. *135:* 68–80 (1972).

Alper, C.A.; Raum, D.; Karp, S.; Awdeh, Z.L.; Yunis, E.T.: Serum complement 'supergenes' of the major histocompatibility complex in man (complotypes). Vox Sang. *45:* 62–67 (1983).

Anand, R.: Pulsed field gel electrophoresis: a technique for fractionating large DNA molecules. Trends Genet. *2:* 278–283 (1986).

Awdeh, D.Z.; Raum, D.D.; Glass, D.; Agnello, V.; Schur, P.; Johnston, R.B.; Gelfand, E.W.; Ballow, M.; Yunis, E.; Alper, C.A.: Complement human histocompatibility antigen haplotypes in C2 deficiency. J. clin. Invest. *67:* 581–583 (1981).

Belt, K.T.; Carroll, M.C.; Porter, R.R.: The structural basis of the multiple forms of human complement component C4. Cell *36:* 907–914 (1984).

Belt, K.T.; Yu, C.Y.; Carroll, M.C.; Porter, R.R.: Polymorphism of human complement C4. Immunogenetics *21:* 173–180 (1985).

Bentley, D.R.: Primary structure of human complement component C2: homology to two unrelated protein families. Biochem. J. *239:* 339–345 (1986).

Bentley, D.R.; Campbell, R.D.; Cross, S.J.: DNA polymorphism of the C2 locus,. Immunogenetics *22:* 377–390 (1985).

Bird, A.P.: CpG-rich islands and the function of DNA methylation. Nature, Lond. *321:* 209–213 (1986).

Brown, W.; Bird, A.P.: Long-range restriction site mapping of mammalian genomic DNA. Nature, Lond. *322:* 477–481 (1986).

Campbell, R.D.: Molecular genetics of C2 and factor B. Br. med. Bull. *43:* 37–49 (1987).

Campbell, R.D.; Bentley, D.R.; Morley, B.J.: The factor B and C2 genes. Phil. Trans. R. Soc. *306:* 367–378 (1984).

Campbell, R.D.; Carroll, M.C.; Porter, R.R.: The molecular genetics of components of complement. Adv. Immunol. *38:* 203–244 (1986).

Campbell, R.D.; Law, S.K.A.; Reid, K.B.M.: Sim, R.B.: Structure, organisation and regulation of the complement genes. A. Rev. Immunol. *6:* 161–196 (1988).

Carroll, M.C.; Alicot, E.A.; Katzman, P.; Klickstein, L.B.; Fearon, D.T.: Organisation of the genes encoding CR1, CR2, DAF and C4bp in the RCA locus on human chromosome 1. Complement *4:* 141 (1987a).

Carroll, M.C.; Alper, C.A.: Polymorphism and molecular genetics of human C4. Br. med. Bull. *43:* 50–65 (1987).

Carroll, M.C.; Belt, K.T.; Palsdottir, A.; Porter, R.R.: Structure and organistion of C4 genes. Phil. Trans. R. Soc. *306:* 379–388 (1984a).

Carroll, M.C.; Campbell, R.D.; Bentley, D.R.; Porter, R.R.: A molecular map of the human major histocompatibility complex class III region linking complement genes C4, C2 and factor B. Nature, Lond. *307:* 237–241 (1984b).

Carroll, M.C.; Campbell, R.D.; Porter, R.R.: The mapping of 21-hydroxylase genes adjacent to complement component C4 genes in HLA, the major histocompatibility complex in man. Proc. natn. Acad. Sci. USA *82:* 521–525 (1985a).

Carroll, M.C.; Katzman, P.; Alicot, E.M.; Koller, B.H.; Geraghty, D.E.; Orr, H.T.; Strominger, J.L.; Spies, T.: A linkage map of the human major histocompatibility complex including the tumor necrosis factor genes. Proc. natn. Acad. Sci. USA *84:* 8535–8539 (1987b).

Carroll, M.C.; Palsdottir, A.; Belt, K.T.; Porter, R.R.: Deletion of complement C4 and steroid 21-hydroxylase genes in the HLA class III region. Eur. molec. Biol. Org. J. *4:* 2547–2552 (1985b).

