# CHARACTERIZATION OF MOUSE MAJOR URINARY PROTEIN GENES

RAYA AL-SHAWI

Thesis presented for the degree of Doctor of Philosophy

University of Edinburgh

1985

To my parents.

I declare that this work is my own, except where otherwise stated.

Raya Al-Shawi

ACKNOWLEDGEMENTS

First of all I would like to express special thanks to my supervisor, John Bishop, for quantities of advice, admirable patience, and for many constructive and stimulating discussions.

Heart-felt thanks go to all my colleagues in the Bishop Laboratory, and especially to John Clark, for his help during the first months of the project and for his interest throughout. Special thanks are also due to Alistair Chave-Cox, Don Anson and Peter Ghazal for enthusiastically discussing their results with me, and to Morag Robertson and Melville Richardson for introducing me to many techniques. I am also grateful to the work-shop, and to George Newall in particular, for maintaining and creating apparatus.

I have had useful discussions with Frank Wright, Bill Held, Noreen Murray, David Leach, Joe Felsenstein and Graham Bullfield, and I greatly appreciate their suggestions and interest.

The manuscript was cheerfully and efficiently typed by Antonia Lovelace, Neil Robertson, Hiyan Al-Shawi, Iain McIntosh and Jackie Bogie, and the photographs were skillfully prepared by Frank Johnson and Alan McEwan. To all of them I am truely grateful.

Finally, I would like to thank David, Marwan, Hiyan, Neil, Antonia and my contemporaries at 42 Marchmont Rd. for their moral support.

## Abstract

Major urinary protein (MUP) genes were isolated from C57 genomic libraries, characterized by restriction enzyme mapping and compared with MUP genes isolated from BALB/c genomic libraries (Clark et al, 1982; Bishop et al, 1982). The conclusions drawn from the characterization of this new set of MUP genes are in agreement with those previously drawn from studies on the BALB/c MUP genes.

Most MUP genes were found to share extensive homology in their transcription units and 5' and 3' flanking regions. Exceptions were those genes whose coding regions have been interupted by insertions and/or deletions. The MUP genes fall into two main groups based on hybridization criteria: group 1 and group 2 (Bishop et al, 1982). With the exception of one group 2 gene (BL-25/CL-2), restriction site homology was found to be greater within groups than between them. Restriction site homologies further divided the group 1 genes into two sub-groups. Sequence data revealed that the two sub-groups have different forms of an A-rich region located ∿1Øbp upstream of the TATA box.

Messenger RNA from tissues that express MUP was shown to be more homologous to group 1 coding sequences than to group 2 coding sequences. In the liver, two forms of MUP mRNA can be distinguished. Group 1 sequences hybridized preferentially to the abundantly transcribed long form of the mRNA, while group 2 sequences hybridized preferentially to the short and rarer form of the mRNA.

Genomic digests illustrated that two types of variation are found between the MUP genes of BALB/c and C57BL/Fa mice. The first relates to the presence of variant restriction fragments. Two cloned MUP genes carrying such fragments were identified. The second relates to variation in the intensity of common restriction fragments. Differences between the strains in the total number of MUP genes were not observed. Variation in the intensity of common restriction fragments are proposed to be the result of different homogenization events that took place in the mouse lineages from which BALB/c and C57BL/Fa were derived.

# Contents

# Tables

F i g u r e s

Results

## The structure of genes transcribed by RNA polymerase II

Many eukaryotic genes contain two distinct types of region, exons and introns. Exons encode the translated sequences as well as 5' and 3' flanking untranslated sequences of the messenger RNA. Introns are sequences present within the transcription unit but not represented in the messenger RNA. Both exons and introns are transcribed. However, the introns are excised or "spliced out" from the newly transcribed RNA and adjacent exons are joined to form a continuous messenger RNA.

Comparison of homologous genes from different species, and comparison of closely related genes within a single species, reveal that the size and coding sequences of exons are well conserved. With the exception of certain discrete regions, the non-coding sequences of exons are less conserved than the coding sequences, and appear to evolve at a similar rate to the silent site divergence of coding sequences (Perler et al, 1980). Introns can undergo considerable size changes during evolution due to insertions, deletions and duplication events (Konkel et al, 1979; Efstratiadis et al, 1980). Where they can be aligned, homologous intronic sequences of the β-globin genes have been found to evolve at a similar rate to the silent site divergence of the coding sequences (Efstratiadis et al, 1980).

Some genes may be alternatively spliced with the result that

intronic sequences of one transcript become part of the translated sequences of another transcript. Splicing may also give rise to differences in the untranslated regions of the mRNA. This is usually associated with the use of alternative polyadenylation or transcription initiation sites, sites which define the beginning and end of the mRNA respectively. Examples of alternatively spliced transcripts are: the αA and $\alpha A^{ins}$ transcripts of the murine αA crystalline gene, which differ in their translated sequences (King and Piatigorsky, 1983); the alkali light myosin gene transcripts, LCI and LC3, which differ in both their 5' untranslated and their translated sequences (Nabeshima et al, 1984); the transcripts coding for the membrane and secreted forms of IgM, which differ in both their 3' untranslated and their translated sequences (Alt et al, 1982) and the mouse α-amylase salivary and main liver transcripts which differ only in their 5' untranslated sequences (Hagenbuchle et al, 1981).

The function of intronic sequences that are not represented in mature mRNA appears to vary. For example, the immunoglobin heavy chain genes and the mature K light chain gene carry a tissue specific transcriptional enhancer in the intron that separates the joining region from the adjacent constant region (Banerji et al, 1983; Queen and Baltimore, 1983; Picard and Schaffner, 1984). However, some introns might not carry sequences with essential functional roles. Thus natural loss of the second intron of the rat preproinsulin I gene has not affected its regulated expression (Lomedico et al, 1979), as judged from the rates of synthesis of insulin I and insulin II in isolated islets of Langerhans (Clark and

Steiner, 1969 ). Also, not all genes transcribed by RNA polymerase II have introns. Known examples of such intronless genes are those that code for histones (Hentschel and Birnstiel, 1983), interferons (Nagata et al, 1980) and protamines (Anson, 1983). The reasons for the evolution of interrupted genes have therefore been the cause of debate.

Some introns appear to separate exons that code for different functional domains. Examples are the introns which define the sequences coding for the haem-binding domain of globin (Craik et al, 1980), the introns which separate immunoglobulin constant and hinge regions (Sakano et al, 1979; Early et al, 1979) and the intron which interrupts the C peptide that connects the A and B chains of mature insulin (Perler et al, 1980). This has led to the theory that novel transcription units may be formed by the bringing together of different functional sequences. It is hypothesised that such "exon shuffling" is made feasible by the separation of coding sequences by introns (Gilbert, 1979; Blake, 1979). Although exon shuffling takes place in the immunoglobulin producing cells (Maki et al, 1980), evidence for exon-shuffling as an important mechanism in the evolution of eukaryotic genes has only recently been provided. This support comes from unravelling the structure of the low-density lipoprotein (LDL) receptor gene (Sudhof et al, 1985a; Sudhof et al, 1985b). The LDL receptor gene has been found to show a clear association between exons and functional domains, and different domains have been found to share homology with the epidermal growth factor (EGF) precursor gene and the complement factor C9 gene. Moreover, the positions at which the introns

interrupt the genes are conserved between the homologous regions. Sudhof et al have therefore concluded that the LDL receptor gene is composed of a "mosaic of exons" derived from different genes.

Introns occasionally separate domains which are functionally related and which have a common evolutionary ancestor. Examples are the three functional domains of the ovomucoid gene, each capable of binding and inhibiting a serine protease (Stein et al, 1980) and the helical domain units of the collagen genes (Yamada et al, 1980; 1984; Chu et al, 1984). The predominant exon sizes of vertebrate collagen genes (54, 45, 99, 108 and 162 bp respectively) has led to the suggestion that these genes evolved by two types of recombinational events involving a 54 bp unit: (1) unequal crossing-over between the 54 bp units, (2) recombination between the genes and their transcribed sequences or cDNAs. Exon duplication may also be an important mechanism in the evolution of genes coding for new products. Thus the homologous region between the LDL precursor and C9 genes is represented seven times in the former gene.

## Cis acting regulatory sequences of genes transcribed by RNA polymerase II

Sequences involved in transcription initiation, transcription regulation and mRNA processing have been identified using both cell-free systems and cultured cells. Different cell-free systems that accurately carry out transcription initiation have been developed from mammalian cell extracts (Weil et al, 1979; Manley et al, 1980; Krainer et al, 1984) Reactions that appear to be coupled

*in vivo* may be uncoupled in different cell-free systems, thus allowing the molecular mechanisms of different transcription and RNA processing steps to be studied. A disadvantage of cell-free systems is that certain factors may be missing that are required *in vivo*. This may partly explain the occasional discrepancies obtained between cell-free systems and cultured cells (Benoist and Chambon, 1981; Mathis and Chambon, 1981).

Several SV40 vectors and a few other viral vectors (for example those based on adenovirus, BPV and retroviruses) have been developed for the introduction of cloned DNA into cells (see Gluzmann, 1983). These vectors have been used for both transient expression studies and long-term expression studies. In transient expression studies, the infecting DNA may remain epigenic, while in long-term expression studies, with the exception of some BPV based vectors (DiMaio and Maniatis, 1982; Matthias *et al*, 1983), the infecting DNA becomes integrated into the genome. Long term studies require continuous selection for cells harbouring the introduced DNA. For this reason, many cell lines which lack functional marker genes (e.g. *tk*, *hprt*) have been developed. The marker gene usually forms part of the vector, although it may be co-transformed with the vector (Pellicer *et al*, 1980). Transient expression studies have two main advantages over long-term studies. The first is that inhibition of expression due to re-arrangements caused by integration of the transforming DNA into the genome are more likely to be avoided in a mixed population of transformed cells. The second is that studies are not restricted to certain cell lines and may even be carried out using primary cell cultures. Their disadvantages are the low levels

of expression obtained and the short-term within which expression can be studied.

Much of our understanding of gene expression is derived from studies in cell culture. It must therefore be emphasized that these represent artificial systems which may occasionally give rise to artifactual results. This is borne out by the observation that cloned genes introduced into cultured cells are often expressed even though the cells are derived from tissues in which these same genes are normally inactive.

Comparison of sequence data from many genes transcribed by RNA polymerase II has revealed certain well-conserved regions that are present within or flanking the transcription units. Deletion mutation studies have been extensively used to identify whether such regions are necessary for transcription and RNA processing. Some artifacts may be obtained in deletion mutation studies due to the altered spatial relationships brought about by the deletions. To avoid such artifacts, McKnight and Kingsbury (1982) studied the effect of deletions within the HSV-tk promoter region by elegantly replacing the deleted HSV-tk sequences with linker DNA. However, deletion replacement studies may also give rise to artifacts due to the introduction of novel sequences within the regions under investigation. For these reasons point mutation studies serve to complement deletion mutation studies, when determining the functional roles of specific sequences.

Fusion genes have often been used to demonstrate the effect of

regulatory sequences that have been removed from their normal gene environment. In the most thorougly conducted studies, short sequences found to be important in the regulation of transcription have been artificially synthesized. The synthetic oligonucleotides are fused to the coding regions of genes which naturally lack these sequences; studies on the expression of the resultant fusion genes are then carried out and may provide final proof for the functional significance of the proposed regulatory sequences.

Sequences required for splicing. Despite the variability in the structure of introns, they share a common feature. With few exceptions, all begin with the dinucleotide 5' GT 3' and end with the dinucleotide 5' AG 3' (Breathnach et al, 1978). A consensus sequence compiled from ∿130 exon and intron boundaries yields the sequence (CA)AG/GT(AG)AGT for the 5' junctions and (TC)nN(CT)AG/G for the 3' junctions (Mount, 1982). These sequences are thought to be important for splicing since mutation of some positions abolishes correct splicing.

The effect of point mutations on the splice junctions of the large intron of the rabbit β-globin gene has been studied in cell culture (Wieringa et al, 1983). This was achieved by infecting HeLa cells with a vector containing the modified β-globin gene linked to the SV40 transcription promoter. Mutation of the G residue at the first intronic position was found to abolish normal splicing and to result in the use of cryptic sites. Naturally-occuring mutations at this position have also been described. The splicing products of a

β-thalassaemia gene carrying such a mutation have been studied *in vitro* in cultured cells. As in the deletion mutation studies, processing of the transcripts involved the use of a cryptic splice site. Moreover, low amounts of mRNA were produced in which the first exon was directly spliced to the third exon (Treisman *et al*, 1982).

Mutations that involve intronic sequences other than the splicing consensus sequences, and that lead to the creation of new splicing patterns, have also been described. A striking example of this is provided by a β-thalassaemia gene which carries a mutation at position 745 of the second intron. Studies in cell culture showed that this gene produced mRNA with an extra exon: the β-mutation was found to create a new 5' splice site which in turn leads to the activation of an internal 3' cryptic splice site. Hence, a 162 bp exon was produced between exons 2 and 3 (Treisman *et al*, 1983). As well as leading to abnormal splicing, mutations at and around the splice junctions have been found to reduce splicing efficiency. This is reflected by the accumulation of unprocessed and semi-processed mRNAs (Busslinger *et al*, 1981).

The molecular mechanism by which splicing occurs is not fully understood. It has been proposed that the association of the small nuclear RNA U1 with the splicing sites is necessary for splicing. This is based on the following observations: (1) the association of U1 with hnRNA; (2) sequence complementarity between U1 and both splice sites (Roger and Wall, 1980; Lerner *et al*, 1980); (3) the inhibition of splicing *in vitro* when anti-(U1) RNP serum is

added to whole cell extracts normally capable of accurate splicing (Lerner and Steitz, 1979; Yang et al, 1981; Padgett et al, 1983), and (4) preferential binding of U1 snRNPs to a 5' splice site in vitro (Mount et al, 1983). Recently, analysis of splicing intermediates in cell-free systems that process exogenously added labelled RNA (Padgett et al, 1984; Krainer et al, 1984) has indicated that introns are excised in the form of lariats (Grabowski et al, 1984; Padgett et al, 1984; Ruskin et al, 1984). A lariat is formed by the joining of the 5' end of an intron to a site close to its 3' end through a 2'-5' phosphodiester bond. The role of U1 RNA in the formation of lariat structures is not known.

The effect of extensive deletions within the large rabbit β-globin intron has also been studied in culture cells (Wieringa et al, 1984). With deletions starting from the centre of the intron and extending in either a 5' or 3' direction, it was found that a minimum of the first six 5' base pairs of the intron or the last twelve 3' base pairs of the intron were required for efficient splicing. When deletions were extended in both directions, efficient splicing was carried out if the first six 5' base pairs and the last twenty-three 3' base pairs were separated by a minimum of 60 bp of heterologous DNA. From this it is concluded that while specific sequences at the 5' and 3' junctions are required for splicing, no strictly conserved internal intronic sequences are required, although a minimum sequence length does appear to be important.

Because the 2'-5' phosphodiester internal bond of the lariat structure is positioned 5' to the conserved 3' intron/exon junction,

it also appears that the formation of the lariat structure does not require highly conserved internal sequences (Wieringa et al, 1984; Weissmann, 1984). Langford et al (1984) studied the effect of deletion mutagenesis on the splicing of yeast transcripts in yeast cells. This was achieved by transforming yeast cells with a yeast-Acanthamoeba hybrid actin gene, the transcripts of which could be distinguished from the wild type actin transcripts. Modification of the fusion gene, by a deletion between 35 - 7Ø base pairs 5' to the 3' end of the yeast actin intron, was found to abolish splicing. Comparison of the deleted region with the sequences of sixteen other yeast introns revealed the presence of the sequence 5'TACTAAC3'. This sequence was found to be positioned approximately 1Ø to 5Ø bp from the 3' exon-intron boundary.

Sequences similar to 5'TACTAAC3' have been found in the 3' regions of higher eukaryotic introns (Keller and Noon,1984). In the large rabbit β-globin intron this sequence coincides with the position where the 5' end of the intron is joined to an internal 3' residue (Wieringa et al, 1984). However, from comparisons of the 5'TACTAAC3'-like sequences in the 3'regions of the introns of higher eukaryotes (Keller and Noon, 1984), and from the deletion mutation studies of Langford et al (1984) and Wieringa et al (1984), it appears that the nucleotide composition of internal intronic sequences used in splicing are less rigidly conserved in higher eukaryotic genes (Wieringa et al, 1984; Weissmann, 1984).

The cap site. The 5' termini of eukaryotic mRNAs are modified with a "cap" structure $M^7G(5')ppp(5')N$. The consensus sequence

PyA(Py)5 for the beginning of the mRNA at the cap site has been derived from a comparison of 22 genes (Breathnach and Chambon, 1981). The cap site is also thought to correspond to the position of transcription initiation. Evidence for this comes from SI mapping of precursor and mature mRNA transcripts (Weaver and Weissmann, 1979; Breathnach and Chambon, 1981). Further confirmation has come from the work of Bunick et al (1982) who used nucleotide triphosphate analogs with phosphohydrolase-resistant beta-gamma phosphate bonds to study cap site formation during the in vitro transcription of the adenovirus Ad2EIV gene. Fingerprint analysis of the gel-purified transcripts with 5' triphosphate ends revealed that the 5' terminal nucleotide triphosphate analog was present in the same oligonucleotide as the cap.

Recently Konarska et al, (1984) have suggested that the cap structure plays an important role in splicing, since the addition of cap analogs to HeLa whole-cell extracts inhibits in vitro splicing of mRNA. The exact role of the cap structure in splicing is not known, although it does not appear to be related to mRNA stability.

Sequences defining the end of the transcription unit and the end of transcription. Most mature transcripts of RNA polymerase II terminate in a poly(A)-tail. The poly(A)-tail is added at a distance up to 30 bases 3' to the sequence AAUAAA (or a closely related version of this sequence) which is found at the 3' end of the mRNA (Proudfoot and Brownlee, 1976). The sequence AAUAAA appears to be important in determining the correct position of polyadenylation.

Deletion of the AATAAA sequence in the SV40 late transcription unit was found to prevent polyadenylation. However, deletion of the DNA between the AATAAA sequence and the wild type polyadenylation site caused the addition of A residues to take place at a site further downstream, thereby maintaining a similar distance between the AAUAAA sequence and the poly(A)-tail to that found in wild type transcripts (Fitzgerald and Shenk, 1981).

Conservation of the sequence AATAAA appears to be necessary for defining the end of the mRNA. Thus point mutation of the AATAAA sequence of the adenovirus early transcription unit to AAGAAA, has been found to greatly reduce the levels of wild-type mRNA. Most transcripts observed extended into the next transcription unit, although in the few that were correctly cleaved, the accuracy of polyadenylation was found to be unaffected (Montell et al, 1983).

Recently Gil and Proudfoot (1984) have identified other regions adjacent to the AATAAA sequence which may also be required for the formation of correct β-globin mRNA termini. They constructed a series of deletions 3' to the AATAAA sequence of the rabbit β-globin gene. The deleted β-globin genes were cloned into a SV40-pBR328 expression vector and their transient expression was assayed in HeLa cells. These studies showed that a ∿35bp deletion which had been positioned 15 bp 3' to the AATAAA sequence, abolished the use of the AATAAA sequence present at the wild type position. Instead, a second AATAAA β-globin sequence was used. This second AATAAA sequence had been introduced further downstream in the vector, and was flanked on its 3'side by 355 bp of β-globin 3'-flanking

sequence. The deleted 35 bp sequence which appears to be required for the formation of correct β-globin termini, contains a G+T rich sequence as well as the sequence CAYUG. This latter sequence has been found at a similar position in other eukaryotic genes (Berget, 1984; Gil and Proudfoot, 1984).

Some genes have more than one AATAAA sequence at the end of their transcription units which are used alternatively. Examples are the mouse dihydrofolate reductase gene (Setzer et al, 1980; 1982) and the ovalbumin gene (Le Meir et al, 1984).

The point of transcription termination in higher eukaryotic genes that are transcribed by RNA polymerase II, is not well defined, and the nature of the sequences that are involved in this process are not known. Hofer et al (1982) found that elongation of labelled RNA in vitro from isolated nuclei did not extend beyond 1.3 kbp 3'to the poly(A) site of the mouse β-globin major gene. Further studies, using SI mapping, showed that transcription terminated at a discrete position located ∿1000 bases downstream of the poly(A) addition site (Salditt-Georgiff and Darnell, 1983). A different situation was observed when similar studies were performed on the $\alpha_2$ amylase gene. In this case transcription was found to terminate at several sites located 2.5 - 4 kbp 3' to the polyadenylation site (Hagenbuchle et al, 1984). Whether one or both of these situations are common to eukaryotic genes with polyadenylated transcripts awaits further studies.

The 3' termini of histone genes and yeast genes differ from those of

other eukaryotic genes that are transcribed by RNA polymerase II. Histone gene transcripts are unique in lacking both a poly(A)-tail as well as a 3' AAUAAA sequence. Transcription termination of the sea urchin H2A gene appears to be dependent on a well conserved palindrome and a short sequence ACCA found in the 3' untranslated region of the mRNA as well as ~80 base pairs of 3' non-transcribed spacer DNA (Birchmeier et al, 1983). Yeast transcripts also lack the sequence AAUAAA although they are polyadenylated. A sequence TTTTTATA appears to be important in determining the end of transcription as well as the site of polyadenylation of some yeast genes transcribed by RNA polymerase II (Henikoff et al, 1983) but it is not ubiquitously found at the 3' ends of all yeast genes. In conclusion it appears that sequences defining the end of transcription for RNA polymerase II have not been rigidly conserved.

The TATA box. Most genes transcribed by RNA polymerase II have a 7 bp sequence known as the 'TATA box', located at about 30 base pairs 5' to the mRNA cap site. A comparison of 60 eukaryotic genes yielded a consensus sequence TATA(AT)A for the TATA box (Breathnach and Chambon, 1981). In vitro deletion mutation studies involving the TATA box usually result in transcription initiating at many points besides the cap site. This has led to the conclusion that the TATA box promotes specific transcription initiation. Deletion of the TATA box or point mutations within the TATA box have also been found to result in a marked reduction in transcription efficiency. The SV40 early region promoter seems to be an exception, for although deletion of its TATA box results in multiple initiations, the transcriptional efficiency is not affected. This

difference may be related to the fact that the SV40 TATA box is unusual in that it is immediately preceded by the AT-rich sequence TAATTTTTTT.

Grosveld et al (1982) studied the effect of 5' flanking deletions on the expression of the rabbit β-globin gene by transforming HeLa cells with a vector containing the modified β-globin gene coupled to the SV40 virus transcriptional enhancer. They found that deletion of the TATA box reduced transcriptional efficiency and resulted in the use of multiple initiation sites. These results contrast with those of Dierks et al (1983) who also studied the effect of 5' deletions on the expression of the rabbit β-globin gene, this time by infecting mouse 3T6 cells with a vector containing the modified β-globin gene coupled to the polyoma virus transcriptional enhancer. This latter group found that although deletion of the TATA box reduced transcription efficiency, the transcripts were initiated at the wild type cap site.

It is possible that the differences between the two groups are due to differences in the expression systems used (e.g. between the polyoma and SV40 enhancers, see de Villiers et al, 1982). However, in both groups deletions upstream of the TATA box allowed the initiation of transcription at the wild type cap site. Also, in both groups, deletions downstream of the TATA box displaced transcription initiation in the 3' direction by a similar number of nucleotides as the deletions. Therefore, the differences are probably due to: (1) differences in the extent of the deletions; and (2) differences in sequence composition introduced by constructing

the deletions. The discrepancies in the results pinpoint some of the disadvantages of deletion mutation studies, which are best interpreted in conjunction with results obtained from point mutation studies.