Chaplin, D.D.: Molecular organisation and in vitro expression of murine class III genes. Immunol. Rev. *87:* 61–80 (1985).

Chaplin, D.D.; Galbraith, L.J.; Seidman, J.G.; White, P.C.; Parker, K.L.: Nucleotide sequence analysis of murine 21-hydroxylase genes: mutations affecting gene expression, Proc. natn. Acad. Sci. USA *83:* 9601–9605 (1986).

Cole, F.S.; Whitehead, A.S.; Auerbach, H.S.; Lint, T.; Zeity, H.J.; Kilbridge, P.; Colten, H.R.: The molecular basis for genetic deficiency of the second component of human complement. New Engl. J. Med. *313:* 11–16 (1985).

Dunham, I.; Sargent, C.A.; Trowsdale, J.; Campbell, R.D.: Molecular map of the human major histocompatibility complex by pulsed field gel electrophoresis. Proc. natn. Acad. Sci. USA *84:* 7237–7241 (1987).

Dupont, B.; Oberfield, S.E.; Smithwick, E.M.; Lee, T.D.; Levine, L.S.: Close genetic linkage between HLA and congenital adrenal hyperplasia (21-hydroxylase deficiency). Lancet *ii:* 1309 (1977).

Dupont, B.; Pollack, M.S.; Levine, L.S.; O'Neill, G.J.; Hawkins, B.R.; New, M.I.: Congenital adrenal hyperplasia. Joint report from the Eighth International Histocompatibility Workshop; in Terasaki, Histocompatibility testing, pp. 693–706 (UCLA Press, Los Angeles 1980).

Dykman, T.R.; Cole, J.L.; Iida, K.; Atkinson, J.P.: Polymorphism of human erythrocyte C3b/C4b receptor. Proc. natn. Acad. Sci. USA *80:* 1698–1702 (1983).

Dykman, T.R.; Hatch, J.A.; Aqua, M.S.; Atkinson, J.P.: Polymorphism of the C3b/C4b receptor (CR1): characterisation of a fourth allele. J. Immun. *134:* 1787–1789 (1985).

Fu, S.M.; Kunkel, H.G.; Brusman, H.P.; Allen, F.H.; Fotino, M.: Evidence for linkage between HL-A histocompatibility genes and those involved in the synthesis of the second component of complement. J. exp. Med. *140:* 1108–1110 (1974).

Giles, C.M.; Uring-Lambert, B.; Boksch, W.; Braun, M.; Goetz, J.; Neumann, R.; Mauff, G.; Hauptmann, G.: The study of a French family with two duplicated C4A haplotypes. Hum. Genet. *77:* 359–365 (1987).

Hardy, D.A.; Bell, J.I.; Long, E.O.; Lindsten, T.; McDevitt, H.O.: Mapping of the class II region of the human major histocompatibility complex by pulsed-field gel electrophoresis. Nature, Lond. *323:* 453–455 (1986).

Hauptmann, G.; Goetz, J.; Uring-Lambert, B.; Grosshans, E.: C4 deficiency. Prog. Allergy, vol. 39, pp. 232–249 (Karger, Basel 1986).

Hauptmann, G.; Grosshans, E.; Heid, E.; Mayer, S.; Basset, A.: Systemic lupus erythematosus and hereditary complement deficiency: a case with total C4 defect. Annls. Derm. Syph *101:* 479–496 (1974).

Hauptmann, G.; Uring-Lambert, B.; Vegnaduzzi-Lamouche, N.; Mascart-Lemone, F.: RFLP studies of 3 complete C4-deficient patients. Complement *4:* 166 (1987).

Higashi, Y.; Yoshioka, H.; Yamane, M.; Gotoh, O.; Fujii-Kuriyama, Y.: Complete nucleotide sequence of two steorid 21-hydroxylase genes tandemly arranged in human chromosome: a pseudogene and a genuine gene. Proc. natn. Acad. Sci. USA *83:* 2841–2845 (1986).

Hobart, M.: Phenotypic genetics of complement components. Phil. Trans. R. Soc. *306:* 325–331 (1984).

Hobart, M.J.; Lachmann, P.J.: Allotypes of complement components in man. Transplant. Rev. *32:* 26–42 (1976).