The CAAT box. A region of homology shared between many genes transcribed by RNA polymerase II is located 70 to 80 base pairs 5' to the cap site. This region is known as the 'CAAT box" and has the consensus sequence 5'-GG(CT)CAATCT-3' (Efstratiadis et al, 1980; Benoist et al, 1980; and Breathnach and Chambon, 1980). By deletion mutation studies on the 5' flanking region of the rabbit β-globin gene in HeLa cells, Grosveld et al (1982) showed that deletion of the CAAT box did not affect transcription specificity although it did lead to a reduction in transcription efficiency. In contrast, deletion of the histone H2A gene CAAT box was found to stimulate this gene's rate of transcription in Xenopus oocytes (Grosschedl and Birnsteil, 1980). The observed difference could be an artifact of the deletion constructs and the different cellular systems used, or could reflect a genuine difference in the functional role of the CAAT box in these two genes.

The 21bp repeat of SV40. A G+C rich sequence, composed of two perfect 21 bp direct repeats and one imperfect direct repeat is located ∿80 bp 5' to the SV40 early transcription unit. This sequence is important for SV40 viability and controls the rate of transcription from the early transcription unit. The 21 base pair repeat forms the in vitro binding site of Sp1, a transcriptional factor required for accurate transcription (Dynan and Tijan, 1983;

Gidoni et al, 1984). Recently, sequence data on the 5' flanking sequence of the 'housekeeping' genes HMG CoA reductase and hypoxanthine phosphoribosyl transferase (hprt) has revealed G+C rich sequences in their promoter regions. These sequences show homology to the SV40 21 bp repeats (Reynolds et al, 1984; Melton et al, 1984), suggesting that such G+C rich sequences may be important promoter elements in a number of cellular genes. However, the functional significance of the G+C rich regions in the HMG CoA reductase and hprt genes is yet to be demonstrated.

Enhancers. Enhancer elements are sequences that are able to enhance transcription initiation when placed in either orientation and in a number of different positions relative to the transcription unit. The most extensively studied of these elements are the two 72 base pair repeats that are present between the early and late transcription units of the SV40 virus. These sequences have been found to retain an enhancing ability when isolated from their viral background and placed in cis relative to other genes. The first such study by Banerji et al (1981) showed that when the rabbit β-globin gene was linked to the SV40 enhancer element, correct transcription initiation of this gene was enhanced by two orders of magnitude. Since then, the SV40 enhancer element has been widely used in expression vectors to enhance the transient expression of cellular genes in cultured cells. Enhancers have been found in other viruses e.g. polyoma, BPV, ASV, adenovirus and many retroviruses (see Gluzman and Shenk, 1983). The retroviral enhancers have come under special scrutiny due to their role in activating cellular proto-oncogenes.

A few cellular enhancer and enhancer-like elements have also been discovered, the best characterized of the former being the enhancer founnd in the first intron of some immunoglobulin (Ig) genes (Banerji et al, 1983; Queen and Baltimore, 1983; Picard and Schaffner, 1984). Studies on the mouse Ig heavy chain gene enhancer (Banerji et al, 1983), showed that this element could enhance the transient expression of the rabbit β-globin gene in mouse myeloma cells when placed 500-2000 base pairs 5' or 3' to the gene. Like the viral enhancers, it was also found to be orientation independent.

Some enhancers appear to show species or tissue specificity, in that transcription is more efficiently enhanced in cell lines derived from the tissues in which the enhancers are normally active. Thus the SV40 enhancer has been found to work more effectively in primate derived cell lines than in mouse derived cell lines (de Villiers et al, 1982; Laimins et al, 1982), while the mouse immunoglobulin heavy chain gene enhancer has been found to enhance transcription in a B-lymphocyte derived cell line but not in HeLa cells (Banerji et al, 1983). Although enhancers do not show extensive sequence homology, a potential consensus core sequence, TGGTT, has been drawn up from 16 enhancer and potential enhancer elements (Laimins et al, 1983).

Regions that affect the rate of transcription initiation have been identified in some cellular genes through deletion-mutation studies. Some of these regions contain or overlap with sequences similar to the consensus enhancer core sequence and may themselves turn out to

be enhancers. Deletion-mutation studies have identified such regions ∿1ØØ base pairs 5' to the transcription units of the HSV-tk and the rabbit β-globin genes (McKnight and Kingsbury, 1982; Grosveld et al, 1982; Laimins et al, 1983). Both these regions contain a C-rich sequence: CC(CT)C(GA)CCC(CT)G. This sequence is similar to the in vitro Spl binding sites (CT)(CT)CCGCCC present in the 21 base pair repeats of SV4Ø (see Dierks et al, 1983). The significance of this limited sequence homology is not known.

Steroid hormone receptor binding sites. Other sequences which influence transcription initiation and which are shared by evolutionarily unrelated genes are those that bind steroid hormone receptors. Of these, the glucocorticoid receptor binding sites have been the most extensively studied. Success in this area largely stems from the discovery that transcription of a natural vector, the mouse mammary tumor virus (MMTV), is glucocorticoid inducible, and from significant improvements in the quality of rat liver glucocorticoid receptor preparations.

Deletion mutation studies on MMTV have suggested that sequences within the LTR are able to enhance in vitro transcription initiation on dexamethasone administration (Lee et al, 1981; Hynes et al, 1983; Pfahl et al, 1983; Chandler et al, 1983). Receptor filter binding studies, DNase I footprinting studies and methylation protection studies (Payvar et al, 1981; Pfahl, 1982; Payvar et al, 1983; Scheidereit et al, 1983) have identified glucocorticoid binding sequences within the MMTV LTR which overlap with sequences found to be important for glucocorticoid regulation

by deletion mutation studies in cultured cells. Taken together, these results suggest that functional glucocorticoid receptor binding sites are present within the LTR of MMTV. *In vitro* glucocoticoid receptor binding sites have also been found in MMTV outside the LTR. The functional significance of these sites is as yet unknown.

Chandler *et al* (1983) studied the dexamethasone induced expression of the *tk* gene in Rat XC*tk-* cells when MMTV LTR sequences were fused 5' to the *tk* promoter region. Detailed studies on an LTR region which contains an *in vitro* glucocorticoid receptor binding site, showed that the functioning of this sequence was relatively unconstrained positionally and was orientation independent. This has led to the suggestion that glucocorticoid response elements may be akin to enhancer elements.

Sequences involved in the glucocorticoid regulation of a cellular gene, the human metallothionein II (hMTIIA) gene, have also been extensively studied. Karin *et al* (1984) constructed a fusion gene consisting of ∿800bp of 5' flanking and untranslated hMTII sequences linked to the HSV *tk* transcription unit. They then compared the transformation efficiency of Rat 2 *tk-* cells infected with the 5' hMTIIA/*tk* fusion gene with that of Rat 2 *tk-* cells infected with 5' deletion variants of the same fusion gene.

It was found that sequences lying between the 5' untranslated region and -268 of the hMTII gene were necessary for glucocorticoid inducibility. DNaseI and methylation protection studies identified

a glucocorticoid receptor binding site within this region at ∿ - 250. These results suggest that a functional glucocorticoid receptor binding site is present at ∿-250bp 5' to the cap site of this cellular gene.

Comparison of the *in vitro* hMIIA gene and MMTV LTR glucocorticoid binding sites gives the consensus 5'TGGTACAAATGTTCT3' (Karin *et al*,1984). This sequence contains the hexamer 5'TGTTCT3' found in other *in vitro* identified glucocorticoid receptor binding sites (Scheidereit *et al*, 1983; Renkawitz *et al*, 1984). However, it is not found in all potential *in vivo* glucocorticoid receptor binding elements. Thus the 5' flanking region of the lysozyme gene, between nucleotides -168 and -203, that is required for dexamethasone and progesterone induction, binds the glucocorticoid receptor only weakly and does not contain the above hexanucleotide (Renkawitz *et al*, 1984). A sequence located between -30 and -74 of the lysozyme gene contains the hexanucleotide 5'TGTTCT3' in the antisense strand and strongly binds the glucorticoid receptor (Renkawitz *et al*, 1984). It is therefore possible that the weak and strong *in vitro* lyoszyme glucocorticoid receptor binding sites have different *in vivo* roles.

Regions involved in the regulated expression of genes by other inducing factors have been identified in several genes. Examples are: the upstream sequences required for heavy metal induction of the mouse and human metallothionein genes (Searle *et al*,1984, Karin *et al*, 1984), the upstream sequences that regulate the expression of the rat prolactin gene by EGF and phorbol esters

(Supowit et al, 1984) and the sequence responsible for heat-shock induction of the Drosophila heat-shock genes (Pelham, 1982; Pelham and Bienz, 1982).

This last example has been the most thoroughly characterized to date. The sequence that confers heat-shock inducibility was originally defined by deletion mutation studies on the Hsp70 heat-shock gene. Sequences of similar nucleotide composition were found to occur 5' to the transcription units of other Drosophila heat-shock genes, and a consensus sequence CTgGAAtnTTCtAGa was derived for the heat-shock response element (Pelham, 1982). Pelham and Bienz,(1982) synthesized oligonucleotide sequences similar to the consensus sequence of the heat-shock response element and linked these synthetic sequences 5' to the TATA box of the HSV-tk structural gene. The expression of the HSV-tk gene was found to be heat-shock inducible in COS cells and Xenopus oocytes when the heat-shock response element was positioned 10 - 20 base pairs 5' to the TATA box. Since the expression of the HSV-tk gene is not normally induced by heat-shock, these results indicate that the synthetic sequences introduced 5' to the HSV-tk gene are sufficient for conferring heat-shock inducibility.

There are few examples of sequences that are known to regulate tissue-specific expression. Failure to identify such sequences may be because cis-acting sequences which are required for tissue-specific expression are not always closely associated with the transcription unit. Another reason may be that cloned DNA does not always take up the appropriate chromatin configuration when

introduced into cells. Tissue-specific expression appears to be under both negative and positive control (Killary and Faurnier, 1984 and references within). A further possible reason for failing to identify sequences that regulate tissue-specific expression may be the lack in cultured cells of the appropriate regulatory factors and/or the presence of factors that inhibit tissue-specific expression. This idea is supported by the following observations. (1) Cultured cells lose their ability to express some genes that are normally expressed in the fully differentiated state of the cells. (2) They sometimes express genes that are only or predominantly expressed during earlier developmental stages of the tissues they are derived from. Recent successes in obtaining tissue-specific gene expression in transgenic mice provide a way to circumvent the difficulties encountered when using cultured cells.

Despite the difficulties, a few regions that appear to be necessary for tissue-specific expression have been identified by studies in cultured cells. Walker et al (1983) identified 5' flanking sequences that are involved in the tissue-specific expression of the rat chymotrypsin B gene, the rat insulin II gene and the human insulin gene, in cell lines derived from pancreas. DNA sequences containing the 5' flanking regions of the insulin and chymotrypsin genes were linked to the coding sequence of the chloramphenicol acetylase (CAT) gene. The effects of deletions within the 5' promoter sequences were assayed during transient expression in pancreas endocrine (HIT) and exocrine (AR4-2J) cell lines. These cell lines were chosen because insulin and chymotrypsin are normally expressed in pancreas endocrine and exocrine cells respectively. It

was found that the CAT gene was expressed preferentially in HIT cells when linked to DNA sequences containing the 5' flanking region of the insulin genes and in AR4-2J cells when linked to DNA sequences containing 5' flanking regions of the chymotrypsin gene.

The deletion studies showed that sequences located somewhere between nucleotides -150 and -300 of the genes were required for efficient expression in the appropriate cell line. The nature of these sequences and their roles in determining tissue-specific expression are currently not known. These experiments of course do not exclude the possibility that sequences within or flanking the insulin and chymotrypsin genes are required for efficient tissue-specific expression in the intact pancreas.

Not all cis acting sequences involved in the regulation of gene expression during development and differentiation are necessarily found 5' to the coding sequences. For example a tissue-specific enhancer is located within the transcription unit of some immunoglobulin genes. Other evidence comes from studies on the expression of cloned globin genes in mouse erythroleukemia (MEL) cells (Spandidos and Paul, 1982). The expression of cloned human β-globin genes in MEL cells is induced when these cells are allowed to differentiate (Wright et al, 1983; Chao et al, 1983). Cloned human α1-globin genes however are expressed at ∿ equivalent levels in MEL cells before and after the cells are induced to differentiate. Charnay et al, (1984) studied the expression of various human α1-globin/human β-globin hybrid genes in MEL cells and

found that sequences 3' to the cap site, are involved in the induced expression of the human β-globin gene. Sequences which allow the α-globin gene to be expressed equivalently before and after MEL cell differentiation are also located 3' to the cap site.

Wright et al (1984) also studied human β-globin gene expression in MEL cells and obtained results which complement those of Charnay et al (1984). They constructed fusion genes containing the coding region of the human β-globin gene linked to 5' promoter sequences of the mouse immunoglobulin H-2k bml gene or of the human γ-globin gene. Expression of these hybrid genes was found to be induced on MEL cell differentiation. Since expression of the H-2k bml gene and the γ-globin gene are not normally inducible in MEL cells, the results indicate that sequences 3' to the translational start point regulate the expression of the human β-globin gene. By following the transcription rate of the hybrid genes in isolated nuclei, Wright et al (1984) showed that the induced expression is at least in part due to an increase in the rate of transcription, and not simply the result of globin mRNA stability.

In addition, this group of researchers investigated whether sequences 5' to the translation start point were involved in the regulated expression of the human β-globin gene in MEL cells. A 1.8 kbp fragment which is present immediately 5' to the translation initiation site of the human β-globin gene was linked to the coding sequences of the H-2k bml gene or the coding sequence of the γ-globin gene. Expression of these hybrid genes was also found to be induced in differentiated MEL cells, indicating that sequences which

regulate the expression of the β-globin gene are present 5' to the translation start point.

To summarize, sequences involved in the regulated expression of globin genes appear to be located both 5' and 3' to the cap site. Whether the regulatory sequences present 3' to the cap site are found within the transcription unit, or within the immediate globin gene 3' flanking sequences, has not yet been determined.

The structure of active genes

The work described in this thesis is largely concerned with the structure of cloned eukaryotic genes. However, for the sake of completeness, it is worth mentioning, briefly, two types of change of state that are associated with active genes in vivo. The first involves changes in chromatin structure. Active chromatin has been found to develop an increased sensitivity to DNaseI and micrococcal nuclease (Weintraub and Groudine, 1976; Wu et al, 1979a; 1979b; 1980). Moderately DNaseI sensitive sites may be found over a large region containing more than one gene. In the ovalbumin gene family the sensitive domain extends for ∿100kbp and thereby encompasses all three members of the gene family (Lawson et al, 1982). Digestion of DNA with very low concentrations of DNaseI causes cleavage to occur at specific sites. These sites may be mapped by purifying the digested DNA, cleaving it with a restriction enzyme and analysing the digestion products by the Southern blot method. Wu et al (1980) demonstrated that a number of sites within

the transcription unit and 5' flanking regions of heat shock genes developed increased sensitivity to DNaseI after induction by heat-shock. Since then the presence of DNaseI hypersensitive sites within and flanking the transcription unit has been described for many expressing genes.

Studies on the chicken globin genes in tsAEV-transformed red blood cells (where globin synthesis is induced upon a temperature shift) have shown that the appearance of DNaseI hypersensitive sites preceeds transcription (Weintraub et al, 1982). DNaseI hypersensitive sites have also been found to be maintained after transcription has ceased. This phenomenon has been demonstrated in the major chicken vitellogenin gene (Burch and Weintraub, 1983) and the chicken globin genes (Groudine and Weintraub, 1982). It therefore appears that while DNaseI hypersensitive sites are associated with expression, they are not a consequence of transcription. Chicken oviduct chromatin contains a tissue specific DNaseI hypersensitive site within the vitellogenin transcription unit. Vitellgenin is synthesized in the liver but not in the oviducts of estrogen induced chickens. The oviduct-specific DNaseI hypersensitive site may therefore relate to the inactivity of the vitellogenin gene in this estrogen regulated tissue.

DNaseI hypersensitive sites are usually found within the 5' promoter region of actively transcribing genes (McGhee et al, 1981; Shermoen and Beckendorf, 1982; Sweet et al, 1982). These hypersensitive sites are considered to represent receptor or enzyme binding sites, since many map to regions which are thought to

contain sequences important for transcription initiation and regulation. For example, a DNaseI hypersensitive site and many restriction endonuclease hypersensitive sites have been found within the promoter region of the HSV-tk gene (Sweet et al, 1982). In the rat insulin II gene, a hypersensitive site within the 5' region has been found to coincide in position with the region required for the tissue specific expression of this gene (Walker et al, 1983). Also, Zaret et al (1984) have described a dexamethasone inducible hypersensitive site within the MMTV LTR which maps to an in vivo glucocorticoid receptor binding site.

More direct evidence comes from studies on the 5' region of the Drosophila heat-shock genes Hsp70 and Hsp83. Wu (1984) found that within the 5' DNaseI hypersensitive regions of these genes, two short stretches of sequence were protected from endonuclease III digestion. One of these sites, which was protected in both heat-shocked and non-heat-shocked embryos, was found to contain the TATA box. The other site, which was protected only in heat-shocked embryos, was found to contain the region that is required for heat-shock induction.

Many actively transcribing genes and their flanking sequences have been found to be relatively undermethylated. Examples are the chicken globin genes (Weintraub et al, 1981; Haigh et al, 1982), the human globin genes (van der Ploeg and Flavell, 1980), the ovalbumin gene (Mandel and Chambon, 1979) and MMTV (Gunzburg and Groner, 1984). The extent of methylation is usually assayed by digesting the genomic DNA with restriction enzymes sensitive to CpG

methylation, and comparing the digestion patterns of DNA derived from expressing and non-expressing cells.

More direct evidence for the association of under-methylation with gene expression has come from studies on cloned genes introduced into cultured cells. Compere and Palmiter (1981) found that a non-inducible cell line for metallothionein I gene expression was made inducible after treatment with 5-azacytidine, a cytidine analogue which cannot be methylated and which appears to block the methylation of cytidine residues. In vitro methylation of cloned genes prior to their introduction into Xenopus oocytes (Vardimin et al, 1982; Waechter and Baserga, 1982; Fradin et al, 1982) or mouse L cells (Stein et al, 1982) has been found to inhibit expression of these genes. Moreover, Busslinger et al (1983) have demonstrated that methylation of the 5' region of the human $\gamma$ globin gene (nucleotides +1ØØ to -76Ø) prevents transcription, while methylation of the coding sequences has no effect. This implies that specific CpG sites in the 5' region of the gene may be involved in regulation. Undermethylation however is not always associated with gene expression. An example of a gene which is not undermethylated when expressed is the Xenopus vitellogenin gene (Gerber-Huber et al, 1983).

Studies on the role of methylation on gene expression have been limited by the fact that the resriction enzymes used only detect a small fraction of the potentially methylated CpG sequences. Genomic sequencing using the Church and Gilbert method (1984) should help to establish whether specific sites are demethylated in actively

transcribing genes. However, this method does not appear to be suitable for studying the genomic variation in methylation between closely related members of a gene family due to cross hybridization of the sequencing probes to more than one gene.

## Gene Families

Many eukaryotic genes are members of gene families. Gene families constitute evolutionarily related genes that share common, although not necessarily identical, functions. Different gene families may vary considerably in size. They may be composed of two or three genes or several hundred genes. The size of a gene family may also vary considerably between different species. This is illustrated quite dramatically by the histone genes of higher eukaryotes; here the number of genes constituting the gene family in some cases differs by two orders of magnitude.

Methods for the preparation of random and representative "gene libraries" have made it possible to clone genes that are present as a few copies or as single copies in the haploid genome. This has led to the identification of new gene families and of unsuspected members of gene families. Complex genomic blot patterns, from DNA digests cleaved with low frequency cutting restriction enzymes, have revealed that genes that were once thought to be single copy genes are in fact members of gene families, and that gene families which were once thought to be smaller, are in fact much larger. Such discoveries have commonly occurred where the products of a gene

family are identical or very similar, or where the unsuspected genes are silent. Genes have also been discovered in cases in which the unsuspected genes are regulated differently from the genes whose products were traditionally studied.

There seem to be several different reasons for the evolution of gene families. Some RNAs and proteins that are required in great quantities during certain developmental stages are found to be encoded by several identical genes. It has therefore been suggested that the synthesis of large amounts of gene product may be facilitated by gene amplification. Experimental support for this suggestion comes from the amplification of appropriate drug resistant genes when cultured cells are maintained in medium containing high concentrations of specific drugs. For example, cells develop resistance to methotrexate by amplifying the dihydrofolate reductase (DHFR) genes and overproducing DHFR (Alt et al, 1978). However, high intracellular RNA concentrations are attained in some cases by the transcription of a single gene copy (see Hentschel and Birnstiel, 1981). For this reason, increased gene dosage due to a requirement for large amounts of product, may not always be a correct explanation.

The members of a gene family often encode different products that fulfil related physiological roles. The different products may be required at different developmental stages. For example, co-ordinate expression of the $\alpha$ and $\beta$ globin gene families gives rise to embryonic, foetal and adult haemoglobins, each with a different polypeptide composition. Different products of a gene family may

also be expressed exclusively or preferentially in certain tissues. Among the better characterized gene families that show such tissue specific expression are those coding for the interferons, tubulins, α-amylases and histones.

Some members of some gene families are pseudogenes. These non-functional genes accumulate base substitutions, insertions and deletions within the transcription unit and flanking regulatory regions. Examples of sequenced pseudogenes harbouring such mutations are the rabbit pseudo-β 2 globin pseudogene and the human pseudo-α 1 pseudogene (Lacy and Maniatis,1980; Proudfoot and Maniatis, 1980). Several pseudogenes have been isolated that lack the introns that are present in other members of the same gene family. Some of these pseudogenes have been found to contain structures which are present only in processed mRNA (Lee et al, 1983; Karin and Richards, 1982; Dudov and Perry, 1984). It has been suggested that such "processed pseudogenes" arise from retroviral reverse-transcription of mRNA into cDNA in the germ line. It is proposed that this cDNA either takes part in gene correction events with homologous genes or itself becomes integrated into the genome (Nishioka et al, 1980; Vanin et al, 1980). Another suggestion has been that gene correction events may take place directly between mRNA and homologous genes in the germ line (Leder et al, 1980). Finally, Lee et al (1983) have suggested that processed pseudogenes may arise most frequently in multigene families that are expressed in the germ line.

Generally speaking, different members of gene families are

structurally homologous in both coding and non-coding sequences. Homologous exons often show extensive sequence similarity in their sequences and are usually of similar size. Introns are less highly conserved and may vary considerably in size. However their positions within the transcription unit are relatively well conserved. For example, silk moth chorion genes are interrupted at a homologous position by a single intron (Jones anf Kafatos, 1980; Iatron and Tsitilou, 1983); functional vertebrate globin genes are interrupted at homologous positions by two introns (see Jeffreys, 1982); the ovalbumin, X and Y genes are interrupted at homologous positions by seven introns (Royal et al, 1979; Heilig et al, 1980) and the vitellogenin A1 and A2 genes are interrupted at homologous positions by ∿33 introns (Wahli et al, 1980). These examples also serve to show that conservation in structure may be maintained regardless of the number of introns that interrupt the coding sequence.

Nevertheless, differences in the structure of the transcription unit within members of gene families have been found. For example, the presence of different poly(A) addition sites in the ovalbumin gene family leads to variation in the size of the last exon. In the mouse major urinary protein (MUP) gene family, variation in both the size of the 3' exon and the number of exons results from the presence of different poly(A) addition sites and different splicing sites (Clark et al, 1984a). In the collagen gene family, differences within the 5' ends of the transcription unit are thought to be due to the presence of different splicing sites and different promoter regions (Chu et al, 1984). Variation in the number of introns can also result from the loss or gain of these sequences during evolution.

Examples of such phenomena are provided by the rat preproinsulin gene family (Lomedico et al, 1979), the chicken histone gene family (Engel et al, 1982; Harvey et al, 1983), the sea urchin actin gene family (Breathnach and Chambon, 1981) and the tick globin gene family (Antoine and Niessing, 1984).

Members of a gene family may be linked, or may be dispersed on different chromosomes. Some gene families consist of both dispersed and linked members. For example, the members of the human α-globin genes are linked and reside within an approximately 50 kb cluster on chromosome 11, while the human β-globin genes are present within an approximately 28 kb cluster on chromosome 16. Genes within a cluster may be linked in a head to tail fashion, or in a head to head fashion. Examples of the former are the vertebrate globin genes (see Jeffreys, 1982), the ovalbumin, X and Y genes (Royal et al, 1979) and the sea urchin histone genes (Hentschel and Birnstiel, 1981). Examples of the latter are the heat-shock genes at 87A and 87C (Leigh Brown and Ish-Horowitz, 1981), some Xenopus and chicken histone genes, the mouse class II MHC α and β genes (Steinmeitz and Hood, 1983) the silk-moth chorion genes and two Drosophila yolk protein genes.

In the silk-moth, divergently orientated gene pairs, formed from different chorion gene subfamilies, are found to be co-ordinately expressed (Jones and Kafatos, 1980a, 1980b). This has led to the suggestion that the shared 5' sequences of these genes may determine their common developmental regulation (Iatrou and Tsitilou, 1983). Co-ordinately expressed chorion gene pairs are also found to be

closely linked. This has led Eickbush and Kafatos (1982) in turn to speculate that co-ordinately regulated chorion genes may constitute separate transcriptionally active domains.

Evidence for *cis* acting regulatory sequences that are shared by linked members of a gene family has come from the *Drosophila* yolk protein genes (Garabedian *et al*, 1985). The 5' ends of the transcription units of the divergently orientated *Drosophila* yolk protiens, *yp1* and *yp2*, are separated by 1225 bp. These closely linked genes were found to share two physically separatable elements that are required for their correct tissue specific expression. One of these elements is necessary for expression of the genes in the ovaries, while the other is necessary for expression of the genes in fat bodies. It will be interesting to identify the sequences that are necessary for expression of the *yp3* yolk protein gene in ovaries and fat bodies: the *yp3* gene lies approximatly 1000 kbp away from *yp1* and *yp2* but is co-ordinately expressed with these genes in the ovaries and fat bodies (Barnett *et al*, 1980).

The organization of the globin gene families of some vertebrates reflects the developmental regulation and functional relatedness of the genes. Human globin genes in both the $\alpha$ and $\beta$ clusters are arranged in an order which coincides with the temporal order in which the genes are expressed during development. Other mammalian globin genes and the chicken $\alpha$-globin genes are also arranged in the temporal order in which they are expressed (Jeffreys, 1982; Dodgson *et al*,1981). However, adult chicken $\beta$-globin and adult *Xenopus laevis* globin genes are flanked by embryonic and larval

globin genes respectively (Dolan et al, 1981; Hosbach et al, 1983). Furthermore, the globin clusters in Xenopus constitute linked α and β genes (Jeffreys et al,1981). It is therefore clear that the organization of genes within this gene family does not have any obvious functional basis. Whether it ever does and whether serial ordering of genes within a cluster is in some way advantageous is not known.

Members within gene families often show homology in their flanking sequences. Barring deletions and disruptions caused by insertions, the extent of these homologies depends on the extent of the duplication events. Linked products from one or more duplications may form a "duplication unit" that in turn may be amplified several times. Examples of such paired-gene and multigene duplication units are found in the silkmoth chorion gene family (Jones and Kafatos, 1980; Iatrou and Tsitilou, 1983) and in eukaryotic histone gene families (Hentschel and Birnstiel, 1981) respectively. Unequal crossing-over in the germ line is believed to be a mechanism by which genes are amplified (Smith et al,1976; Zimmer et al,1980). Evidence for unequal crossing-over as a mechanism by which gene families evolve comes from variation in the length of the rDNA repeats of Drosophila and X.laevis (Wellauer et al, 1976; Wellauer and Dawid, 1977) and from rare alterations in the number and structure of globin genes within the human globin clusters (See Lewin, 1983).

Within a species, the degree of homology between members of a gene family can be remarkably high. The high sequence homology may be

shared by all members or it may be confined to a specific region and shared by only two or three genes. In contrast, gene families often show substantial variation between species. For these reasons it is thought that mechanisms must act that lead to the homogenization of sequences within gene families. One of these mechanisms is unequal crossing-over (Smith, 1976). Independent amplification events of a gene family in different species, can result in the replacement of an ancestral array by different sets of genes in each species. Unequal crossing-over can also serve to exchange sequences between non-allelic genes, if the cross-over points take place within the genes. Examples of such chimeric genes are found in some β-thalassemias (see Lewin, 1983).

Another mechanism which leads to gene homogenization is gene conversion. Gene conversion is the non-reciprocal exchange between two homologous sequences. This phenomenon occurs when DNA strands from two allelic, or non-allelic but homologous genes, form a heteroduplex, and correction of mismatched bases takes place. Such events within the germ-line lead to heritable change. Gene conversion has been shown to take place in yeast and other fungi (Fink and Styles, 1974; Radding, 1978; Klein and Petes, 1981). In these organisms the molecular mechanism of gene conversion has also been extensively studied (Jackson and Fink, 1981; Klar and Strathern, 1984; Klein, 1984). In higher eukaryotes it has been suggested that gene conversion occurs in the globin genes (Slighton et al, 1980), the mouse major histocompatibility genes (Weiss et al, 1983; McIntyre and Seidman, 1984), and the immunoglobulin genes (Bentley and Rabbitts, 1983; Ollo and Rougeon, 1983; Baltimore,

1981).

Limited sequence data often does not allow one to distinguish between the products of gene conversion and unequal crossing-over. This is made especially difficult by the fact that reciprocal exchange due to double cross-overs may take place between non-allelic genes, albeit at lower frequencies than reciprocal exchange due to single cross-overs. For these reasons, the term "gene conversion" has often been loosely used to describe genetic exchange between non-allelic genes, where change in sequence length is not observed and where all the products of a single recombination event cannot be recovered.

## MUP Genes

Studies on recently diverged gene families are fruitful for two reasons. The first is that subtle differences which lead to important changes in regulation are not obscured and may easily be identified. The second is that structural differences between members can contribute to our understanding of the mechanisms involved in eukaryotic genome evolution. This thesis describes the characterization of genes from a recently diverged gene family: the major urinary protein (MUP) gene family of the mouse.

The major urinary proteins of the mouse are a family of closely related, small, acidic proteins (Mr $\sim$19,000) that are synthesized in large quantities in the liver. The MUP genes are also expressed in

the mammary, lachrymal, submaxillary, parotid and sublingual glands, but at much lower levels than in the liver (Shaw et al, 1983). In vitro translation of liver mRNA shows that there are at least twelve different MUP species expressed in this tissue (Clissold and Bishop, 1982; Shaw et al, 1983; Shahan and Derman, 1984). The in vitro translation products of the submaxillary, parotid, sublingual and mammary glands largely comprise different subsets of the liver products, while the lachrymal gland mRNA gives rise to a different set of MUP proteins (Shaw et al, 1983). The MUPs are under multihormonal control and variation in hormonal responsiveness is detected between MUPs that are expressed in the same tissue as well as between MUPs that are expressed in different tissues.

Liver MUPs are secreted into the plasma and excreted into the urine, where MUP excretion may be as much as 20 mg per mouse per day. Protein in the urine of mice was first described by Parfentjev in 1932. Subsequent work demonstrated that the urinary protein was of hepatic origin and that both sex differences and strain differences were found in the amounts of urinary protein excreted (Finlayson and Baumann, 1958; Finlayson et al, 1963; Ruemke and Thung, 1964; Finlayson et al, 1965).

Agarose and acrylamide gel electrophoresis resolved the urinary MUPs into three components: MUP1, MUP2 and MUP3 (Finlayson and Baumann, 1958; Finlayson et al, 1963; Finlayson et al, 1974). Inbred strains were found to fall into two classes: those excreting MUP1 and MUP3, and those excreting MUP2 and MUP3. Hudson et al (1967) identified a genetic locus responsible for the observed strain

variability which was linked to the brown locus on chromosome 4. Finlayson *et al* (1974) sequenced the N-terminal regions of the MUP components, MUP1, MUP2 and MUP3, and found that MUP1 and MUP2 were nearly identical. These results led to the suggestion that MUP1 and MUP2 represented the products of two allelic genes *Mup-1*$^{c}$ and *Mup-1*$^{b}$ located on chromosome 4.

Later it was shown that the pattern of urinary MUP excretion was more complex. First of all it was found that all strains investigated excrete the three MUPs (MUP1, MUP2 and MUP3) when induced by testosterone (Szoka and Paigen, 1978). This observation led Szoka and Paigen (1978) to propose that *Mup-1* is a regulatory locus. Secondly, it was found that a much larger set of distinguishable MUPs is present in the urine of inbred mice (Hoffman, 1982, Hainey and Bishop,1982). Thirdly, and more recently, IEF resolution of urinary MUPs and MUP mRNA translation products demonstrated that although variation between strains is predominantly quantitative, qualitative differences are also observed (Clissold and Bishop, 1982).

In BALB/c adult male mice, MUP mRNA makes up ∿8% of total liver poly(A) mRNA, this level being five times higher than that of adult female mice (Hastie and Held, 1978; Hastie *et al*, 1979). In BALB/c male mice, MUP mRNA is the most abundant class of liver mRNA, while in female BALB/c mice, serum albumin is the most abundant class of mRNA in the liver (Clissold and Bishop, 1981). Female mice show a simpler liver MUP pattern than male mice, although treatment with testosterone induces a male-like pattern and results in the

TABLE I.1. Relative level of MUP mRNA in expressing tissues

| Tissue | Max Level[a] of MUP mRNA (copies per cell) | Hormonal Regulation | Influencing Hormones | First Detectable Time of Expression |
|---|---|---|---|---|
| Liver (male) | 30,000 | + | T, $T_4$, GH | 3 weeks |
| Lachrymal (male) | 6,000 | + | T | 2 weeks |
| Submaxillary gland | 1,250 | - | none | 1 week |
| Mammary gland | 1,000 | ? | unknown | 1st pregnancy |

(From Shaw et al., 1983).

synthesis of male-like levels of MUP mRNA (Finlayson et al, 1963; Szoka and Paigen, 1978; Clissold et al, 1984).

The expression of MUP in different tissues is under different developmental control. In the liver, MUP mRNA is first detected in 3 week old male mice, full expression being reached only 6 - 7 weeks after birth. Derman, (1982) showed that the different levels of MUP mRNA in the livers of mice of different ages and sex, are reflected in differences in the rate of transcription. The lachrymal glands, like the liver, show sexual dimorphism, with male lachrymal glands having approximately five times as much MUP mRNA per cell as female lachrymal galnds. However, unlike the liver, adult MUP mRNA levels in this tissue are already established at two weeks of age. This corresponds to the earliest stage at which lachrymal glands can be identified for dissection. The submaxillary gland does not show sexual dimorphism with respect to MUP expression. MUP mRNA in this tissue is detectable in one week old mice, maximal levels being achieved between 4-7 weeks of age. MUP expression in the mammary glands is detected at the first pregnancy (Shaw et al, 1983; see Fig.I.1).

The hormonal regulation of MUP is different in the different tissues. Liver MUP mRNA is regulated by testosterone, thyroxine, growth hormone and, in some strains, glucocorticoid (Knopf et al, 1983; Norstedt and Palmiter, 1984). Using thyroidectomized and hypophysectomized female mice and mutant mice (*little* male mice and *tfm*/Y mice), Knopf et al (1983) have found that testosterone, growth hormone and thyroxine modulate MUP synthesis. Shaw et al

(1983) have also found that different liver MUP components are regulated differently by testosterone, thyroxine and growth hormone. In the lachrymal glands, testosterone induction of MUP appears to be independent of growth hormone and thyroxine, while in the submaxillary gland, MUP expression does not appear to be under hormonal regulation. The hormonal regulation of the mammary gland MUP(s) has not yet been studied.

The MUPs are the products of a large gene family consisting of ∿35 closely related genes (Bishop et al, 1982) clustered on chromosome 4 (Bennett et al, 1982; Bishop et al, 1982; Krauter et al, 1982). The predominant duplication unit of the MUP genes appears to consist of two genes, linked in a head to head fashion and separated by 15 kbp of DNA (Clark et al, 1984b). The divergently orientated paired genes with their 3' flanking sequences, encompass 45 kbp of DNA and constitute a huge imperfect palindrome. The paired genes within the duplication unit are members of two different groups, as originally defined by hybridization criteria (Bishop et al, 1982). MUP expression appears to be largely the result of transcription from group 1 genes, and sequencing data has suggested that most, if not all, group 2 genes are pseudogenes, sharing a common mutation (a stop codon in the seventh amino acid of the mature protein).

From genomic blot analyses, it has been estimated that group 1 and group 2 consist of ∿15 genes each and that there is a similar number of 45 kbp duplication units within the genome of the BALB/c mouse (Bishop et al, 1982; Clark et al, 1984b). Several group 1 and group 2 genes have been isolated (Clark et al, 1982; Clark et

al, 1984b). A few genes that do not fall into either group 1 or group 2 have also been isolated. Some of these are pseudogenes that have resulted from a number of different rearrangements.

The transcription unit of MUP genes is 3.9 kbp long and contains seven exons (Clark et al, 1984a). The first six exons contain the coding region sequences, while the last exon consists entirely of non-coding sequences. Three different splicing configurations have been found, which result from the presence of alternative splice sites within the untranslated region of exon 6 (Clark et al, 1984a; A.Chave-Cox, unpublished results). The most abundant liver transcripts contain part of exon 6 and all of exon 7. The less abundant and smaller liver transcripts entirely lack exon 7.

Several partially cloned group 1 cDNA clones have been isolated from cDNA libraries prepared from the livers of different mouse strains (Hastie et al, 1979; Clissold and Bishop, 1981; Derman, 1982; Kuhn et al, 1984). The sequences of four nearly full-length cDNA clones are known (Kuhn et al, 1984; A.Chave-Cox, unpublished results). The sequences of the mRNA-specifying regions of four different BALB/c group 1 genes are also known (Clark et al, unpublished results). These appear to code for a signal peptide, 18 amino acids long, and a mature protein, 162 amino acids long. Although nucleotide homology between the different group 1 sequences is on average 99.6%, they nevertheless specify different proteins. At present it is not known which genes code for which proteins.

Two non-group 1 cDNA clones, p199 and MUP 15, have been isolated

(Kuhn et al, 1984; A.Chave-Cox, unpublished). These are identical in their overlapping cloned sequences (707 bp), and are thought to correspond to transcripts of a low copy number gene (possibly single copy) within the mouse genome. pl99, isolated from a C57BL/6J male liver cDNA library (Kuhn et al, 1984), and MUP 15, isolated from a BALB/c female liver cDNA library (A.Chave-Cox, unpublished), specify a protein which is considerably different in amino acid composition from the proteins specified by the group 1 genes. The pl99/MUP15 protein sequence contains a potential glycosylation site. MUPs, generally, have not been found to be glycosylated (Szoka and Paigen, 1978), although a protein component whose mRNA is preferentially selected by a pl99 5' subclone, may be glycosylated. This component is synthesized in the livers of both male and female mice.

The transcription units of a group 1 gene (BS-6) and a group 2 gene (BS-2) have been sequenced, and the homology between the two genes has been found to be 90% at the nucleotide level. Although there are several insertions and deletions within the introns, the overall structures of the two transcription units are very similar. Sequence data from the 5' flanking regions of BS-6 and BS-2 has also shown that these regions are generally quite similar. Both genes carry a TATA box and an A-rich region at ∿-80. The length and sequence composition of the A-rich region, however, differs between the two genes. Another significant difference is the absence in BS-2, of one of the glucocorticoid consensus sequences found in the 5' flanking region of BS-6 (J.Clark, unpublished results).

The locations of DNaseI hypersensitive sites within the 45 kbp duplication unit have been mapped in the livers of male and female mice at various stages of development (J.Clark, unpublished results). Eight hypersensitive sites are arranged in similar positions around the group 1 and group 2 genes. These are present Ø.5 kbp 3' to the poly(A) addition sites and Ø.75, 2.25 and 7 kbp 5' to the cap sites of the group 1 and group 2 genes. Two non-symmetrical hypersensitive sites have also been found. These are present Ø.5 kbp and 5.5 kbp 5' to the cap sites of the group 2 genes. The hypersensitive sites are only fully established in the livers of 3-week-old mice and are not present in the kidney, a tissue which does not express MUP. Because DNaseI hypersensitive sites are associated with gene expression, it is thought that at least some of these sites are involved in the tissue-specific expression of the MUP genes.

Differences in the ratio of group 1 like sequences and p199/MUP15 like sequences are found within the genomes of the wild mouse strains M.musculus, M.castaneus, M.hortulanus, M.caroli and M.cervicolor (Sampsell and Held, 1984). The differences may reflect variation in copy number of group 1 and p199/MUP15-like sequences, and/or variation in homology to the group 1 and p199/MUP15 probes. Differences in the hepatic mRNA ratio of group 1-like sequences and p199/MUP15-like sequences have also been found among the wild mouse strains. These may be due to a combination of differences in regulation, sequence homology and gene dosage. More detailed studies on the MUPs of wild mice may give us some interesting insight into the evolution of the MUP gene family in the

mouse.

In the rat, a homologous gene family codes for the $\alpha_{2u}$globulins (Kurtz, 1981; Dolan et al, 1982). $\alpha_{2u}$globulins are synthesized in the liver in male rats (Laperche et al, 1983). Hepatic $\alpha_{2u}$globulin is regulated by testosterone, glucocorticoid, thyroxine, growth hormone, insulin and estrogen, unlike submaxillary $\alpha_{2u}$globulin which does not appear to be under hormonal regulation (Motwani et al, 1980; Lynch et al, 1982; Ray et al, 1983; Laperche et al, 1983). Dexamethasone-induced expression of $\alpha_{2u}$globulin genes introduced into $Ltk^-$ cells has been reported (Kurtz, 1981).

The $\alpha_{2u}$globulins are thought to be encoded by $\sim$20 genes (Kurtz et al, 1981). Comparison of a rat $\alpha_{2u}$globulin gene (207) and a mouse group 1 gene (BS-6) has revealed that their transcription units are similar in structure and that their exonic sequences are 81% homologous.

Assuming that the mouse-rat divergence took place 30 million years ago, the silent site and replacement site divergence rates between the $\alpha_{2u}$globulin gene and the group 1 MUP gene are estimated to be 0.33 M years and 1.3 M years respectively (Clark et al, 1984a). From these divergence rates, it is concluded that while the genes are undergoing rapid evolution, the protein sequences are being conserved. The functions of the MUPs and the $\alpha_{2u}$globulins are not known. Based on their sites of expression, Shaw et al (1983) suggested that the MUPs may be involved in behavioural communication. Recently it has been found that MUP genes share

significant homology to $\beta$-lactoglobulin, a secretory protein found in the milk of ruminants, and $\alpha_1$ microglobulin, a low molecular weight human plasma protein of unknown function.

Aims of project

In the above section I have described our current knowledge of the MUP gene family and the proteins it encodes. The project described in this thesis is largely concerned with the characterization of cloned MUP genes by restriction enzyme mapping. In the following section I will first briefly describe our knowledge of the MUP gene family at the time the project was initiated. I will then outline the main aims the project wished to achieve.

In 1981, BALB/c MUP genes were isolated from a liver genomic library and a sperm genomic library (see Clark et al, 1982). These were found to fall into two main groups, based on their hybridization reactions to the sub-clones, BS-6-5-5 and BS-2-2-2, which contain homologous fragments of the MUP genomic clones BS-6 and BS-2 respectively. Isolated MUP genes which formed stable hybrids with BS-6-5-5 at Ø.2 x SET, 68°C, were classified as group 1 genes while those which formed stable hybrids with BS-2-2-2 were classified as group 2 genes. The hybridization studies coupled with restriction enzyme mapping showed that four different group 1 genes (BS-6/BL-14/BL-7, BS-5, BL-1, BS-1), three different group 2 genes (BS-2/BS-3/BS-4, BL-25, BL-15), and three genes belonging to neither group (BL-2, BL-8, BL-6) had been isolated. BL-6, one of the

clones which did not form a stable hybrid with either the group 1 probe or the group 2 probe, was classified as a pseudogene. This was based on its interrupted pattern of hybridization to two overlapping cDNA clones.

A high degree of restriction enzyme homology in the coding regions and 5' and 3' flanking regions was found between the isolated MUP clones. Restriction site homologies were found to extend into the flanking region for at least 7 kbp in both directions (Clark et al, 1982). Hybridization of genomic blots with the group 1 and group 2 probes showed that there were approximately 35 MUP genes in the haploid BALB/c genome consisting of ∿15 group 1 genes, ∿15 group 2 genes and ∿5 genes belonging to neither group (Bishop et al, 1982).

In 1981, Bennett et al found that variation in the urinary MUP phenotype between inbred mouse strains was paralleled by variation in the genomic restriction patterns of the MUP genes. Strains with the same urinary MUP phenotype also had the same MUP restriction enzyme pattern. By comparing the urinary MUP phenotype and MUP genomic restriction fragment patterns of two different sets of recombinant inbred strains (CXB and AKXL), Bennett et al also showed that the MUP structural locus was linked to the proposed regulatory locus Mup-1 (Szoka and Paigen, 1980). These results suggested that at least some of the variation in MUP phenotype between inbred strains may be due to differences in the structural genes.

The inbred mouse strains BALB/c and C57BL/Fa have Mup-$1^c$ and Mup-$1^b$ genotypes respectively. Using solution hybridization, Clissold and Bishop (1982) estimated that BALB/c and C57BL/Fa mouse genomes carried an equal number of MUP genes. IEF resolution of the unprocessed *in vitro* translation products of hybrid selected MUP mRNA revealed that an equal number of distinguishable components (∿2Ø) were synthesized in the livers of BALB/c and C57BL/Fa male mice. The number of MUP components synthesized in the livers of BALB/c and C57BL/Fa mice was found to be similar to the estimated number of group 1 genes present in the haploid genome of BALB/c mice. Because liver mRNA formed stable hybrids with a group 1 cDNA clone at high stringency, it was concluded that most of the liver MUP components were probably the products of group 1 genes (see Clark *et al*, 1984). The *in vitro* translation studies also revealed that although the differences between the inbred strains in the liver MUP components were mainly quantitative, some qualitative differences were found. Taken together the results of Bennett *et al* (1982) and Clissold and Bishop (1982) suggested that some variant genes were expressed in inbred mouse strains with different Mup-1 genotypes.

The MUP clones isolated from the BALB/c genomic libraries represented ∿1/4 of the estimated number of MUP genes in the haploid genomes of inbred mouse strains. The first aim of the project was to increase the pool of isolated and characterized MUP genes. This was done for two reasons. (a) It would allow the identification of subtle variations between different MUP genes, which potentially lead to differences in their hormonal and tissue

specific regulation. (b) It would give a more comprehensive understanding of the evolution of the MUP gene family.

The second aim of the project was to isolate a functional variant MUP gene that was present in one strain but not the other. Such a gene would be used to study the regulation of MUP genes in the mouse. A variant gene from one strain would be micro-injected into the pronuclei of fertilized eggs of a mouse strain lacking the variant gene. The product of the variant gene would be distinguishable from the products of other MUP genes in the transgenic mice and studies on the expression of the variant gene would be facilitated by developing a specific probe to its mRNA. The introduction of deletions or point mutations within specific regions of its coding and flanking sequences, and fusion of its coding sequences to different promoters from other MUP genes, would allow the definition of MUP regulatory elements. Because controlling elements are sometimes found within the transcription units of genes, the use of a natural variant would be superior to the use of a synthetic fusion gene. Also, due to potential species differences, the use of a natural variant would be superior to the use of a rat $\alpha_{2u}$globulin gene.

For the reasons outlined above, C57 MUP genomic clones were isolated and characterized.