Holers, V.M.; Cole, J.L.; Lublin, D.M.; Seya, T.; Atkinson, J.P.: Human C3b- and C4b-regulatory proteins: a new multi-gene family. Immunol. Today *6:* 188–192 (1985).

Holers, V.M.; Chaplin, D.D.; Leykam, J.F.; Gruner, B.A.; Kumar, V.; Atkinson, J.P.: Human CR1 mRNA polymorphism correlates with the CR1 allelic molecular weight polymorphism. Proc. natn. Acad. Sci. USA *84:* 2459–2463 (1987).

Inoko, H.; Trowsdale, J.: Linkage of the TNF genes to the HLA-B locus. Nucl. Acids Res. *15:* 8957–8962 (1987).

Isenman, D.E.; Young, J.R.: The molecular basis for the difference in immune hemolysis activity of the Chido and Rodgers isotype of human complement compoment C4. J. Immun. *132:* 3019–3027 (1984).

Klickstein, L.B.; Rabson, L.D.; Wong, W.W.; Smith, J.A.; Fearon, D.T.: CR1 5′ cDNA sequences contain a fourth LHR and identify a new B cell-specific mRNA. Complement *4:* 180 (1987a).

Klickstein, L.B.; Wong, W.W.; Smith, J:A.; Weiss, J.H.; Wilson, J.G.; Fearon, D.T.: Human C3b/C4b receptor (CR1): Demonstration of long homologous repeating domains. J. exp. Med. *165:* 1095–1112 (1987b).

Kristensen, T.; D'Eustachio, P.; Ogata, R.T.; Chung, L.P.; Reid, K.B.M.; Tack, B.F.: The superfamily of C3b/C4b-binding proteins. Fed. Proc. *46:* 2463–2469 (1987).

Lachmann, P.J.: Inherited complement deficiencies. Phil. Trans. R. Soc. *306:* 419–430 (1984).

Lamm, L.U.; Olaisen, B.: Report of the committee on the genetic constitution of chromosome 6. Cytogenet. Cell. Genet. *40:* 128–155 (1985).

Law, S.K.A.; Dodds, A.W.; Porter, R.R.: A comparison of the properties of the two classes, C4A and C4B, of the human complement component C4. Eur. molec. Biol. Org. J. *30:* 1819–1823 (1984).

Law, S.K.A.; Reid, K.B.M.: Complement; in Focus series (IRL Press, Oxford, in press 1988).

Lawrance, S.K.; Smith, C.L.; Srivastava, R.; Cantor, C.R.; Weissman, S.M.: Megabase-scale mapping of the HLA gene complex by pulsed field gel electrophoresis. Science *235:* 1387–1390 (1987).

Lintin, S.J.; Lewin, A.; Reid, K.B.M.: Studies on the structure of the human C4b-binding protein gene. Complement *4:* 186 (1987).

Lintin, S.J.; Reid, J.B.M.: Studies on the structure of the human C4b-binding protein gene. FEBS Lett. *204:* 77–81 (1986).

McAleer, M.A.; Hauptmann, G.; Brai, M.; Misiano, G.; Sim, R.B.: Restriction fragment length studies for factor H. Complement *4:* 191 (1987).

Meo, T.; Atkinson, J.P.; Bernoco, M.; Bernoco, D.; Ceppellini, R.: Structural heterogeneity of C2 complement protein and its genetic variants in man: A new polymorphism of the HLA region. Proc. natn. Acad. Sci. USA *74:* 1672–1675 (1977).

Mole, J.E.; Anderson, J.K.; Davison, E.A.; Woods, D.E.: Complete primary structure of the zymogen of human complement factor B. J. biol. Chem. *259:* 3407–3412 (1984).

Morley, B.J.; Campbell, R.D.: Internal homologies of the Ba fragment from human complement component factor B, a class III MHC antigen. Eur. molec. Biol. Org. J. *3:* 153–157 (1984).

Müller, U.; Jongeneel, C.V.; Nedospasov, S.A.; Lindahl, K.F.; Steinmetz, M.: Tumour necrosis factor and lymphotoxin genes map close to H-2D in the mouse major histocompatibility complex. Nature, Lond. *325:* 265–267 (1987a).