## Methods

### Preparation of C57 unamplified library: Ligation and Packaging

EcoRI cut Charon 4A arms were first annealed in 66mM Tris pH 7.6, 1mM EDTA pH 7.Ø, 1ØmM $MgCl_2$ and 4Ø mM NaCl for three hours at 42°C. A total of Ø.125 units of BRL T4 Ligase was used to ligate 1.8µg of mouse EcoRI* fragments with 4µg of Charon 4A arms. Ligation was carried out at 1Ø°C, overnight, in a reaction buffer containing 5ØmM Tris pH 7.6, Ø.8mM EDTA pH 7.Ø, 8mM $MgCl_2$, 32mM NaCl, 1ØØµgml$^{-1}$ BSA, 1ØmM DTT and Ø.2mM ATP.

The amount of T4 ligase needed to give optimum ligation had been previously determined by a series of test ligations. Successful ligation was identified as described by Maniatis et al (1978).

Packaging was carried out as described by Grosveld et al (1981) using packaging extracts prepared by Melville Richardson in John Bishop's laboratory.

### Plating Charon 4A and its recombinant derivatives.

The host used was E. coli ED8654 [supE, supF, hsd $R^-M^+S^+$, met$^-$, trpR (Murray et al, 1977)]. A 1:5Ø dilution of stationary phase cells to LB was prepared and the cells allowed to grow with shaking to an $O.D_{540nm}$ of Ø.5. Cells were pelleted by centrifugation and resuspended in an equal volume of cold 1ØmM $MgSO_4$.

Ø.5ml of bacteriophage suspended in phage buffer was incubated with Ø.5ml of the prepared host cells for 2Ø minutes at 37°C. 3ml of LB top at 42°C made 1ØmM in $MgSO_4$ was added after the incubation period and the mixture poured onto 9cm diameter, Ø.5cm deep, LB bottom plates. After the top agar had set, the plates were inverted and incubated at 37°C overnight.

All volumes were scaled up when using larger plates.

X-gal plates

To estimate the number of non-recombinant bacteriophage in Charon 4A genomic libraries, the bacteriophage were plated on a lawn of *E. coli lac z* (C344) grown on an X-gal indicator plate (Blattner *et al*, 1977).

X-gal: 5 chloro 4 bromo 3 inodolyl-β-galactoside

C3344: *thr*, *leu*, Bl, *supE*, *tonA*, *hsdR*$^-$ M$^-$, *lacZ*.

Storage of bacteriophage as plate lysates

Bacteriophage from a single plaque were plated at a density of $\sim 10^4$ pfu cm$^{-2}$ to give confluent lysis.

The top agar was scraped into 2 volumes of phage buffer, 1/50 volume of $CHCl_3$ added and the mixture vortexed, to ensure complete lysis. Cell debris was pelleted by centrifugation (10,000 rev. $min^{-1}$ 4°C) and the clear phage suspension decanted into a fresh container and stored at 4°C.

## Isolation and Purification of nucleic acids.

### Quick bacteriophage DNA preparations.

These were prepared as described by Cameron et al (1977).

Analysis by electrophoresing restriction digests of these DNA preparations allowed early identification of identical phages picked from amplified genomic libraries.

### Pure bacteriophage DNA preparations.

These were prepared as described by Clark et al (1982).

DNA from these preparations was used for restriction enzyme mapping procedures.

### Plasmid DNA preparations.

HB101 was transformed with recombinant derivatives of the plasmid pPH207 (Bishop and Davis, 1980) and grown with the appropriate

plasmid selecting antibiotic.

Transformation and isolation of plasmid DNA were carried out as described by Bishop (1977), except that the plasmid was put over a Sepharose 2B column as a further purification step.

HB101: F-, hsdS20($R^-M^-$), recA13, ara-14, proA2, lacY1, galK2, rpsL20($Sm^r$), xyl-5, mtl-1, supE44

Genomic DNA.

DNA from male mouse liver nuclei was prepared as described by Clissold and Bishop (1982).

Preparation of E.coli DNA.

The carrier DNA used in genomic library screening hybridizations was that of the E.coli host ED8654.

Stationary phase bacteria were subcultured into six, 1 litre flasks each containing 200ml LB made 1% in glucose and grown overnight, with shaking, at 37°C.

The cells were pelleted by centrifugation for 15 minutes at 10,000 rev. $min^{-1}$ and resuspended in 400ml of 10mM Tris, 1mM EDTA, pH 7.4.

The cells were pelleted again and resuspended and homogenized in

100ml of sucrose mix. Lysozyme was added to a final concentration of 2.3mgml$^{-1}$ and the solution incubated for 15 minutes on ice. 30ml of 0.5M EDTA pH 8.1 was added and incubation continued for 5 minutes. 270ml of triton mix was added, and after a further 10 minute incubation period on ice, the nucleoids were pelleted by centrifugation for 30 minutes at 25000 rev. min$^{-1}$.

The pellets were taken up in 50ml of 10mM Tris, 0.1M EDTA, pH 8.0 and incubated for 30' on ice with RNase at a final concentration of 0.2mgml$^{-1}$. 10ml of pronase was added and the solution incubated on ice for four hours. 1 volume of phenol was added and the phases gently shaken overnight at room temperature. 0.5 vol. of $CHCl_3$ was added, the phases separated and the aqueous phase reextracted once or twice with $CHCl_3$.

0.1 vol. 3M NaCl and 1 vol. cold ethanol were added and the DNA spooled out and dissolved in 50ml of 10mM Tris pH 8.0. The DNA solution was made 0.3M in NaCl, sonicated and ethanol precipitated.

Isolation of mRNA

7-10 week old BALB/c male mice were sacrificed for the preparation of submaxillary gland and lachrymal gland RNA. Eight week old 18 day pregnant female BALB/c mice were sacrificed for the preparation of mammary tissue RNA. RNA was extracted as described by Chirgwin et al (1975), including the CsCl ultracentrifugation step to separate DNA from RNA. Poly(A) mRNA was isolated as described by Aviv and Leder (1972) except that LiCl was substituted for NaCL.

Agarose gel electrophoresis

1/ Mapping MUP recombinant bacteriophages

Two types of gel were prepared:

(A) Ø.4% agarose horizontal gels made 1 x in TA. DNA samples of Ø.25µg - Ø.5µg were applied to 5mm deep, 1mm thick, 2mm wide wells and electrophoresed at $3Vcm^{-1}$ for 16 hours.

(B) Ø.8% agarose vertical gels made 1 x in TB. DNA samples of Ø.5µg were applied to 8mm thick, 5mm deep, 5mm wide wells and electrophoresed at $2.2Vcm^{-1}$ for 16 hours.

2/ Electrophoresis of genomic DNA.

Ø.7% - 2.Ø% vertical agarose gels made 1 x in TB were used. Up to 2Øµg of DNA was applied to 8mm thick, 5mm deep, 1Ømm wide wells. Electrophoresis was carried out at $1.9Vcm^{-1}$ for 2Ø - 24 hours.

3/ Electroelution of DNA from agarose gels.

Ø.6% - 1% vertical agarose gels made 1 x in TB were used. Up to 1Øµg of DNA was applied to 8mm thick, 5mm deep, 2Ømm wide wells, and electrophoresed at $2.2Vcm^{-1}$ for 16 hours. After staining (see below) the gel was placed horizontally and a trough cut directly in front of the band to be eluted. The trough was filled with

electrophoresis buffer and a piece of sterile dialysis membrane was placed in the trough over and under the band. Electrophoresis of the band onto the dialysis membrane was carried out at 12 $Vcm^{-1}$ for 1 hour. With a voltage still being applied, the membrane was then removed into a small volume (2 - 3ml) of buffer and the DNA allowed to wash off the dialysis membrane. To purify the DNA from agarose derived enzyme inhibitors, the DNA was passed over a Schleicher and Schuell Elutip-d Column and recovered by ethanol precipitation.

Visualization of electrophoresed DNA

$1\mu gml^{-1}$ ethidium bromide was added to TA horizontal agarose gels before being cast. TB vertical gels were stained after electrophoresis in 1 x TB, containing $1\mu gml^{-1}$ ethidium bromide, for 30 - 60 minutes. The DNA bands were visualized under a short wavelength UV transilluminator.

RNA denaturing gels

Two types of vertical gel were used.

(A) 1.4% formaldehyde agarose gels made 1 x in MOPS. These were prepared as described by Clissold and Bishop (1982). Poly(A) RNA samples, denatured with formamide and formaldehyde as described by Rave et al (1979), were applied to 8mm thick, 5mm deep, 5mm wide wells and electrophoresed at $1.9Vcm^{-1}$ for 20 hours.

(B) 1.4% formaldehyde agarose gels made 1 x in PBS. The gels were

prepared and run exactly as in (A) except that PBS buffer was substituted for MOPS.

Screening genomic libraries

Plaque transfers were prepared by the method of Benton and Davis (1977) as modified by Maniatis et al (1978).

The prehybridization and hybridization steps were carried out as described by Maniatis et al (1978) except that 150$\mu g ml^{-1}$ of sonicated E.coli DNA (ED8654) was substituted for sonicated salmon sperm DNA in the prehybridization and hybridization steps and 10% Dextran Sulphate was added at the hybridization step (Wahl et al, 1979).

Preparation and hybridization of Southern transfers

The methods used for the transfer of DNA to nitrocellulose and its subsequent hybridization were essentially those of Wahl et al (1979) and Maniatis et al (1978), respectively. The modifications are described by Clissold and Bishop (1982).

Low stringency washes were in 1 x SET, 68°C. High stringency washes were in 0.2 x SET, 68°C. 1 x SET = 150 mM NaCl, 30mM Tris, 1mM EDTA pH 8.0.

Preparation of Southern transfers for rehybridization

It was often desirable to rehybridize Southern transfers with another probe. To prepare Southern transfers for rehybridization, the filters were wetted in 4 x SET for 2Ø minutes at room temperature, and the hybridized probe removed by dipping the filters in Ø.1N NaOH, 1.5M NaCl for 3Ø seconds. The filters were then neutralized by dipping in Ø.2M Tris, pH 7.5 and 2 x SSCP for 2Ø seconds, blotted, and dried by baking at 8Ø°C in a vacuum oven for 3Ø minutes.

Northern transfers

Northern transfers were prepared essentially as described by Thomas (198Ø).

Schleicher and Schüll nitrocellulose membrane filters (Ø.45µm pore size) were used to transfer RNA from MOPS/formaldehyde agarose gels. These were hybridized exactly as Southern transfers (see above). Pall Biodyne A nylon membrane filters were used to transfer RNA from PBS/Formaldehyde agarose gels. These were hybridized as outlined below. Filters were pre-treated for four hours at 42°C in 5Ø% formamide, 2.5 x Denhardts, 2.5 x SSPE, Ø.25% SDS and 25Øµgml$^{-1}$ of sonicated and denatured, salmon sperm DNA. They were then hybridized overnight at 42°C in a solution identical to that used in the prehybridization step except that 1Ø% Dextran Sulphate and the denatured probe were included. The prehybridization and hybridization steps were carried out in sealed plastic bags submerged in a shaking water bath. After hybridization, the filters were washed four times in 2 x SSC, Ø.1% SDS at room temperature for

15 minutes. Stringency washes were carried out by washing the filters in Ø.5 - Ø.2 x SET at room temperature for 15 minutes followed by washing in Ø.5 - Ø.2 x SET at 68°C for 3Ø minutes.

Restriction digests

Restriction of DNA with EcoRI, BamHI, HindIII, PvuII, PstI SalI and SstI were carried out in EcoRI buffer (1ØmM Tris pH 7.5, 1ØmM $MgCl_2$, 1ØØmM NaCl, 1ØmM β-mercaptoethanol) at 37°C. Restriction of DNA with other enzymes was carried out in the buffers and at the temperatures recommended by the manufacturers. In double digests with enzymes requiring different buffers, restriction was first carried out with the low salt requiring enzyme, the buffer re-adjusted and restriction continued with the high salt requiring enzyme.

Digestion was terminated by adding 1/5 volume of FDE (3Ø% Ficoll, Ø.Ø5% bromophenol blue, 1ØmM EDTA pH 7.Ø).

Digests of lambda derivatives were heated for five minutes at 68°C to melt the cohesive termini before electrophoresis.

Extra care was taken to ensure the complete digestion of genomic DNA. Several small samples (Ø.5µg) of the reaction were removed during the course of the incubation. These were run on a test-gel and full digestion assumed if the pattern of the final two samples was identical.

TABLE M.1 Specific activities of hybridization probes.

| Type of filter | cpm/μg DNA | Concentration of probe in hybridization solution / $ngml^{-1}$ |
|---|---|---|
| Genomic library 1st screen | $2x10^7$-$5x10^7$ | 10 |
| " " 2nd screen | $2x10^7$ | 10 |
| " " 3rd screen | $1x10^7$ | 10 |
| Southern transfers of Charon 4A MUP recombinants | $7x10^6$-$1x10^7$ | 10 |
| Northern transfers | $3x10^7$-$4x10^7$ | 20 |
| Southern transfers of mouse genomic DNA | $3x10^7$-$4x10^7$ | 20 |

Labelling DNA by nick translation

DNA was nicked using DNase I under conditions that gave approximately one nick per 500 - 1000 bp. Dilutions of DNase I were prepared in 50mM Tris pH 7.4, 100$\mu gml^{-1}$ BSA. Nicking reactions, with different concentrations of DNase I, were carried out in 66mM Tris pH 7.5, 6mM $MgCl_2$, 20$\mu gml^{-1}$ BSA for 7 minutes at 20°C. The reactions were terminated by adding EDTA, pH 7.5, to a final concentration of 4mM and the DNA recovered by phenol/chloroform extraction and ethanol precipitation. To identify the most favourably nicked DNA, a sample from each of the reactions was run on a 0.8% agarose gel. Nick translation was carried out using E.coli Polymerase I as described by Bishop (1979) except that the T4 DNA ligase step was omitted. [$^{32}$P]dCTP was the sole labelled nucleotide.

Hybridization probes were labelled by nick translation. Table (M.1) summarizes the specific activities achieved for each type of hybridization.

DNA sequencing

DNA fragments to be sequenced were cloned into M13mp9 (Messing and Viera, 1982). The dideoxynucleotide sequencing method of Sanger et al (1977) was used to sequence the single stranded templates essentially as described by Coulson and Winter(1982).

[$^{32}$P]dCTP was substituted for [$^{32}$P]dATP and the synthetic

"universal" primer (17 nucleotides long) was purchased from Uniscience.

## General methods

### TCA precipitation and scintillation counting of labelled DNA

DNA samples were added to 1.3ml of Ø.2M $Na_4PPi$, 115µgml-1 BSA, 15% W/V TCA and precipitated on ice for 15 minutes. The precipitated DNA was recovered on GF/C filters by vacuum filtration and the filters rinsed free from any residual unincorporated nucleotides with 5% TCA. Dried GF/C filters were counted in PPO/POPOP toluene counting fluid.

### G5Ø Spun Columns

To separate unincorporated nucleotides from labelled DNA, spun Sephadex G5Ø columns as described by Maniatis, Fritsch and Sambrook (1982) were used.

### Phenol/Chloroform extraction

This was carried out as described by Maniatis, Fritsch and Sambrook (1982).

Composition of solution and media not specified in text.

LB : 1% Difco Bacto tryptone, Ø.5% Difco Bacto yeast extract, 1% NaCl.

LB top : LB + 1% Difco agar

LB bottom : LB + 1.5% Difco agar

Phage buffer : Ø.3% $KH_2PO_4$, Ø.7% $Na_2HPO_4$ (anhydrous) Ø.5% NaCl, 1ØmM $MgSO_4$, 1mM $CaCl_2$, Ø.ØØ1% gelatin.

1 x TA : 5ØmM Tris, 2ØmM NaOAc, 1ØmM NaCl, 2mM EDTA, pH 7.9.

1 x TB : 9ØmM Tris, 9ØmM boric acid, 2mM EDTA, pH 8.3.

1 x MOPS : 2ØmM MOPS (Morpholinopropanesulphonic acid), 1mM EDTA, 5mM NaOAc, pH 7.Ø

1 x PB : 12mM $Na_2HPO_4$ (anhydrous), 8mM $NaH_2PO_4.2H_2O$.

1 x Denhardts : Ø.4% Ficoll, Ø.4% Polyvinylpyrrolidone, Ø.4% bovine serum albumin.

1 x SSCP : 12ØmM NaCl, 15mM $KH_2PO_4$, 2mM EDTA, pH 7.2.

1 x SSPE : 18ØmM NaCl, 1ØmM $Na_2HPO_4$ (anhydrous), 1mM EDTA,

pH 8.3.

1 x SSC : 150mM NaCl, 15mM NaCitrate, pH 7.0.

# Results

## Section 1 : Screening C57 genomic libraries for MUP genes.

Screening amplified pools A1 and A2. The library used was prepared by John Bishop by the method of Kemp et al (1979) from male C57 genomic DNA using Charon 4A as a vector. In brief, liver DNA was methylated to completion using EcoRI methylase to protect canonical EcoRI sites. The DNA was then partially digested with EcoRI under EcoRI* conditions so that the recognition specificity was reduced to the tetranucleotide NAATTN'. The DNA was sized on sucrose gradients, and fractions containing ∿15 kbp EcoRI* genomic fragments were ligated into Charon 4A EcoRI arms (Maniatis et al, 1978), and packaged by the method of Grosveld et al (1981).

Two pools, each of $2.4 \times 10^5$ recombinant bacteriophages (∿1.2 genome equivalents) were separately amplified in the host ED8654 (Murray et al, 1977) to give libraries A1 and A2. $8.5 \times 10^4$ and $1.5 \times 10^5$ recombinant bacteriophages from the respective libraries were screened with the MUP cDNA plasmid LVA325 as a probe. LVA325 consists of the 3' half of exon 4 and all of exons 5 and 6 as well as intron 5 of a group 1 gene cloned into the plasmid pPH207 (Clissold and Bishop, 1981).

The hybridization washes were carried out at low stringency (1 x SET,

TABLE R.1.1. Probability of finding single copy sequence in C57 unamplified library.

| Library | Total number of recombinant phage (corrected for "blues") | Phages per µg of eukaryotic DNA | Average length of inserts | Approximate size of mouse genome | Probability of finding single copy sequence* |
|---|---|---|---|---|---|
| Unamplified ♂ liver C57 genomic | $8.8 \times 10^5$ | $4.9 \times 10^5$ | 13 kb | $3 \times 10^6$ kb | 0.98 |

* Probability (P) was calculated by the method of Clarke and Carbon (1976) using the equation $P = 1 - (1-f)^N$, where f = fraction that each fragment represents of the genome and N = total number of recombinant phage.

68°C) so as not to discriminate against any members of the gene family. Seven positive plaques from library pool A1 and eight positive plaques from library pool A2 were picked and the bacteriophage purified. On characterization by restriction mapping and Southern blotting (discussed in the next section) it was found that several of the bacteriophages were identical in all aspects of cloning and mapping and that only 3 out of 7 bacteriophages from pool A1 and 1 out of 8 bacteriophages from pool A2 were different.

Screening of unamplified C57 library. The non-randomness of library pools A1 and A2 was attributed to preferential replication of certain recombinants at the amplification steps. It was therefore decided that an unamplified pool of the library should be prepared and screened.

1.8 µg of ∿15 kbp size-fractionated C57 genomic DNA was ligated to 4 µg of EcoRI Charon 4A arms, the DNA packaged and aliquots assayed by plating on the host ED8654. A total of $8.8 \times 10^5$ recombinants were obtained (∿4.4 genomic equivalents) giving a high efficiency of $4.9 \times 10^5$ recombinants per µg of eukaryotic DNA (Maniatis et al, 1978). The probability of finding a single copy sequence was calculated by the method of Clarke and Carbon (1976) and found to be 0.98 (Table R.1.1).

$8 \times 10^4$ bacteriophages of this unamplified library were plated and screened. Eight positive plaques were isolated and characterized as before. Six of the recombinant MUP bacteriophages were found to be different and two pairs of identical clones were recovered. This

TABLE R.1.2 MUP clones isolated from the C57 and BALB/c libraries.

| Library | Clone | Total number isolated | Size (kb) | EcoRI sites at phage insert boundary | | Comment |
|---|---|---|---|---|---|---|
| | | | | left arm | right arm | |
| Amplified C57♂ Liver | | | | | | |
| Pool $A_1$ | CL-1 | 2 | 13.5 | - | - | |
| | CL-2 | 2 | 13.3 | + | + | = BL-25 |
| | CL-4 | 2 | 11.8 | + | - | = BL-8 |
| Pool $A_2$ | CL-3 | 8 | 14.2 | + | - | |
| Unamplified C57♂ liver | CL-5 | 1 | 12.2 | - | - | |
| | CL-6;CL-13 | 1 + 1 | 16.6 | - | + | CL-6 = CL-13 |
| | CL-8;CL-9 | 1 + 1 | 12.0 | + | - | CL-8 = CL-9 |
| | CL-10 | 1 | 12.8 | - | - | |
| | CL-11 | 1 | 11.8 | - | - | |
| | CL-12 | 1 | 11.8 | + | + | |
| Amplified BALB/c ♂ liver | BL-1 | 2 | 14.2 | - | - | |
| | BL-2 | | 13.8 | + | - | |
| | BL-6 | | 11.4 | + | - | |
| | BL-7 | | 10.2 | - | - | |
| | BL-14 | | 13.5 | - | - | |
| | BL-15 | | 12.8 | + | - | |
| | BL-8 | | 11.8 | + | - | = CL-4 |
| | BL-25 | | 13.3 | + | + | = CL-2 |

indicated that amplification alone was not responsible for the observed non-randomness.

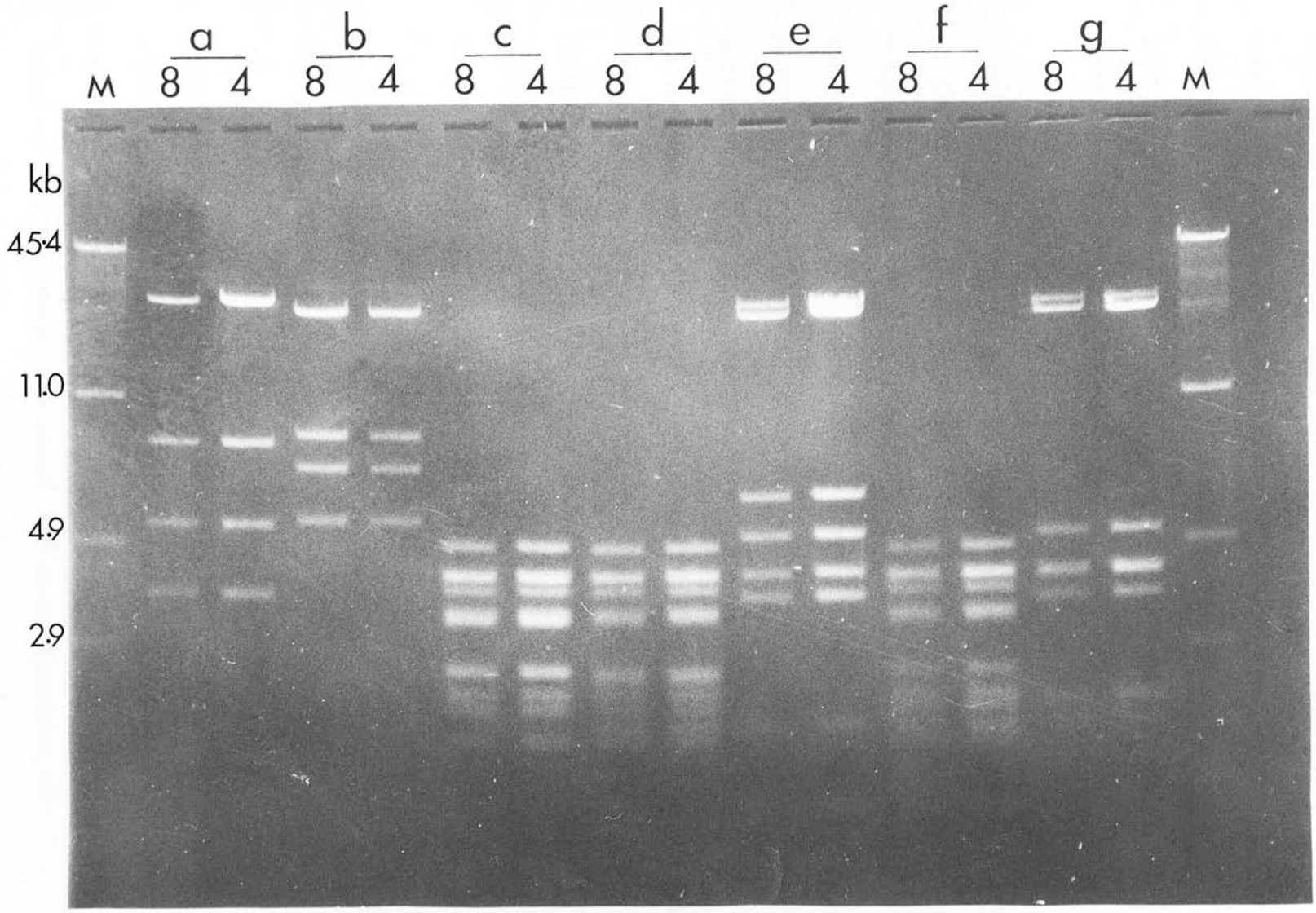
The recombinant MUP clones isolated from the C57 libraries were compared to clones isolated from a similarly constructed BALB/c genomic library (Clark, Clissold and Bishop, 1982) and the results are summarized in Table R.1.2.

Out of 20 clones which should have arisen from different molecular cloning events (i.e. excluding duplications within amplified pools) four different duplicate pairs were obtained. Three of these pairs included one bacteriophage from a C57 library and one from a BALB/c library: CL-2/BL-25, CL-4/BL-8, CL-1/BL-14 (Figs.R.1.1 and R.2.2). It is therefore concluded that the BALB/c library as well as the C57 libraries is non-random and that the cause of the non-randomness is common.

Library non-randomness. Possible reasons for the non-randomness of the libraries were suggested by the high number of reconstituted EcoRI sites at the insert/bacteriophage boundaries: 40% compared with the expected 20%. This could have arisen in a number of different ways as discussed below.

- Incomplete methylation. Under EcoRI* conditions the enzyme specificity becomes reduced from the full hexamer GAATTC to the central tetramer NAATTN'. The most rapidly cleaved sequence under these conditions is the canonical EcoRI site. Undermethylation however results in rapid cleavage of the unmethylated EcoRI sites.

Figure R.1.1. Restriction digests of BL-8 and CL-4. 0.25µg of the digested DNAs were electrophoresed on a 0.4% agarose gel and stained with ethidium bromide. Tracks labelled (8) and (4) correspond to the clones BL-8 and CL-4 respectively. The restriction digests are as follows: (a) BamHI, (b) HindIII, (c) PvuII, (d) HindIII + PvuII, (e) SstI, (f) SstI + PvuII and (g) HindIII + SstI. Tracks labelled (M) correspond to 0.25µg of marker DNA (Charon 4A partially digested with EcoRI and pCM2 digested with BamHI and EcoRI + BamHI).


a
b
c
d
e
f
g
M
8
4
8
4
8
4
8
4
8
4
8
4
8
4
M
kb
45·4
11·0
4·9
2·9

Undermethylation probably did not occur during the construction of the C57 and BALB/c libraries because after methylation, the DNA samples were incubated with excess EcoRI under normal EcoRI conditions and found to be unaffected. Furthermore, the reconstituted EcoRI sites at the ends of the inserts do not coincide with any of the EcoRI sites mapped on the MUP genes, except perhaps in the case of CL-8/CL-9.

- Preferential cleavage of a sub-population of NAATTN'. Polisky et al (1975) examined the cleavage of EcoRI* sites by analysing nearest-neighbour data obtained from ASV polymerase repair-synthesis of EcoRI* termini, and concluded that there was a hierarchy of recognition. Goodman et al (1977) explored this point further and found that the hierarchy for N in the EcoRI* recognition sequence NAATTN' was C>>T>A>>G. This means that the most rapidly cleaved EcoRI* site after the canonical EcoRI site is GAATTT = AAATTC. If the mouse DNA was cleaved mainly at these sites, then we would expect around 50% of all the termini to reconstitute EcoRI sites, in close agreement with the data.

The sequence GAATTT=AAATTC is expected to occur every 1029 bp in 60% AT-rich DNA. Let us suppose that MUP clones can be produced by one cleavage within a 9 kbp region to one side of the gene and a second cleavage within a 2 kbp region to the other. Given that cleavage occurs only at GAATTT sites, each gene would be found in about 36 different fragments. If there are 8 distinguishably different group 1 genes (see later) and an equal number of distinguishably different group 2 genes, we would have a potential

pool of about 576 clonable fragments. The probability of picking a duplicate clone from a sample of 8 (out of an unamplified library) would be 1/85. This calculation assumes a random distribution of sites and would easily be distorted, especially in a closely related gene family, if certain members were unexpectedly rich or poor in GAATTT sites or if sequences flanking these sites affected their rate of cleavage. However, the calculation gives a probability of =< 1/11400 for obtaining two or more pairs out of a sample of 8. Since two pairs were obtained from the unamplified library, it appears that while preferential cleavage at GAATTT may be a contributory cause to library non-randomness, additional factors must be involved in order to explain the results satisfactorily. Possible candidates are the rapid cleavage of non-conventional EcoRI* sites and the occurrence of certain sequences within the ligated fragments that influence the viability of the recombinant bacteriophage. These are discussed below.

- Cleavage of sequences other than NAATTN'. Woodbury et al (1980) reported that under EcoRI* conditions the hexanucleotide GGATTT was cleaved in preference to conventional EcoRI* sites. Gardner et al (1982) sequenced the genome of CaMV (strain CM184) by cloning CaMV EcoRI* fragments into M13mp2. By comparing the full sequence of CaMV with the ends of the CaMV EcoRI* cloned fragments, they were able to deduce the sequences of the EcoRI* sites that had been cleaved and successfully ligated. Cleavage at GGATTT was not confirmed by these authors although they detected other non-conventional EcoRI* sites: G(GC)ATTC and GA(CG)TTC.

The non-conventional EcoRI* sites G(GC)ATTC and GA(GC)TTC detected by Gardner et al, had been ligated with the EcoRI cut vector and true EcoRI sites had often been reconstituted by mismatch-repair in the E.coli host. The frequency of recovery of these sites at the junctions of cloned fragments was much less than that of conventional EcoRI* sites. It therefore seems unlikely that these sequences would have contributed significantly to our observed reconstituted EcoRI sites. However, if cleaved, they could contribute to the non-randomness of the libraries by forming fragment ends with a low cloning efficiency. The effect of cleavage of these sites is impossible to estimate quantitatively since there is no available data on their relative rates of cleavage.

- Chi sites. The four pairs of identical clones: BL-8/CL-4; BL-25/CL-2; CL-6/CL-13 and CL-8/CL-9, were also found to be identical with respect to the orientation in which they were cloned into the vector. This suggested that these recombinant bacteriophage may carry Chi sites. A Chi site is a short non-palindromic sequence (5'GCTGGTGG3') that has been shown to stimulate recombination in derivatives of lambda. Chi recombination is specific to and dependent on the RecA, RecBC pathway of E.coli (Stahl, 1979). It is orientation dependent in that if an active Chi site is inverted, its activity is greatly reduced (Faulds et al, 1979). This means that when a fragment containing a Chi sequence is cloned into a lambda vector, the Chi sequence is much more effective in one of the two possible cloning orientations. The orientation dependency of Chi is due to interaction between the non-palindromic sequence and other lambda sequences, notably the cos site (Kobayashi et al, 1983).

Chi test. Bacteriophages that are red$^-$ gam$^-$ are dependent on the Rec system of E.coli to form infective particles and to lyse their host. gam$^-$ bacteriophages do not enter the rolling circle mode of replication (since the RecBC product, Exonuclease V, is no longer inactivated) and the only substrate for packaging is the double stranded bacteriophage genome derived from theta replication followed by recombination. This process is normally inefficient and very small plaques are obtained on a Rec$^+$ host. The presence of a Chi site enhances recombination so that large, normal sized, plaques are obtained on the Rec$^+$ host (Stahl, 1979).

Since all our mouse genomic libraries were plated on a Rec+ host and because the vector Charon 4A is Chi$^-$, some of the non-randomness of the libraries could be due to selection of bacteriophage carrying functional Chi sites in the cloned fragment. This was tested by plating the clones on a Rec$^+$ host and comparing the sizes of the plaques to control bacteriophage identical to each other except for the presence (positive control) and absence (negative control) of a Chi site. To demonstrate that any variation in plaque size was due to recombination deficiencies, the bacteriophage were also grown on a recB$^-$ sbcA$^-$ host. The sbcA$^-$ mutation activates the RecE pathway so that the strain has a Rec+ phenotype and the recB mutation allows the transition to the rolling circle mode of replication. Therefore large plaques are formed on this host by red$^-$ gam$^-$ bacteriophage whether or not they carry active Chi sites. Both hosts used for the test were supF as the vector Charon 4A carries the genetic markers Aam32, Baml, in its left arm.

TABLE R.1.3 Chi test.

| λ phage | E. coli host | |
|---|---|---|
| | JM1 ($recB_{21}$,$sbcA_{20}$,supF) | QD5003 ($rec^+$,supF) |
| MMS 659 (b/453,cI857) | L | S |
| MMS 885 (b/453,cI857,$\chi^+$D) | L | L |
| C57 liver library | | |
| CL-1 | M | M |
| CL-3 | M | M |
| CL-6 | M | M |
| CL-13 | M | M |
| CL-5 | M | M |
| CL-10 | M | M |
| CL-8 | M | M |
| CL-9 | M | M |
| CL-11 | M | M |
| CL-12 | M | M |
| BALB/c liver library | | |
| BL-1 | M | M |
| BL-2 | M | M |
| BL-8 | S | S |
| BL-15 | M | M |
| BL-25 | M | M |
| BL-7 | M | M |
| BALB/c sperm library | | |
| BS-2 | M | M |
| BS-6 | M | M |
| BS-5 | M | M |
| BS-1 | M | M |
| BS-107 | M | M |

Key:- L = large plaques,
M = medium plaques,
S = small plaques.

Recombinant MUP bacteriophage from the BALB/c liver, the C57 liver and BALB/c sperm libraries were grown on the two bacterial strains at concentrations that allowed single plaques to be distinguished. Plaque sizes were compared with those of the controls and the results are summarized in Table R.1.3. All the recombinant bacteriophage that were tested produced plaques which were larger than those produced by the bacteriophage lacking a Chi site indicating that they probably carry Chi sites. The medium sized (as opposed to large) plaques given by these bacteriophages is thought to be due to the presence of two amber mutations in the vector whose effects are not completely overcome by the supF hosts. Bacteriophage BL-8 gave very small plaques on both hosts indicating that it is likely to have acquired a mutation unrelated to Chi and of unknown nature, sometime after the initial screen.

On average, a Chi site occurs every 70 kbp in eukaryotic DNA, and for a library composed of 12 kbp inserts one might expect one fifth of the recombinants to carry Chi sites . However, the MUP recombinants do not represent a random sample since they originate from a closely related gene family. Also selection for MUP recombinants carrying Chi sites would have taken place during library amplification, while in the case of the unamplified library the very small plaques produced by $Chi^-$ MUP recombinants would have made them difficult to detect during screening. If some MUP fragments are $Chi^+$ while others are $Chi^-$, then this could partly explain the strong bias in recoveries that was found.

Selection for Chi$^{+}$ recombinants, coupled with the non-randomness caused by EcoRI$^{*}$ sequence recognition in the case of the liver libraries, may explain why from a total of 25 clones (isolated using the same probe: LVA325) 3 distinguishably different group 2 MUP genes were isolated compared to 13 group 1 genes, when it has been estimated that there are equal numbers of group 1 and group 2 genes in the mouse genome.

Section 2: Restriction Analysis

The same restriction enzymes which had been used to map the MUP bacteriophages isolated from the BALB/c libraries (Clark et al, 1982 and Bishop et al, 1982) were used to map the C57 clones. When additional restriction enzymes were used, both the C57 and the BALB/c derived clones were mapped. Restriction sites were mapped by running single and double digests of the DNA on agarose gels. Most of the analysis was carried out on low percentage (Ø.4%) agarose gels. Where the restriction fragments of interest were under 2 kbp, Ø.8% agarose gels were used. Different clones were run in parallel so that small deletions/insertions (1ØØ bp - 5ØØ bp) in what appeared to be homologous fragments could be detected. Southern blots of the gels were hybridized to MUP cDNA clones (Clissold and Bishop, 1982) and to subclones previously prepared from the BALB/c MUP bacteriophages (Clark et al, 1982 and Bishop et al, 1982). The cDNA clones and subclones used and their origins are illustrated in Fig.R.2.1. The restriction maps of C57 and BALB/c genomic clones characterized with eight 6 base-pair restriction enzymes are shown in Fig.R.2.2. The structure of the transcription unit as determined for BS-6 (Clark et al, 1984a) is shown above the (bacteriophage) maps.

Some of the C57 clones are identical to each other and most are identical or similar to bacteriophages isolated from the BALB/c libraries. Thus CL-1 is identical in all respects to BL-14 and both these bacteriophages show no differences from BL-7, BS-6 and CL-3 in the regions where the clones overlap. Similarly CL-6 and CL-13 are

Figure R.2.1. cDNA clones (LVA325 and LVA132) and genomic subclones (BS-6-2, BS-6-5, BS-6-1-1, BS-6-3, BS-5B-3 and BL-1-4) used for library screening and characterization of isolated MUP genomic clones. All clones were constructed in pPH207 (Clissold and Bishop, 1981, 1982; Clark et al., 1982). LVA325 etc. correspond to known limits of hybridization. Arrow shows orientation and extent of the transcription unit (T.U.) of BS-6 (see Clark et al., 1984a).

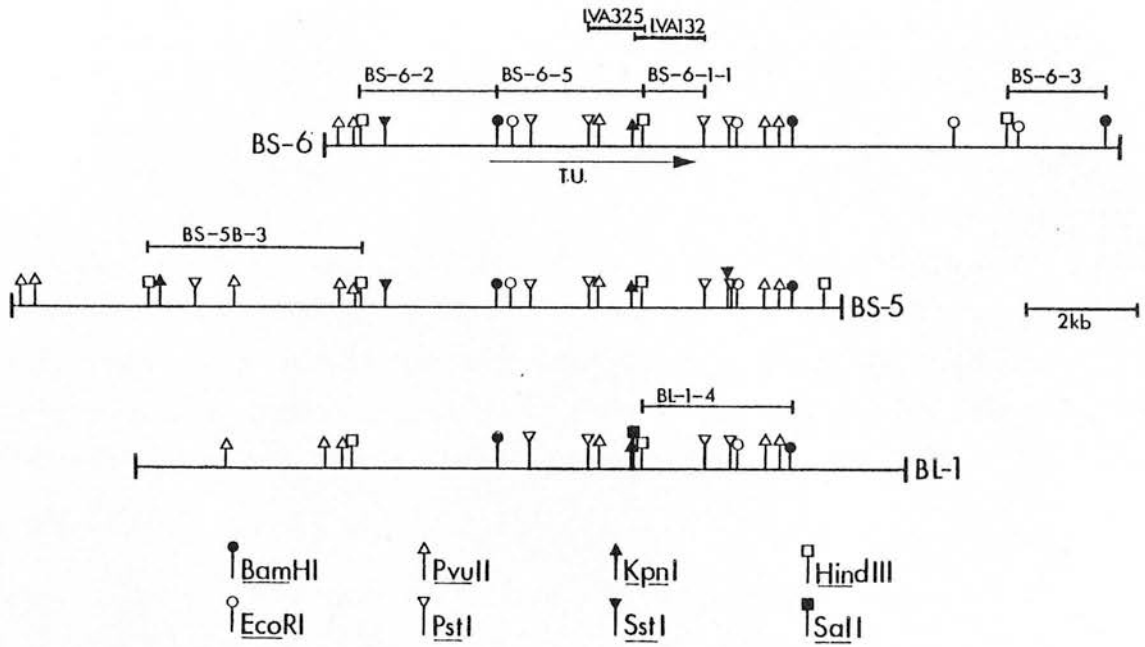
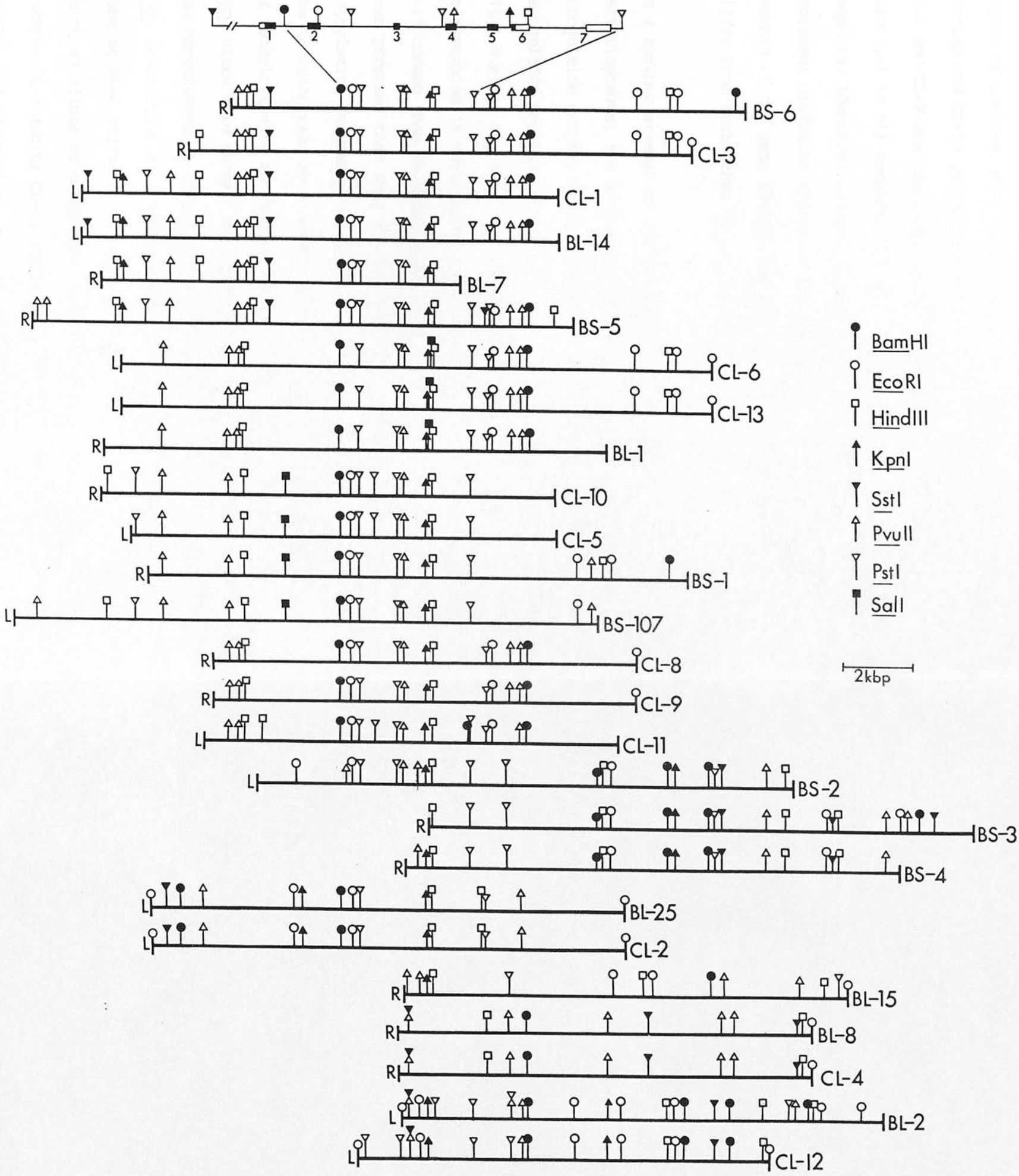

LVA325
LVA132
BS-6-2
BS-6-5
BS-6-1-1
BS-6-3
BS-6
T.U.
BS-5B-3
BS-5
2kb
BL-1-4
BL-1
BamHI
EcoRI
PvuII
PstI
KpnI
SstI
HindIII
SalI

Figure R.2.2.

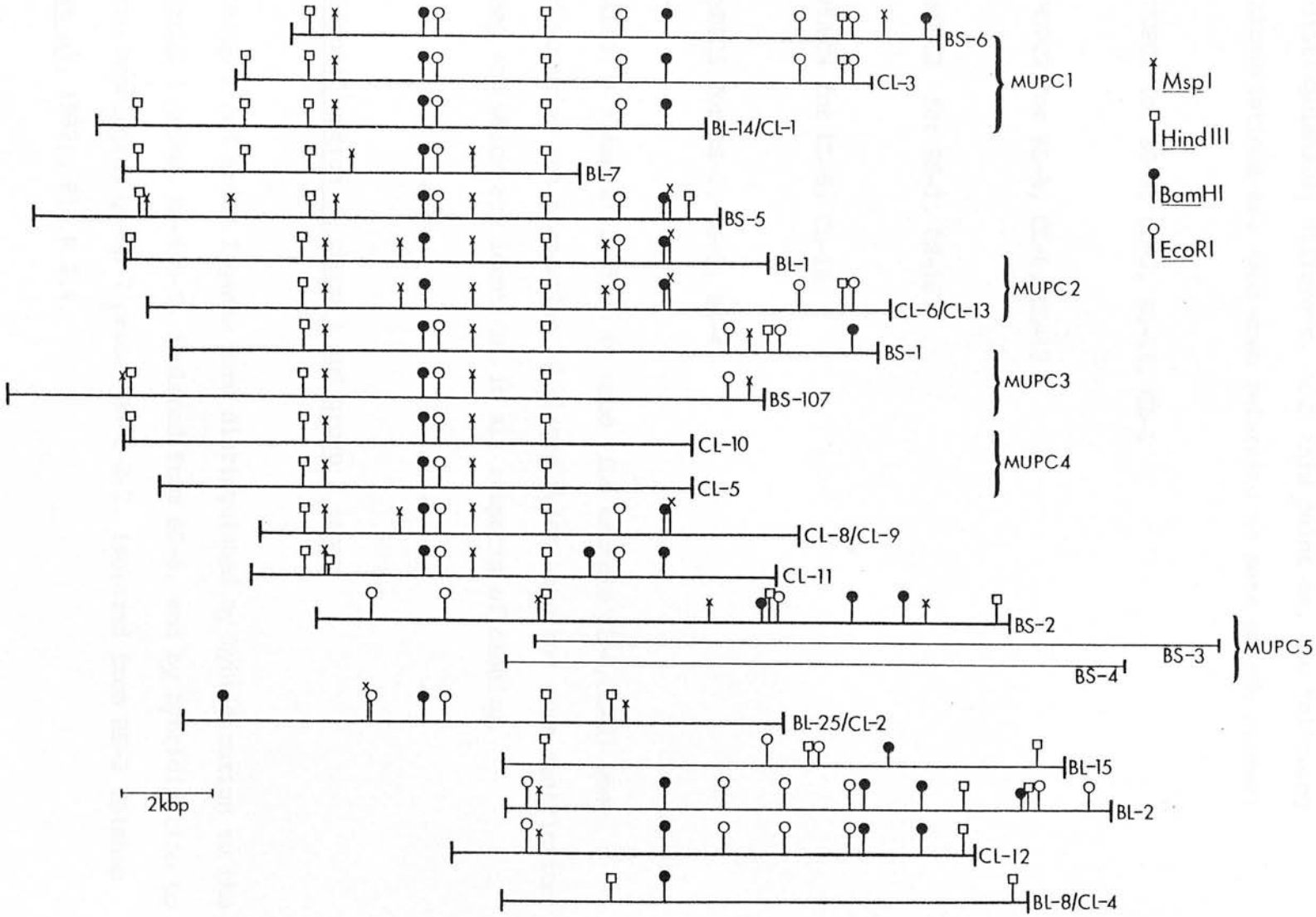
Restriction map of genomic clones isolated from the BALB/c and C57BL/Fa libraries. BS-clones were derived from a BALB/c sperm library, BL-clones were derived from a BALB/c liver library, and CL-clones were derived from C57 liver libraries. L, left arm of vector; R, right arm of vector. Reconstituted EcoRI sites at the insert/Charon 4A arm boundary are shown for the genomic clones isolated from the liver libraries.