Müller, U.; Stephan, D.; Phillippsen, P.; Steinmetz, M.: Orientation and molecular map position of the complement genes in the mouse MHC. Eur. molec. Biol. Org. J. *6:* 369–373 (1987b).

Müller-Eberhard, H.J.: The membrane attack com-

plex of complement. Annu. Rev. Immunol. *4:* 503–528 (1986).

Müller-Eberhard, H.J.; Miescher, P.A.: Complement (Springer, Heidelberg 1985).

Olaisen, B.; Teisberg, P.; Jonassen, R.; Thorsby, E.; Gedde-Dahl, T.: Gene order and gene distances in the HLA region studied by the haplotype method. Ann. hum. Genet. *47:* 285–292 (1983).

O'Neill, G.J.; Yang, S.Y.; Dupont, B.: Two HLA-linked loci controlling the fourth component of human complement. Proc. natn Acad. Sci USA *75:* 5165–5169 (1978a).

O'Neill, G.J.; Yang, S.Y.; Tegoli, J.; Berger, R.; Dupont, B.: Chido and Rodgers blood groups are distinct antigenic components of human complement C4. Nature, Lond. *273:* 668–670 (1978b).

Palsdottir, A.; Arnason, A.; Fossdal, R.; Jensson, O.: Gene organisation of haplotypes expressing two different C4A allotypes. Hum. Genet. *76:* 220–224 (1987a).

Palsdottir, A.; Fossdal, R.; Arnason, A.; Edwards, J.H.; Jensson, O.: Heterogeneity of human C4 gene size. Immunogenetics *25:* 299–304 (1987b).

Partanen, J.; Koskimies, S.: Human MHC class III genes, Bf and C4. Polymorphism, complotypes and association with MHC class I genes in the Finnish population. Hum. Hered. *36:* 269–275 (1986).

Ragoussis, J.; Bliek, A. van der; Trowsdale, J.; Ziegler, A.: Mapping of HLA genes using pulsed-field gradient electrophoresis. FEBS Lett. *204:* 1–4 (1986).

Raum, D.; Awdeh, S.L.; Anderson, J.; Strong, L.; Granados, J.; Pevan, L.; Giblett, E.; Yunis, E.J.; Alper, C.A.: Human C4 haplotypes with duplicated C4A or C4B. Am. J. hum. Genet. *36:* 72–79 (1984).

Reid, K.B.M.: Application of molecular cloning studies on the complement system. Immunology *55:* 185–196 (1985).

Reid, K.B.M.: Activation and control of the complement systems. Essays Biochem. *22:* 27–68 (1986).

Reid, K.B.M.; Bentley, D.R.; Campbell, R.D.; Chung, L.P.; Sim, R.B.; Kristensen, T.; Tack, B.F.: Complement system proteins which interact with C3b or C4b. Immunol. Today *7:* 230–233 (1986).

Reid, K.B.M.; Porter, R.R.: The proteolytic activation systems of complement. A. Rev. Biochem. *50:* 433–464 (1981).

Rey-Campos, J.; Rubinstein, P.; Rodriguez de Cordoba, S.: Mapping of DAF to the RCA gene cluster in humans. Complement *4:* 217 (1987).

Rey-Campos, J.; Rubinstein, P.; Rodriguez de Cordoba, S.: A physical map of the human regulator of complement activation gene cluster linking the complement genes CR1, CR2, DAF and C4BP. J. exp. Med. *167:* 664–669 (1988).

Rittner, C.; Giles, C.M.; Roos, M.L.H.; Demant, P.; Mollenhauer, E.: Genetics of human C4 polymorphism: detection and segregation of rare and duplicated haplotypes. Immunogenetics *19:* 321–323 (1984).

Rittner, C.; Hauptmann, G.; Gross-Wilde, H.; Grosshans, E.; Tongio, M.M.; Mayer, S.: Linkage between HL-A (major histocompatibility complex) and genes controlling synthesis of the fourth component of complement, in Histocompatibility testing, pp. 945–953 (Munksgaard, Copenhagen 1975).

Robinson, M.A.; Carroll, M.C.; Johnson, A.H.; Harzman, R.J.; Belt, K.T.; Kindt, T.J.: Localisation of C4 genes within the HLA complex by molecular genotyping. Immunogenetics *21:* 143–152 (1985).