1
2
3
4
5
6
7
R
BS-6
CL-3
L
CL-1
BL-14
R
BL-7
BS-5
L
CL-6
CL-13
R
BL-1
CL-10
L
CL-5
R
BS-1
L
BS-107
R
CL-8
CL-9
L
CL-11
BS-2
R
BS-3
BS-4
L
BL-25
CL-2
R
BL-15
BL-8
CL-4
L
BL-2
CL-12
BamHI
EcoRI
HindIII
KpnI
SstI
PvuII
PstI
SalI
2kbp

identical to eachother in all respects and identical to BL-1 in the region of overlap. BS-1 is identical to BS-1Ø7 in the region of overlap and CL-1Ø is identical to CL-5 in the region of overlap. CL-8 and CL-9 are identical in all respects, BL-25 and CL-2 are identical in all respects and BL-8 and CL-4 are identical in all respects. Identical clones isolated from the same strain may represent different clones of the same gene or closely related members of the gene family. They may also represent genes which differ from each other in uncloned regions.

In a further attempt to detect differences between the bacteriophages, the DNAs were restricted with MspI, a 4 base-pair recognition enzyme. The rare doublet CpG is present in the recognition sequence of this enzyme, making the frequency of MspI sites similar to that of 6 base pair recognition enzymes. Polymorphism in restriction sites carrying CpG in their sequences is more common than in those lacking CpG (Barker et al, 1984). It has been proposed that this may be partly attributed to the high methylation frequency of CpG to $^{m}$CpG. Methylation could increase the mutation rate by a number of mechanisms, including de-amination of methylcytosine to thymine (Salser, 1977; Barker et al, 1984). MspI sites were mapped onto the entire cloned regions of most of the bacteriophages (Fig.R.2.3). BL-7 was found to have a unique MspI restriction site and is therefore not a clone of the same gene as BS-6 and/or BL-14. However, this does not exclude it from being an allele of a MUP gene isolated from the C57 library (for example CL-3 and/or CL-1, see Fig.R.2.2). Thus from a total of twenty-four clones that were characterized with MspI and eight 6

Figure R.2.3. MspI restriction maps of the MUP genomic clones. BL-15 and BL-8/CL-4 do not contain MspI sites.


BS-6
CL-3
BL-14/CL-1
BL-7
BS-5
BL-1
CL-6/CL-13
BS-1
BS-107
CL-10
CL-5
CL-8/CL-9
CL-11
BS-2
BS-3
BS-4
BL-25/CL-2
BL-15
BL-2
CL-12
BL-8/CL-4
MUPC1
MUPC2
MUPC3
MUPC4
MUPC5
MspI
HindIII
BamHI
EcoRI
2kbp

base-pair recognition restriction enzymes, fourteen were distinguishably different. From this point on, the following abbreviations are used when referring to sets of MUP clones:

MUPC1 for BS-6, CL-3, BL-14, CL-1

MUPC2 for BL-1, CL-6, CL-13

MUPC3 for BS-1, BS-107

MUPC4 for CL-5, CL-10

MUPC5 for BS-2, BS-3, BS-4.

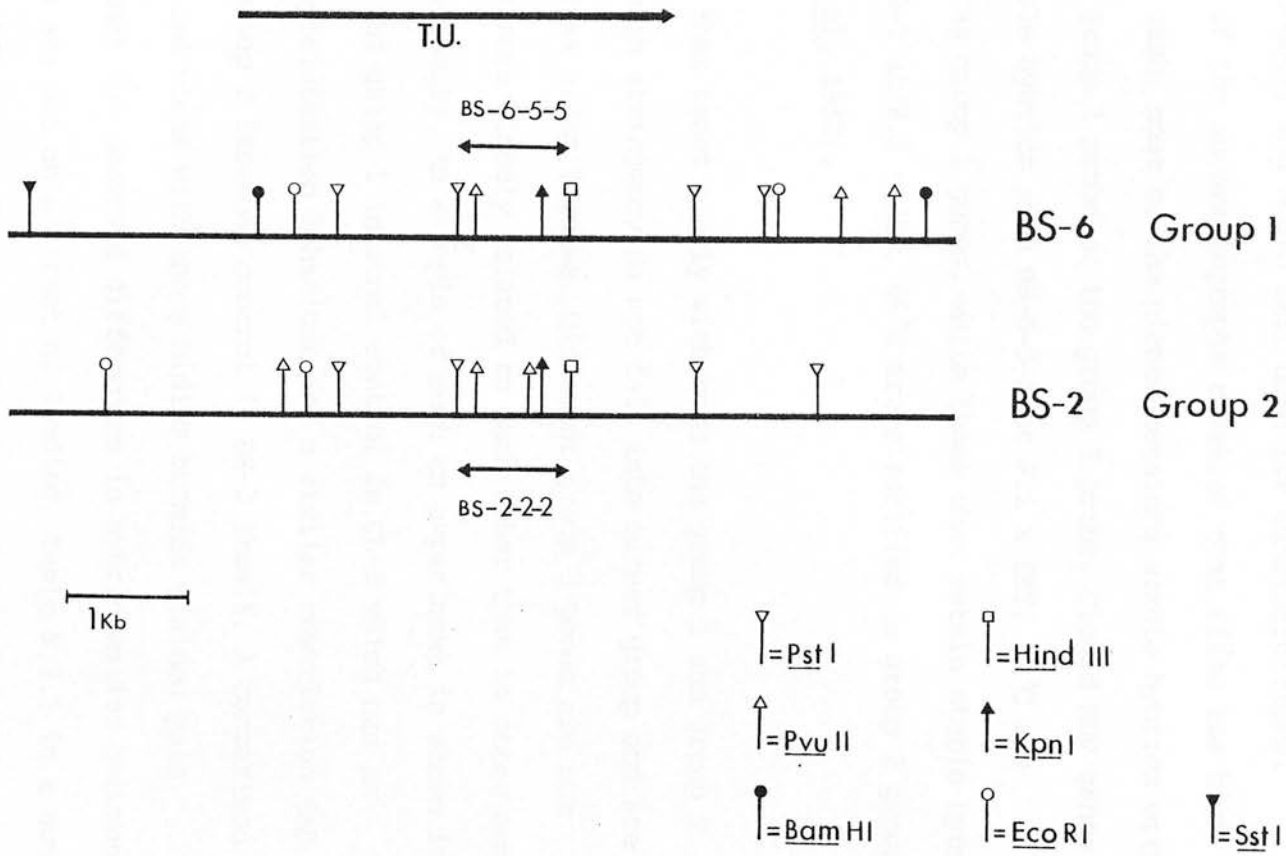
Also, a slash will often be used (for example CL-8/CL-9) when discussing the properties of clones which have the same restriction map and which are identical in all respects of cloning.

Identification of group 1 and group 2 genes

Group 1 and group 2 genes were distinguished by hybridization to the group 1 probe, BS-6-5-5, isolated from BS-6, and by hybridization to the homologous group 2 probe, BS-2-2-2, isolated from BS-2 (Bishop et al, 1982), Fig.R.2.4.

Duplicate filters of restricted bacteriophages were prepared and hybridized to either the nick translated group 1 or group 2 probe.

Figure R.2.4. Origin of the group 1 and group 2 probes (see Bishop et al., 1982). Arrow indicates orientation and extent of transcription unit.


T.U.
BS-6-5-5
BS-6
Group 1
BS-2
Group 2
BS-2-2-2
1Kb
= Pst I
= Hind III
= Pvu II
= Kpn I
= Bam HI
= Eco RI
= Sst I

Probes were nick translated to equal specific activities and an equal number of counts were sealed into the hybridization bags. The filters were washed down to 1 x SET, 68 °C (low stringency) and laid down for autoradiography. They were then washed to Ø.2 x SET, 68 °C (high stringency) and again laid down for autoradiography. Comparison of the autoradiographs revealed that after the high stringency wash, most of the clones retained stable hybrids with either the group 1 probe or the group 2 probe. Cloned MUP genes that retain stable hybrids with BS-6-5-5 at Ø.2 x SET, 68 °C are classified as group 1 genes, while those that retain stable hybrids with BS-2-2-2 at Ø.2 x SET, 68 °C are classified as group 2 genes (Bishop et al, 1982).

The clones that react poorly with both the group 1 and group 2 probes at high stringency do not fall into either group and are referred to as group 3 genes. Different group 3 genes are not necessarily more closely related to each other than to other members of the gene family. An example of such an experiment is shown in Fig.R.2.5. The group 1 internal control is CL-3 which has an identical hybridization behaviour and a similar restriction map to BS-6. The group 2 internal control is BS-2 itself. A comparison of the hybridized blots with the ethidium bromide stained gels indicates that the observed differences in hybridization between the two filters are not an artifact of loading. Table R.2.1 is a summary of the classification of MUP genomic clones isolated from the BALB/c and C57 libraries based on these hybridization studies.

Figure R.2.5. Hybridization of MUP genomic clones with the group 1 probe and the group 2 probe. 0.5µg DNA samples of the Charon 4A clones were digested with PstI and electrophoresed on a 0.4% agarose gel. The Southern transfers were hybridized with either the group 2 probe (a) or the group 1 probe (b) and washed under high stringency conditions (0.2 x SET, 68°C). (a') and (b') are photographs of the ethidium bromide stains of the gels prior to transfer. Lanes 1-15, DNA from Charon 4A clones BL-6 (1), BL-8 (2), BL-15 (3), BL-25 (4), BS-2 (5), BS-5 (6), BS-201 (7), BS107 (8), CL-3 (9), CL-5 (10), CL-6 (11), CL-8 (12), CL-10 (13), CL-11 (14) and CL-12 (15).

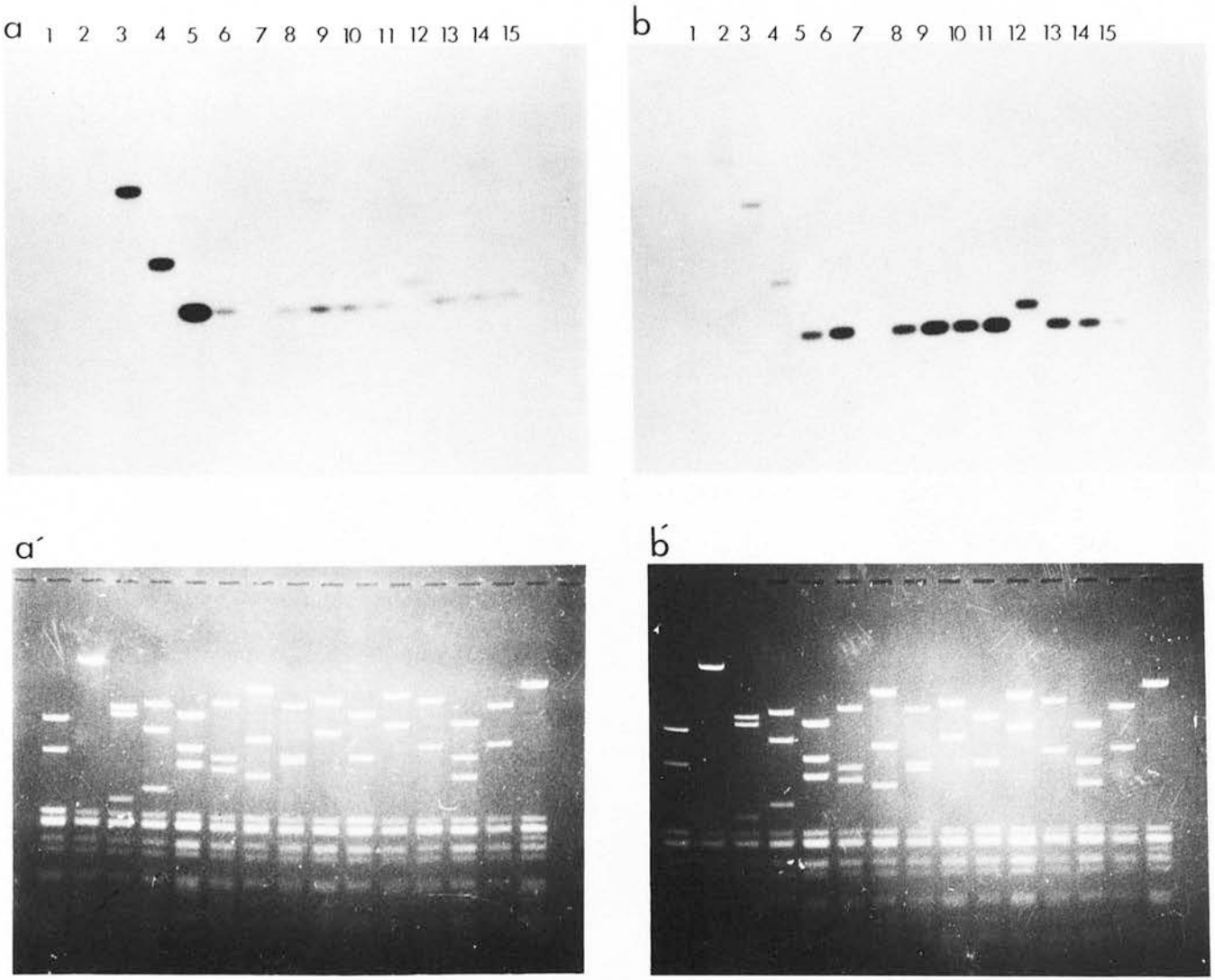

a
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
b
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
a´
b´

Table R.2.1 Classification of MUP genes based on hybridization criteria.

Group 1

MUPC1 (BS-6, CL-3, BL-14/CL-1)
BL-7
BS-5
MUPC2 (CL-6/CL-13, BL-1)
MUPC3 (BS-1, BS-107)
MUPC4 (CL-5, CL-10)
CL-8/CL-9
CL-11

Group 2

MUPC5 ( BS-2, BS-3, BS-4)
BL-25/CL-2
BL-15

Group 3

BL-8/CL-4
BL-2
CL-12

Comparison of the restriction maps of MUP genes

The coding region. Many of the restriction sites mapping to within the transcription unit of BS-6 are common to different cloned members of the gene family: the EcoRI site that maps to exon 2 is present in all group 1 and group 2 genes with the exception of genes identical and similar to CL-6; the PstI site mapping to intron II is common to all group 1 and group 2 genes; the PstI site mapping to intron III is common to all isolated MUP genes with the exception of BL-25/CL-2 and BL-8/CL-4; the PvuII site mapping to exon 4 is only lacking in BL-25/CL-2; the KpnI site mapping to intron V is only lacking in BL-8/CL-4 and the HindIII site mapping to the 3' end of exon 6 is present in all group 1 and group 2 genes, although it is not present in the group 3 genes BL-2, CL-12 and BL-8/CL-4.

In addition, group 1 genes share a BamHI site mapping to intron I (the BamHI site located at a similar position in BL-25/CL-2 is not homologous), and group 2 genes with the exception of BL-25/CL-2, share a PvuII site mapping to intron IV. The group 3 genes BL-2, CL-12 and BL-8/CL-4 all share a SstI site mapping to exon 4.

All MUP clones which were isolated hybridized to the three probes spanning the seven exons of BS-6 (BS-6-2, BS-6-5 and BL-1-4) unless they were truncated through cloning. This coupled with the homology of the restriction sites within the transcription unit of BS-6 suggests that most MUP genes have a similar structure.

The 5' flanking region

No differences in hybridization are detected between group 1 genes when hybridized to 5' flanking region probes, BS-6-2 and BS-5B-3. Restriction site homology in the 5'-flanking regions of the isolated group 1 genes extends to the bacteriophage arm boundaries, so that homology between CL-1/BL-14 and BS-5 is observed 6 kbp upstream of the cap site and is extended a further 2 kbp between BS-5 and BS-1Ø7. In some cases the homologies are interrupted by the presence of small insertions or deletions. A small insertion/deletion of ∿2ØØ bp is present in group 1 genes between the beginning of the transcription unit and the HindIII site positioned 2 kbp upstream from the cap site (Clark et al, 1982). This results in BL-7, BS-5 and MUPC1 having a "small" 5' HindIII fragment as opposed to the "large" 5' HindIII fragment present in all the other isolated group 1 genes. Homologies may also be interrupted by the accumulation of point mutations in a specific region. Such regions were not detected. However it must be pointed out that the detection of these is dependent on the presence of appropriately positioned restriction sites, since the extent of hybridization to a probe is determined by the proximity of flanking restriction sites to the homologous area. If such regions are present 5' to the group 1 genes then they are likely to be quite small, bearing in mind the homology of the restriction sites.

To summarize, the 5' flanking regions of group 1 genes are well conserved over a distance of at least 6 - 8 kbp upstream of the cap site.

The group 2 clones BL-25/CL-2 and BS-2 hybridize to the subclone BS-6-2, demonstrating homology within this group and between group 1 and group 2 genes in the 5' flanking region. BL-25/CL-2 also hybridizes to BS-5-B3. Thus although restriction enzyme polymorphism is observed between BL-25/CL-2 and isolated group 1 genes, 5'flanking homology between group 1 and some group 2 genes must extend at least 3.5 kpb upstream of the cap site. This has now been confirmed by electron microscope studies on the two more recently isolated MUP bacteriophages BS-1Ø2 and BS-1Ø9, each showing 5' linkage of a group 1 and a group 2 gene. Self annealing within each bacteriophage results in a stem of ∿4 kbp corresponding mainly to the homologous regions at the 5' ends of the group 1 and group 2 genes (Clark et al, 1984). This 5' flanking region therefore appears to be ancestral to the divergence of group 1 and group 2 genes and to the duplication event.

Whether or not the homologies found between group 1 and group 2 genes in the 5' flanking region are shared by the group 3 genes is not known since the corresponding regions of these genes have not been cloned.

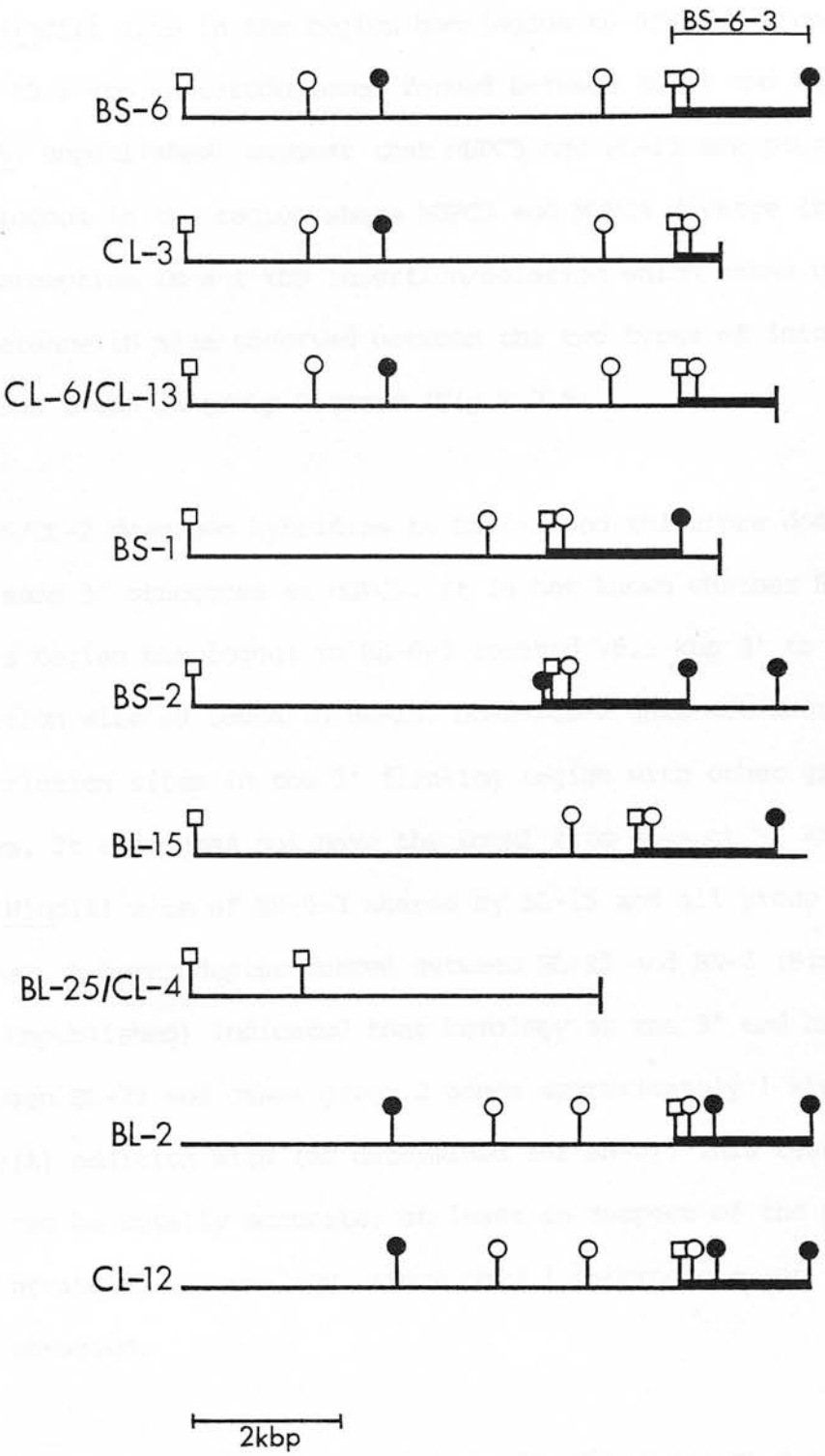
The 3' flanking region.

Homologies in the 3' flanking region also extend over a large distance away from the transcription unit. Amongst group 1 genes, homology is observed for at least 7 kpb between BS-6 and CL-6/CL-13. Homology breaks down approximately Ø.5 kbp 3' to the poly(A) addi-

tion site (as defined for BS-6) between genes with a 3' flanking region similar to BS-6 (MUPC1, MUPC2, BS-5, CL-8/CL-9, CL-11) and those with a 3' flanking region similar to BS-1 (MUPC3, MUPC4). It is regained at least 1 kbp 5' to the region that hybridizes with BS-6-3 (the precise point could not be accurately determined with the probes used). The distance from the poly(A) addition site to the HindIII site of the region homologous to BS-6-3 is shorter in genes similar to BS-1 (4 kbp) than in genes similar to BS-6 (5.7 kbp). This suggests that an insertion and/or deletion has taken place in the intervening region (Fig.2.6).

Homologies are also shared between group 1 and group 2 genes in the 3' flanking region (Clark et al, 1982). Thus, group 2 genes hybridize to the group 1 3'flanking region probes BL-1-4 and BS-6-3. The region homologous to BS-6-3 in MUPC5 is positioned at a distance from the end of the transcription unit equal to that in MUPC3 and MUPC4. However, heteroduplexes observed under the electron microscope between BS-1 and BS-2, show that the intervening regions are not homologous. Heteroduplexes formed between BS-2 and clones with a 3' region similar to BS-6 show that homology is maintained along the entire region that hybridizes to BL-1-4, while heteroduplexes formed between BS-1 and clones with a 3' region similar to BS-6 show that homology breaks down at $\sim$0.5 kbp from the end of the transcription unit (Bishop et al, unpublished). It therefore appears that the events that led to the equal positioning of the BS-6-3 homologous region in MUPC5 on the one hand and MUPC3 and MUPC4 on the other, are not the same.