Rodrigues, N.R.; Dunham, I.; Yu, C.Y.; Carroll, M.C.; Porter, R.R.; Campbell, R.D.: Molecular characterisation of the HLA-linked steroid 21-hydroxylase β gene from an individual with congenital adrenal hyperplasia Eur. molec. Biol. Org. J. *6:* 1653–1661 (1987).

Rodriguez de Cordoba, S.; Dykman, T.R.; Ginsberg-Fellner, F.; Ercilla, G.; Aqua, M.; Atkinson, J.P.; Rubinstein, P.: Evidence for linkage between the loci coding for the binding protein for the fourth component of human complement (C4BP) and for the C3b/C4b receptor. Proc. natn. Acad. Sci. USA *81:* 7890–7892 (1984).

Rodriguez de Cordoba, S.; Ferreira, A.; Nussenzweig, V.; Rubinstein, P.: Genetic polymorphism of human C4-binding protein. J. Immun. *131:* 1565–1569 (1983).

Rodriguez de Cordoba, S.; Lublin, D.M.; Rubinstein, P.; Atkinson, J.P.: Human genes for three complement components that regulate the activation of C3 are tightly linked. J. exp. Med. *161:* 1189–1195 (1985).

Rodriguez de Cordoba, S.; Rubinstein, P.: Genetic polymorphism of human factor H (βIH). J. Immun. *132:* 1906–1908 (1984).

Rodriguez de Cordoba, S.; Rubinstein, P.: Quantitative variations of the C3b/C4b receptor (CR1) in human erythrocytes are controlled by genes within the regulator of complement activation (RCA) gene cluster. J. exp. Med. *164:* 1274–1283 (1986).

Rodriguez de Cordoba, S.; Rubinstein, P.: New alleles of C4-binding protein and factor H. Immunogenetics *25:* 267–268 (1987).

Schendel, D.J.; O'Neill, G.; Wank, R.: MHC-linked class III genes. Analyses of C4 gene frequencies, complotypes and associations with distinct HLA haplotypes in German Caucasians. Immunogenetics *20:* 23–31 (1984).

Schneider, P.M.; Carroll, M.C.; Alper, C.A.; Rittner, C.; Whitehead, A.S.; Yunis, E.J.; Colten, H.R.: Polymorphism of human complement C4 and steroid 21-hydroxylase gene. Restriction fragment length polymorphisms revealing structural deletions, homoduplications, and size variants. J. clin. Invest. *78:* 650–657 (1986).

Schumaker, V.N.; Zavodsky, P.; Poon, P.H.: Activation of the first component of complement. Annu. Rev. Immunol. *5:* 21–42 (1987).

Schwartz, D.C.; Cantor, C.R.: Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell *37:* 67–75 (1984).

Sim, R.B.; Malhotra, V.; Ripoche, J.; Day, A.J.; Micklem, K.J.; Sim, E.: Complement receptors and related complement control proteins. Biochem. Soc. Symp. *51:* 83–96 (1986).

Singer, M.F.; Skowronski, J.: Making sense out of LINES: long interspersed repeat sequences in mammalian genomes. Trends biochem. Sci. *10:* 119–122 (1985).

Spies, T.; Morton, C.C.; Nedospasov, S.A:; Fiers, W.; Pious, D.; Strominger, J.L.: Genes for the tumor necrosis factor α and β are linked to the human major histocompatibility complex. Proc. natn. Acad. Sci. USA *83:* 8699–8702 (1986).

Stafford, H.A.; Tykocinski, M.L.; Holers, V.M.; Lublin, D.M.; Atkinson, J.P.; Medof, M.E.: Polymorphism of the DAF gene. Complement *4.* 227 (1987).

Strachan, T.: Molecular genetics and polymorphism of class I HLA antigens. Br. med. Bull. *43:* 1–14 (1987).

Teisberg, P.; Akesson, I.; Olaisen, B.; Gedde-Dahl, T., Jr.; Thorsby, E.: Genetic polymorphism of C4 in man and localization of a structural C4 locus to the HLA gene complex of chromosome 6. Nature, Lond. *264:* 253–254 (1976).