Figure R.2.6. The 3' flanking region of MUP genomic clones. The regions shown extend from the end of the sixth exon (as determined for BS-6) to the 3' hybridization limit of BS-6-3 to BS-6. The heavy lines indicate the limits of hybridization of BS-6-3 to each of the clones. For symbols see Figure R.2.2.


BS-6-3
BS-6
CL-3
CL-6/CL-13
BS-1
BS-2
BL-15
BL-25/CL-4
BL-2
CL-12
2kbp

BL-15, a group 2 gene, hybridizes to both BL-1-4 and to BS-6-3. The distance from the poly(A) addition site (as determined for BS-6) to the HindIII site in the region homologous to BS-6-3 is unique, viz. ∿5.1 kbp. Heteroduplexes formed between BL-15 and BS-4 (Bishop et al, unpublished) suggest that MUPC5 and BL-15 are partially homologous in the region where MUPC3 and MUPC4 diverge from MUPC5. The exception is a 1 kbp insertion/deletion which makes up the difference in size observed between the two types of intervening regions found in group 2 genes (Fig.R.2.6).

BL-25/CL-2 does not hybridize to BS-6-3 and therefore does not share the same 3' structure as MUPC5. It is not known whether BL-25/CL-2 has a region homologous to BS-6-3 located ∿5.1 kbp 3' to the poly(A) addition site as found in BL-15. BL-25/CL-2 does not share any restriction sites in the 3' flanking region with other group 2 genes. It also does not have the EcoRI site present ∿1 kbp 5' to the HindIII site of BS-6-3 shared by BL-15 and all group 1 genes cloned. A heteroduplex formed between BL-25 and BS-2 (Bishop et al, unpublished) indicated that homology at the 3' end breaks down between BL-25 and other group 2 genes approximately 1 kbp 3' to the poly(A) addition site (as determined for BS-6). This result however may not be totally accurate, at least in respect of the position of the breakdown in homology, since only 1 heteroduplex of this type was observed.

Homologies in restriction sites are found between BL-2 and CL-12 over their entire cloned 3' ends. Homologies in restriction sites are also found between these two genes and group 1 and group 2 genes

in the regions that hybridize to the probes BL-1-4 and BS-6-3. The distance from the poly(A) addition site (as determined for BS-6) to the HindIII site in the regions homologous to BS-6-3 is identical to that observed in MUPC1 and MUPC2. BL-8/CL-4 have unique 3' restriction sites and do not hybridize to BS-6-3.

To summarize, homology in the 3' flanking regions is often interrupted. The points of breakdown in homology are usually different and the distance of the poly(A) addition site (as determined for BS-6) from the well-conserved region that hybridizes to the subclone BS-6-3 is variable. It is therefore concluded that interruption of homology in the 3' flanking region is largely due to a number of different insertion/deletion events. Such a large breakdown in homology is not observed in the more conserved 5' flanking region.

Section 3: Sequencing the 5' ends of some group 1 genes.

The 5' ends of three MUP clones: BL-7, CL-8 and CL-11 were sequenced. The region sequenced in each case consisted of the first exon and 80 - 100 bp of 5' untranscribed flanking DNA. The cloning strategy is outlined in Fig.R.3.1. The vector used was M13mp9 cut with HindIII and BamHI.

BL-7 was digested with BamHI and the 12 kbp BamHI fragment extracted by electroelution from a gel and purified as described in the Methods section. This fragment was then digested with HindIII, and the resulting BamHI-HindIII fragments (2.4 kbp and 5.5 kbp) were ligated with the vector using a 3 molar excess of fragments.

CL-8 was digested with BamHI and the 8.5 kbp fragment was extracted and digested with HindIII. The resulting BamHI-HindIII fragments (6 kbp and 2.6 kbp) were ligated with the vector.

CL-11 was digested with HindIII and the 4.8 kbp fragment was extracted and digested with BamHI. After purification, the BamHI-HindIII fragments (2.1 kbp and 2.7 kbp) were ligated with the vector.

The ligation mixes were separately used to transform competent JM101 cells. Single-stranded templates were prepared and clones of the appropriate sizes identified by electrophoresis in agarose gels, and sequenced by the method of Coulson and Winter, (1982). The

Figure R.3.1. Cloning strategy for sequencing the first exon and immediate 5' flanking region of BL-7, CL-8 and CL-11. Open arrows indicate cleavage with second restriction enzyme. Closed arrows indicate direction in which the *Hind* III-*Bam*HI fragments were sequenced after cloning into M13mp9. For restriction enzyme symbols not given, see Figure R.2.2.

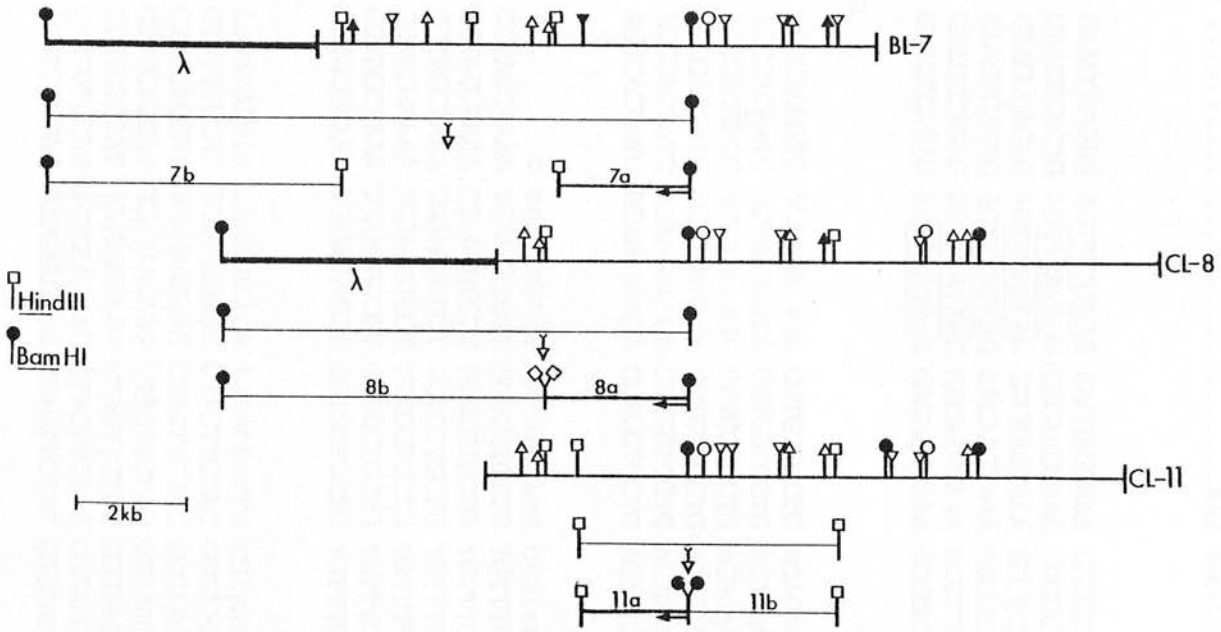

λ
BL-7
7b
7a
λ
CL-8
HindIII
BamHI
8b
8a
CL-11
2kb
11a
11b

Figure R.3.2. 5' sequences of group 1 genes.

```
BS-6   GAAGAGGGAA AAAAAAAAAA ACAAAACAAA CAACAACAAC .AAAAAAAAA
BL-7   GAAGAGGGAA AAAAAAAAAA ACAAAACAAA CAACAACAAC .AAAAAAAAA
BS-5   GAAGAGGG.. .AAAAAAAAA ACAAAACAAA CAACAAGAAC AACAAAAAAA
BL-1   GGAAGAGGG. .......... .......... .....AAAAA AAAAAAAAAA
CL-8   GGAAGAGGG. .......... .......... ......AAAA AAAAAAAAAA
CL-11  GGAAGAGGG. .......... .......... .......AAA AAAAAAAAAA


BS-6   AAA.CCCGCT GAACCCAGAG AGTATATAAG GACAAGCAAA GGGGCTGGGG
BL-7   ...CCCCGCT GAACCCAGAG AGTATATAAG GACAAGCAAA GGGGCTGGGG
BS-5   AAA.CCCGCT GAACCCAGAG AGTATATAAG GACAAGCAAA GGGGCTGGGG
BL-1   AAA...CGCT GAACCCAGAG AGTATATAAG GACAAGCAAA GGGGCTGGGG
CL-8   AAA...CGCT GAACCCAGAG AGTATATAAG GACAAGCAAA GGGGCTGGGG
CL-11  AAA.CCCGCT GAACCCAGAG AGTATATAAG GACAAGCAAA GGGGCTGGGG

          cap site
          *
BS-6   AGTGGAGTGT AGCCACGATC ACAAGAAAGA CGTGGTCCTG ACAGACAGAC
BL-7   AGTGGAGTAT AGCCACAATC ACAAGAAAGA CGTGGTCCTG ACAGACAGAC
BS-5   AGTGGAGTGT AGCCACGATC ACAAGAAAGA CGTGGTCCTG ACAGACAGAC
BL-1   AGTGGAGTGT AGCCACGATC ACAAGAAAGA CGTGGTCCTG ACAGACAGAC
CL-8   AGTGGAGTGT AGCCACGATC ACAAGAAAGA CGTGGTCCTG ACAGACAGAC
CL-11  AGTGGAGTGT AGCCACGATC ACAAGAAAGA CGTGGTCCTG ACAGACAGAC


                              m etlysmetle uleuleuleu cysleuglyl
BS-6   AATCCTATTC CCTACCAAAA TGAAGATGCT GCTGCTGCTG TGTTTGGGAC
BL-7   AATCCTATTC CCTACCAAAA TGAAGATGCT GCTGCTGCTG TGTTTGGGAC
BS-5   AATCCTATTC CCTACCAAAA TGAAGATGCT GCTGCTGCTG TGTTTGGGAC
BL-1   AATCCTATTC CCTACCAAAA TGAAGATGCT GCTGCTGCTG TGTTTGGGAC
CL-8   AATCCTATTC CCTACCAAAA TGAAGATGCT GCTGCTGCTG TGTTTGGGGC
CL-11  AATCCTATTC CCTACCAAAA TGAAGATGCT GCTGCTGCTG TGTTTGGGGC


          euthrleuva lcysvalhis alaGluGluA laSerSerTh rGlyArgAsn
BS-6   TGACCCTAGT CTGTGTCCAT GCAGAAGAAG CTAGTTCTAC GGGAAGGAAC
BL-7   TGACCCTAGT CTGTGTCCAT GCAGAAGAAG CTAGTTCTAC GGGAAGGAAC
BS-5   TGACCCTAGT CTGTGTCCAT GCAGAAGAAG CTAGTTCTAC GGGAAGGAAC
BL-1   TGACCCTAGT CTGTGTCCAT GCAGAAGAAG CTAGTTCTAC GGGAAGGAAC
CL-8   TGACCCTAGT GTGTGTCCAT GCAGAAGAAG CTAGTTCTAC GGGAAGGAAC
CL-11  TGACCCTAGT CTGTGTCCAT GCTGAAGAAG CTAGTTCTAC GGGAAGGAAC


          PheAsnValG luLys
BS-6   TTTAATGTAG AAAAGGTATG ATCACTGAAT AGTAGCTTCT GACTCAGAAT
BL-7   TTTAATGTAG AAAAGGTATG ATCACTGAAT AGTAGCTTCT GACTCAGAAT
BS-5   TTTAATGTAG AAAAGGTATG ATCACTGAAT AGTAGCTTCT GACTCAGAAT
BL-1   TTTAATGTAG AAAAGGTATG ATCACTGAAT AGTAGCTTCT GACTCAGAAT
CL-8   TTTAATGTAG AAAAGGTATG ATCACTGAAT AGTAGCTTCT GACTCAGAAT
CL-11  TTTAATGTAG AAAAGGTATG ATCACTGAAT TGTAGCTTCT GACTCAGAAT
```

sequences are given in Fig.R.3.2 along with the sequences of homologous regions from three other group 1 genes (Peter Ghazal and John Clark, unpublished). To confirm observed differences, parallel sequencing reactions of the two clones in question were run as illustrated by Fig.R.3.3. This was possible since the sequences of the 5' ends of all the genes were derived by cloning the 5' HindIII-BamHI fragment into M13mp9 and M13mp8. For the comparative gels the M13mp9 recombinants were used.

All six 5' group 1 sequences were found to be nearly identical. The 7 bp TATA box, present in almost all sequenced eukaryotic genes, was positioned 25 bp upstream of the cap site and was found to have an identical sequence (TATATAA) in all six group 1 genes. None were found to have the sequence GG(CT)CAATCT, which is located 7Ø-8Ø bp upstream from the transcription unit of many eukaryotic genes. Instead an A-rich region was found at this position (see Fig.R.3.2). The length and/or sequence composition of the A-rich region varied in all six genes causing the sequences to diverge at this point.

Further upstream of the A-rich region, precise homology was regained. The points at which the sequences of CL-8 and BS-6 diverge and come together again are marked on the sequencing gel illustrated by Fig.R.3.3. The major differences between the group 1 gene sequences were found to fall within the A-rich region. It is possible that some of the variation in expression between different members of the MUP gene family, may be attributed to variation within this region, bearing in mind its position relative to the cap site.

Figure R.3.3. Sequencing gel comparing the 5' flanking sequences of BS-6 and CL-8. Samples were sequenced according to Coulson and Winter (1982), and run on a 6% sequencing gel (pH 8.8) for five hours. Lanes from left to right, are: CL-8, G track; BS-6, G track; CL-8, A track; BS-6, A track; CL-8, T track; BS-6, T track; CL-8, C track; BS-6, C track. The point where the sequences diverge is indicated by arrow (a). The points where sequence homology is regained are indicated by arrows (b) and (b').

a
b
c

## Section 4: Phylogenic relationships of MUP genes based on restriction data

In the following section an attempt is made to draw up a phylogeny, based on restriction enzyme data, for some of the cloned MUP genes. This serves as an aid to understanding the evolution of the MUP gene family. The phylogenic relationships described here were derived using a parsimony method.

Parsimony methods find the evolutionary trees that would have been constructed through the least evolutionary change (see Felsenstein, 1982). Restriction enzyme data are best analyzed by the Dollo parsimony method (Felsenstein, 1983), which is based on the theory that in evolution complex structures are less likely to be gained than lost (Farris, 1977). To analyze the restriction data by the Dollo parsimony method, a computer program obtained from J.Felsenstein was used. The program is suitable for analyzing discrete two-state characters (0,1) where "0" indicates the absence of a character and "1" indicates its presence. If the character state is unknown, "?" may be substituted. The program is based on two main assumptions: (1) the ancestral state is 0, (2) the probability of a change of the form 1->0 is small but the probability of a change of the form 0->1 is even smaller. Other assumptions are that different characters evolve independently, and that lineages evolve independently. The algorithm allows up to one forward change (0->1) and as many reversions (1->0) as necessary. The program minimises the number of reversions (1->0). In these terms, the most parsimonious tree is that requiring the minimum

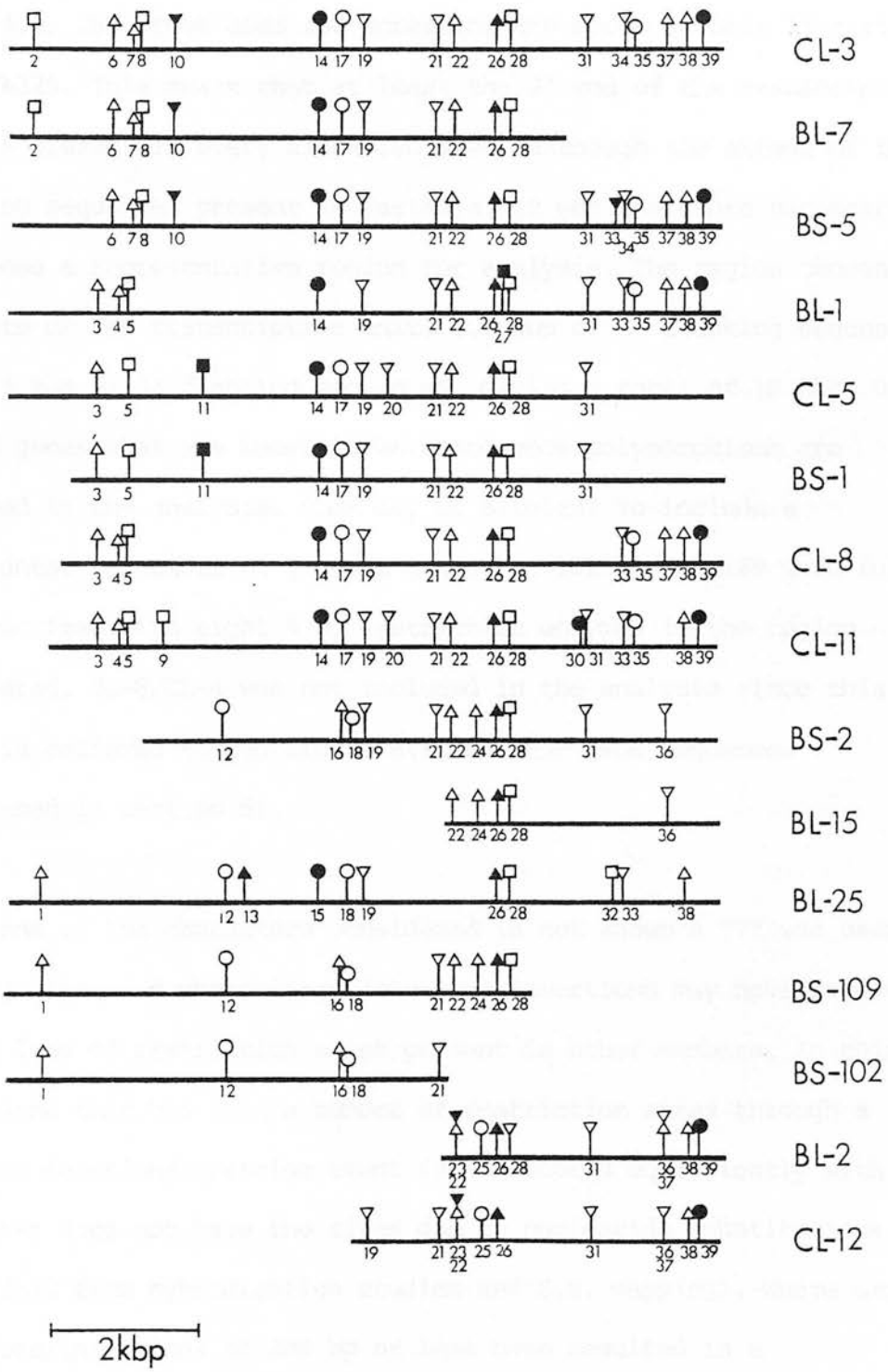
number of reversions.

The program constructs a tree as follows. The first two species in the input data are arranged into a bifurcating tree. The third species is added to one of three possible positions that maintain a bifurcating tree in such a way as to give the most parsimonious tree with three species. The fourth species is then considered, and its addition to one of the 5 possible positions that still result in a bifurcating tree is evaluated. Again the most parsimonious tree is chosen. The program continues in this manner except that after the addition of each new species a number of possible local re-arrangements of the internal segments of the tree are tried in an attempt to improve it. Because of the nature of the program , it is necessary to perform several runs changing the order of input of the same set of data. It is possible that the most parsimonious tree is not found, although the probability of this is small if only one or two forms of the tree are constructed, and if the data set is small. It is also possible to obtain more than one equally parsimonious tree. To choose between these, other data may have to be considered.

Finally, it must be pointed out the most parsimonious tree is not necessarily the most evolutionarily correct tree, since the assumptions may not always hold. For example, gene conversion could lead to the exchange of sites between members of the family which are not very closely related, but which share sufficient homology for conversion events to occur. The advantage of the method, however, is that it is possible to develop an objective phylogeny based on well defined assumptions which are thought to approximate

Figure R·4·1

Restriction enzyme data used to construct Dollo parsimony phylogenies.



most closely to the biological state.

The restriction site data which was analyzed by the Dollo parsimony method is given in Fig.R.4.1, and a table of characters is shown in Fig.R.4.2. The probe used for screening the mouse genomic libraries was LVA325. This means that at least the 3' end of the transcription unit is present in every clone isolated, although the extent of the flanking sequences present is variable. It was therefore necessary to choose a representative region for analysis. The region chosen consists of the transcription unit, 3.7 kbp of 5' flanking sequences and 2.5 kbp of 3' flanking sequences, giving a total of 10 kbp. Only cloned genes that are known to have sequence polymorphisms are included in the analysis. Some may be alleles. To include a representative number of group 2 genes, BS-102 and BS-109 were fully characterized with eight 6 bp restriction enzymes in the region considered. BL-8/CL-4 was not included in the analysis since this clone is believed to contain re-arranged MUP gene sequences (discussed in section 6).

Where one of the characters considered is not known a "?" was used. "?" was also used where large deletions/insertions may have resulted in the loss of restriction sites present in other members. In this way a gene that has lost a number of restriction sites through a possible deletion/insertion event is not scored equivalently with a gene that does not have the sites due to nucleotide substitutions (as judged from hybridization studies and E.M. mapping). Where small deletions/insertions of 200 bp or less have resulted in a displacement of sites, and where homology is otherwise maintained

(as judged from E.M. mapping), the site(s) are scored as new ones (see Fig.R.4.1). Scoring the group 1 sites that had been displaced by the proposed insertion/deletion event at the 5' flanking sequence as new sites did not affect the relationship of group 1 genes to other members of the gene family. This is because these sites are not shared by other members and because they represent a small proportion of the total number of group 1 restriction sites considered.

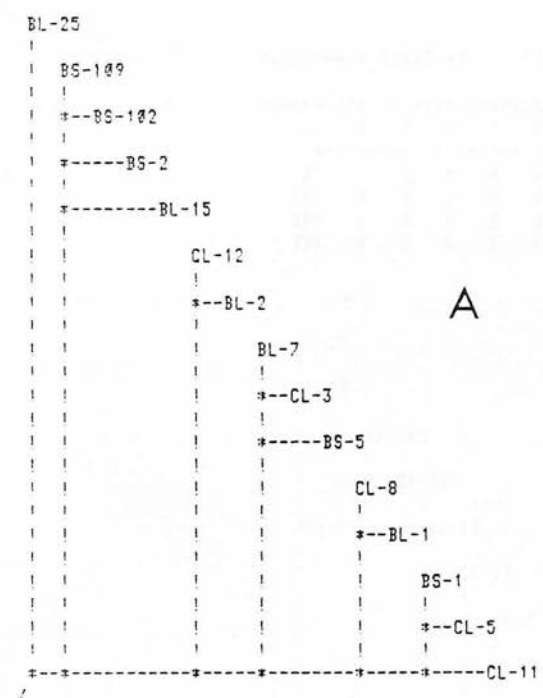
A minimum of 1Ø runs were considered for each set of data.The output format of the program is that of a rooted tree that "grows" from the bottom left hand corner of the diagram. The lengths of the branches are not proportional to evolutionary time. The total number of reversions required to construct the tree is given, as well as a table of the number of reversions experienced by each character.

Using all 39 characters, two equally parsimonious trees, A and B, were obtained (Fig.R.4.2 and R.4.3). These differ mainly in the position of the branch leading to BL-2 and CL-12, genes which hybridize equally poorly to group 1 and group 2 probes. In tree A, BL-2 and CL-12 are more closely related to group 1 genes. In tree B, BL-2 and CL-12 are more closely related to group 2 genes. In both trees A and B, the branch leading to BL-25, a group 2 gene, separates away from the rest of the gene family at the first fork.