Teisberg, P.; Olaisen, B.; Jonassen, R.; Gedde-Dahl, T., Jr.; Thorsby, E.: The genetic polymorphism of the fourth component of human complement: Methodological aspects and a presentation of linkage and association data relevant to its localization in the HLA region. J. exp. Med. *146:* 1380–1389 (1977).

Trowsdale, J.: Genetics and polymorphism: class II antigens. Br. med. Bull. *43:* 15–36 (1987).

Uring-Lambert, B.; Goety, J.; Tongio, M.M.; Mayer, S.; Hauptmann, G.: C4 haplotypes with duplications at the C4A or C4B loci: frequency and associations with Bf, C2 and HLA-A, B, C, DR alleles. Tissue Antigens *24:* 70–72 (1984).

Weiss, J.J.; Fearon, D.T.; Klickstein, L.B.; Wong, W.W:; Richards, S.A.; deBruyn Kops, A.; Smith, J.A.; Weiss, J.H.: Identification of a partial cDNA clone for the C3d/Epstein-Barr virus receptor of human B lymphoytes. Proc. natn. Acad. Sci. USA *83:* 5639–5642 (1986).

Weiss, J.H.; Morton, C.C.; Bruns, G.A.P.; Weis, J.J.; Klickstein, L.B.; Wong, W.W.; Fearon, D.T.: A complement receptor locus: genes encoding C3b/C4b receptor and C3d/Epstein-Barr virus receptor map to 1q32. J. Immun. *138:* 312–315 (1987).

White, P.C.; Grossberger, D.; Onufer, B.J.; Chaplin, D.D.; New M.I.; Dupont, B.; Strominger, J.L.: Two genes encoding steroid 21-hydroxylase are located near the genes encoding the fourth component of complement in man. Proc. natn. Acad. Sci. USA *82:* 1089–1093 (1985).

White, P.C.; New, M.I.; Dupont, B.: HLA-linked congenital adrenal hyperplasia results from a defective gene encoding a cytochrome P-450. Proc. natn. Acad. Sci. USA *81:* 7505–7509 (1984).

White, P.C.; New, M.I.; Dupont, B.: Structure of human steroid 21-hydroxylase genes. Proc. natn. Acad. Sci. USA *83:* 5111–5115 (1986).

Whitehead, A.S.; Colten, H.R.; Chang, C.C.; Demars, R.: Localisation of the human MHC-linked complement genes between HLA-B and HLA-DR by using HLA mutant cell lines. J. Immun. *134:* 641–643 (1985).

Wong, W.; Cahill, J.; Kennedy, C.; Bonaccio, E.; Wilson, J.; Klickstein, L.; Ralson, L.; Fearon, D.: Analysis of genomic polymorphism in the human CR1 gene. Complement *4:* 240 (1987).

Wong, W.W.; Wilson, J.G.; Fearon, D.T.: Genetic regulation of a structural polymorphism of human C3b receptor. J. clin. Invest. *72:* 685–693 (1983).

Wu, L.C.; Morley, B.J.; Campbell, R.D.: Expression of the human complement protein factor B gene: evidence for the role of two distinct 5′ flanking elements. Cell *48:* 331–342 (1987).

Yu, C.Y.; Belt, K.T.; Giles, C.M.; Campbell, R.D.; Porter, R.R.: Structural basis of the polymorphism of human complement components C4A and C4B: gene size, reactivity and antigenicity. Eur. molec. Biol. Org. J. *5:* 2873–2881 (1986).

Yu, C.Y.; Campbell, R.D.: Definitive RFLPs to distinguish between the human complement C4A/C4B isotypes and the major Rodgers/Chido determinants. Application to the study of C4 null alleles. Immunogenetics *25:* 383–390 (1987).

Yunis, E.J.; Awdeh, Z.; Johnson, A.; Suciu-Foca, N.; Robinson, M.A.; Hartzman, R.; Raum, D.; Fleischnick, E.; Alper, C.A.: Complotype genetic loci segregate more frequently with HLA-DR than with HLA-B. Immunogenetics *21:* 25–31 (1985).

R.D. Campbell
MRC Immunochemistry Unit
Department of Biochemistry
University of Oxford
South Parks Road
Oxford OX1 3QU (UK)