A less parsimonious tree requiring one extra reversion is given in C (Fig R.4.3). In this case, BL-25 is associated more closely with other group 2 genes, while the branch leading to BL-2 and CL-12

Figure R·4·2

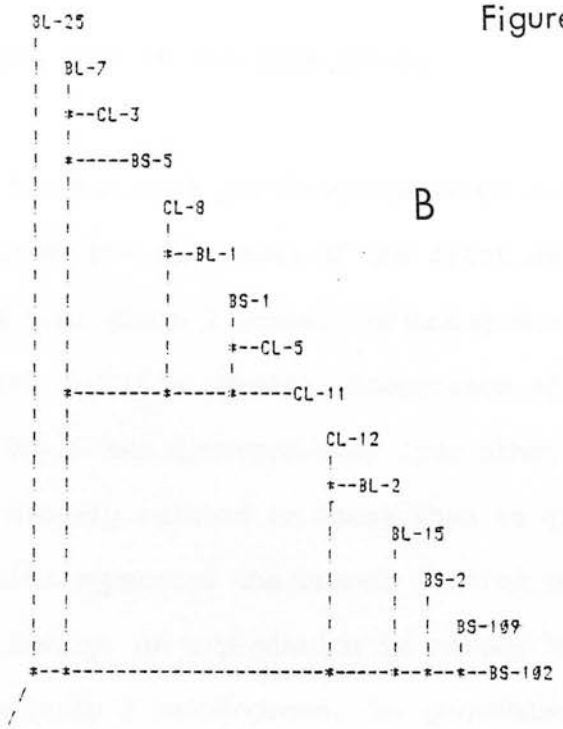
```
Dollo parsimony algorithm, version 2.4

15 species,  39  characters

Character-state data:

BL-7          01000 11101 00010 01010 11000 1010? ????? ????
CL-3          01000 11101 00010 01010 11000 10100 10011 0111
BS-5          00000 11101 00010 01010 11000 10100 10111 0111
BS-1          00101 00000 10010 01010 11000 10100 ????? ????
CL-5          00101 00000 10010 01011 11000 10100 ????? ????
BL-1          00111 00000 00010 00010 11000 11100 10011 0111
CL-8          ??111 00000 00010 01010 11000 10100 00011 0111
CL-11         ??111 00010 00010 01011 11000 10101 10011 0011
BS-2          ????? ????? ?1000 10110 11010 10100 10000 1???
BS-102        10000 00000 01000 10100 ????? ????? ????? ????
BS-109        10000 00000 01000 10100 11010 1010? ????? ????
BL-15         ????? ????? ????? ????? ?1010 10100 00000 1000
BL-2          ????? ????? ????? ????? ?1101 10010 10000 1111
CL-12         ????? ????? ????? ????0 11101 10000 10000 1011
BL-25         10000 00000 01101 00110 00000 10100 01010 0010
```




A

```
requires a total of     16.000

reversions in each character:
         0   1   2   3   4   5   6   7   8   9
     +-----------------------------------------
   0!        1   0   0   1   0   0   0   0   0
  10!    0   0   1   0   0   0   0   1   1   1
  20!    1   0   0   0   0   0   0   0   1   0
  30!    0   2   0   0   2   0   1   2   1   0
```

Figure R·4·3

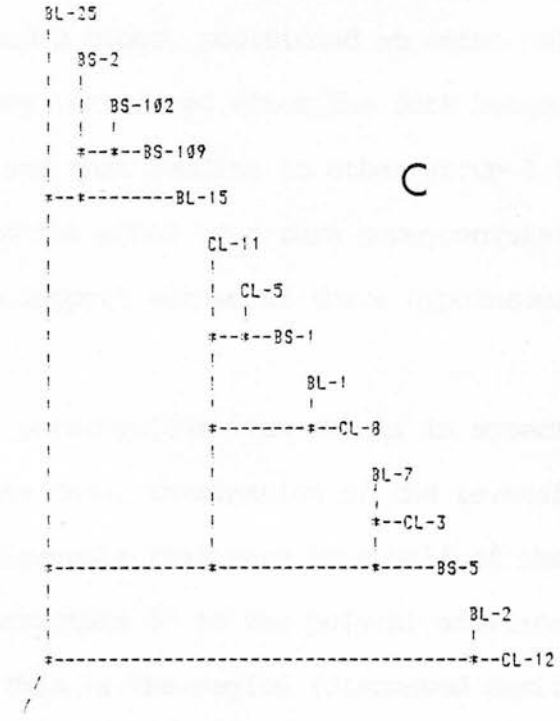



B

requires a total of 16.000

reversions in each character:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0! | | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10! | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| 20! | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 30! | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 3 | 1 | 1 |




C

requires a total of 17.000

reversions in each character:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0! | | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10! | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 20! | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30! | 0 | 3 | 0 | 0 | 1 | 0 | 2 | 3 | 1 | 1 |

forks away from the rest of the gene family.

The evolution of the two most parsimonious trees A and B is difficult to explain. The sequences of the first exon of eight group 1 genes and four group 2 genes, including BL-25, are known (Ghazal et al, 1985 and this thesis). Comparison of the sequences shows that while BL-25 has diverged away from other group 2 genes, it is still more closely related to these than to group 1 genes. If an early duplication separated the branch leading to BL-25 from the rest of the gene family, an explanation is needed for its greater similarity to the group 2 pseudogenes. One possible explanation is that group 1 genes underwent accelerated evolution compared to group 2 genes, and that group 2 genes are not pseudogenes (for example the hexapeptide encoded by group 2 genes could be functional, or group 2 genes could be spliced differently from group 1 genes). Another explanation would involve postulating that the stop codon common to all sequenced group 2 genes, positioned at amino acid 7 of the mature protein, was introduced after the fork between the branch leading to BL-25 and that leading to other group 2 genes but spread from one branch to the other by a rare gene conversion event. There is no evidence to support either of these hypotheses.

The slightly less parsimonious tree (C) is in agreement with all presently available data. Examination of the reversion tables for trees A, B, and C reveals that more than half of the reversions are experienced by characters 3' to the poly(A) addition site (A and B, 8/16; C, 11/17). This is the region (discussed earlier) that has diverged considerably in the gene family, possibly through a series

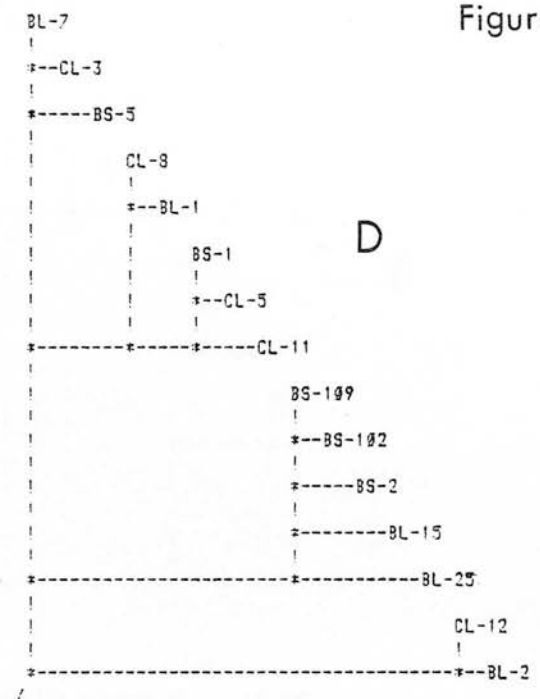
of different insertions and/or deletions.

If the 3' flanking region restriction data is omitted and only 5' flanking restriction sites and restriction sites present in the coding region are used for analysis, a single most parsimonious tree, D, is obtained (Fig.R.4.4). D has the same form as C: BL-25 does not separate away from the other group 2 genes before the split between group 1 and group 2 genes takes place, and BL-2 and CL-12 diverge away at an early stage from the rest of the gene family. Similarly, when the 5' 3Ø characters are used and BL-2 and CL-12 are omitted from the data set, only a tree of the form illustrated in E (Fig.R.4.4) is obtained. When BL-2 and CL-12 are omitted from the data set and all 39 characters are considered, two equally parsimonious trees, F and G, are obtained (Fig.R.4.5). F has the same general form as trees A and B, while G has the same form as C.

Examination of restriction sites present at the 3' end of the genes shows that BL-25 shares sites 33 and 38 (PstI and PvuII respectively) with group 1 genes. Site 38 is also shared by BL-2 and CL-12. It is not possible to determine whether this is simply the result of coincidence or the result of an event such as gene conversion. Sequencing of this region may allow us to distinguish between these possibilities.

In all the trees described so far, local re-arrangements within group 1 and group 2 sub-branches that do not affect the number of reversions or the general form of the trees, are observed. The two different arrangements observed between group 2 genes involve the

Figure R·4·4

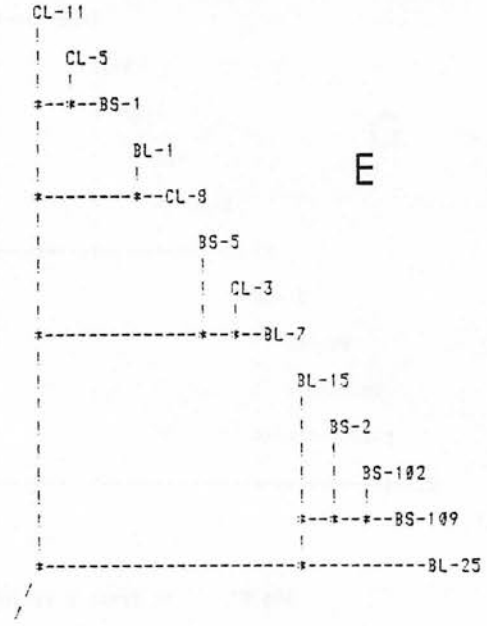



D

requires a total of 7.000

reversions in each character:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0! | | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10! | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 20! | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30! | 1 | | | | | | | | | |



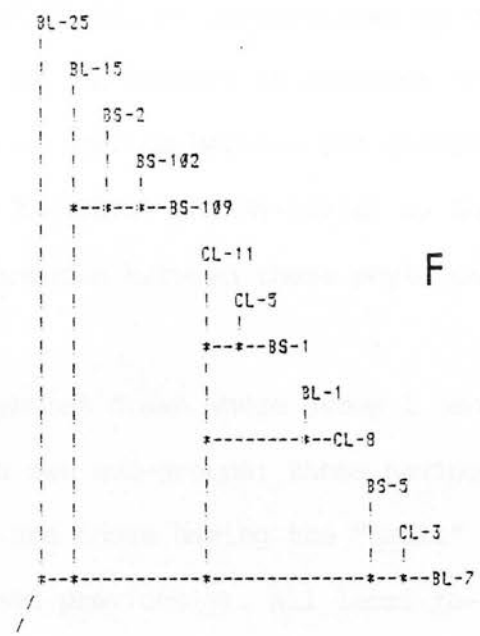

E

requires a total of 6.000

reversions in each character:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0! | | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10! | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 20! | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30! | 0 | | | | | | | | | |

Figure R·4·5

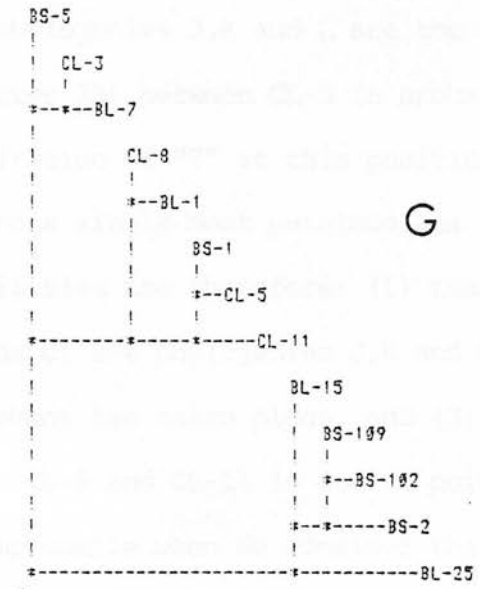



F

requires a total of 12.000

reversions in each character:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0! | | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10! | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| 20! | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30! | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |




G

requires a total of 12.000

reversions in each character:

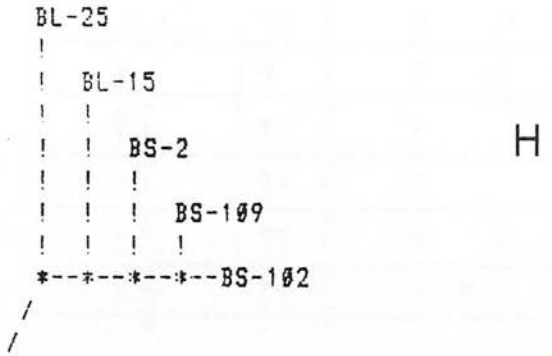
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0! | | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10! | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 20! | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30! | 0 | 3 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |

branching point of BL-15, as demonstrated by phylogenies H and I (Fig.R.4.6). Due to the absence of sequence data on BL-15 and due to the small amount of overlap between the cloned regions of BL-15 on the one hand and BS-102(2) and BS-109(2) on the other, it is not possible to distinguish between these phylogenies.

In all the phylogenies drawn where group 1 genes have been included, these divide into two sub-groups: those having the "large" 5' HindIII fragment and those having the "small" 5' HindIII fragment (discussed previously). All local re-arrangements take place within the sub-group of genes containing the large 5' HindIII fragment. Addition of MspI restriction site data to the character sets (Fig.R.4.7) still results in 3 equally parsimonious trees being drawn (Figs. R.4.7 and R.4.8).

The alternative phylogenies J,K and L are the result of a shared PstI site (character 20) between CL-5 (a probable allele of BS-1), and CL-11. Substitution of "?" at this position in either CL-11 or CL-5 gives rise to a single most parsimonious tree identical in form to K. The possibilities are therefore: (1) that the reversions illustrated by one of the phylogenies J,K and L took place, (2) a gene conversion event has taken place, and (3) the occurrence of the PstI site in both CL-5 and CL-11 is due to coincidence. The third possibility is improbable when we consider that substitution between group 1 genes is less than 1%. BS-1 and CL-5 share a 3' deletion/ insertion not common to other members of the gene family. This deletion/insertion event must have taken place after the split between the group 1 genes into the two sub-groups and is not

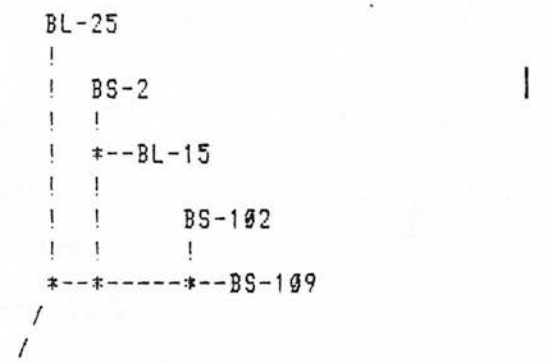
Figure R-4-6




requires a total of 1.000

reversions in each character:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

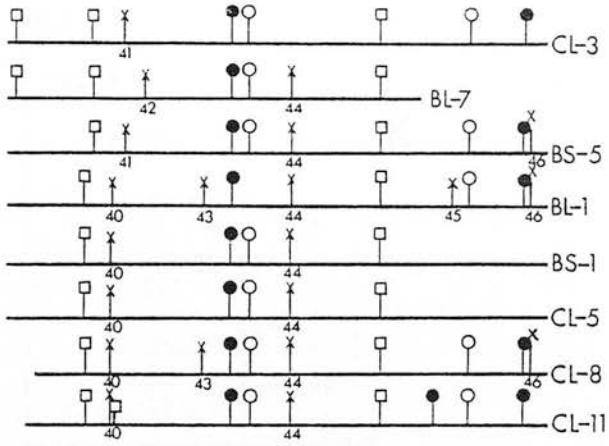



requires a total of 1.000

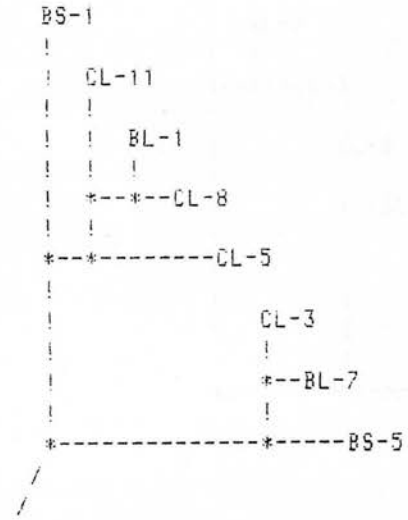
reversions in each character:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure R·4·7 MspI restrictio enzyme data

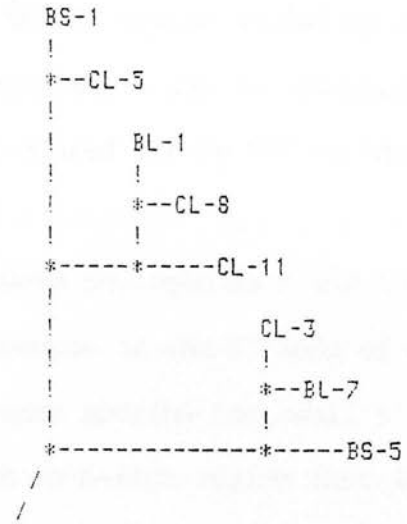



```
Dollo parsimony algorithm, version 2.4

 8 species,  46  characters

Character-state data:

CL-5          00101 00000 10010 01011 11000 10100 1???? ????1 0001? ?
BL-1          00111 00000 00010 00010 11000 11100 10011 01111 00111 1
CL-11         ??111 00010 00010 01011 11000 10101 10011 00111 00010 0
CL-8          ??111 00000 00010 01010 11000 10100 00011 01111 00110 1
BL-7          01000 11101 00010 01010 11000 1010? ????? ????0 0101? ?
CL-3          01000 11101 00010 01010 11000 10100 10011 01110 10000 0
BS-5          00000 11101 00010 01010 11000 10100 10111 01110 10010 1
BS-1          00101 00000 10010 01010 11000 10100 1???? ????1 0001? ?
```




```
requires a total of      9.000

reversions in each character:
          0   1   2   3   4   5   6   7   8   9
     *-----------------------------------------
    0!        0   0   0   0   0   0   0   0   0
   10!   0    1   0   0   0   0   0   1   0   0
   20!   1    0   0   0   0   0   0   0   0   0
   30!   0    1   0   0   0   0   0   1   0   0
   40!   0    1   0   0   1   0   2
```

Figure R-4-8

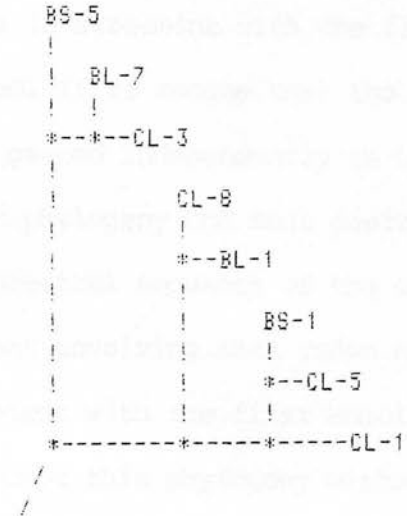



K

requires a total of 9.000

reversions in each character:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0! | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10! | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 20! | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30! | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 40! | 0 | 1 | 0 | 0 | 1 | 0 | 2 | | | |




L

requires a total of 9.000

reversions in each character:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0! | | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10! | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 20! | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30! | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 40! | 0 | 1 | 0 | 0 | 1 | 0 | 2 | | | |

ancestral to the 3' region shared by all other group 1 genes. On this basis phylogeny J may be rejected (the characters within the 3' deletion were scored for by "?" in BS-1 and CL-5).

To choose between phylogenies K and L we must turn to sequencing data. The sequences of the 5' ends of group 1 genes (Fig.R.3.2) reveal that genes sharing the small 5' HindIII fragment also have in common an A-rich region that is ∿42bp long, 10 bp upstream of TATA. On the other hand CL-11, CL-1 and CL-8 share an A-rich region of ∿17 bp, and BS-1 has an A-rich region of ∿55 bp (the 5' region of CL-5 has not been sequenced). This data is in agreement with both phylogenies K and L. Other shared homologies are: (1) a 2 bp substitution immediately 3' to the A-rich region common to CL-8 and BL-1 and (2) a silent substitution in the 10th amino acid (glycine) of the signal peptide common to CL-11 and CL-8.

Phylogeny K is in agreement with the first shared homology but not with the second. If we assume that the alternative glycine codon usage was not gained independently in CL-8 and CL-11, then in order to accept this phylogeny one must postulate that BL-1 has reverted back to the ancestral sequence of the glycine codon, or that a gene conversion event involving that codon has taken place. Phylogeny L is also consistent with the first homology but not with the second. In order to accept this phylogeny without suggesting that the alternative glycine codon arose independently in CL-8 and CL-11, one must postulate that both BL-1 and BS-1 have reverted to the ancestral glycine codon usage or that once again a gene conversion event involving that codon has taken place.

In the H-2 locus, a gene conversion event involvng a minimum of 13 and a maximum of 32 bases has been convincingly described (Weiss et al, 1983). If a gene conversion event has taken place between CL-8 and CL-11 then it would involve a maximum of 154 bases. However, due to the high homology between group 1 genes ( >= 99% ) it is not possible to prove or disprove whether such a 'micro' conversion event has taken place between the two group 1 genes.

The GGG glycine codon for amino acid 1Ø of the signal peptides of CL-8 and CL-11 could have arisen independently, particularly if there was a preference for glycine codon usage in MUP genes. An examination of the coding sequences of BS-6 and MUP15 shows that there is no significant difference from the random expectation of a frequency of Ø.25 for each of the four codons. However the sample number is small. [ BS-6 shares identical glycine codons with three other sequenced group 1 genes (J.Clark, unpublished) and with the liver group 1 cDNA clones p499 and plØ57 (Kuhn et al, 1984). MUP15 is a group 3 liver cDNA clone (isolated by A.Chave-Cox and discussed in following sections).]

In summary, the Dollo parsimony phylogenies are in agreement with the hybridization results, in that the three groups established by hybridization to the 1 kbp group 1 and group 2 probes are also distinguishable on the basis of restriction enzyme homology covering a much larger region. In addition, the group 1 genes are found to form two sub-groups based on restriction site polymorphisms in their 5' flanking sequences. BL-25 appears to have diverged from other

cloned group 2 genes as confirmed by a limited amount of sequencing data.

The results argue against extensive "homogenization" events between group 1 and group 2 genes. However, that such events may occur between groups is suggested by some restriction sites located 3' to the transcription unit. Also the possibility exists of gene conversion between group 1 genes sharing the large 5' HindIII fragment.

The phylogenies also suggest that BL-2 and CL-12 have diverged equally from group 1 and group 2 genes, although further 5' sequences are necessary to confirm this.

Section 5: Hybridization of group 1 and group 2 probes to mRNA isolated from different tissues.

Shaw et al (1983) reported the transcription of MUP genes in five tissues other than liver: lachrymal, submaxillary, sublingual, parotid and mammary. None expressed MUP at a level equivalent to that found in adult male liver. Estimates for the more abundantly transcribing glands at developmental stages where MUP expression was at a maximum, indicated that 1/10th, 1/24th and 1/30th of the steady state level was achieved in lachrymal, submaxillary and mammary glands respectively compared with male adult liver. In their sexual and developmental patterns and in their hormonal regulation of MUP mRNA, these three tissues were found to be different from the liver.

In vitro translation of hybrid-selected mRNA in the presence of dog pancreas membranes indicated that different MUP proteins were expressed in different tissues. Hybrid-selected mRNA from the submaxillary gland appeared to code for a MUP protein(s) that co-migrated with one from male liver, while hybrid-selected mRNA from mammary tissue coded for a protein(s) that co-migrated with the most predominant female liver protein, also detected in male liver. Unlike other tissues, the lachrymal gland exhibited a new set of MUP proteins which had higher pIs ( 5.6 - 6.5 ) than those of the liver ( 4.4 - 4.8 ) suggesting that these were coded for by a different set of MUP genes. Since there was no evidence at the time to suggest that the group 2 genes are pseudogenes, it was of particular interest to determine whether the lachrymal gland proteins were coded for by group 2 genes. It was also of interest to determine

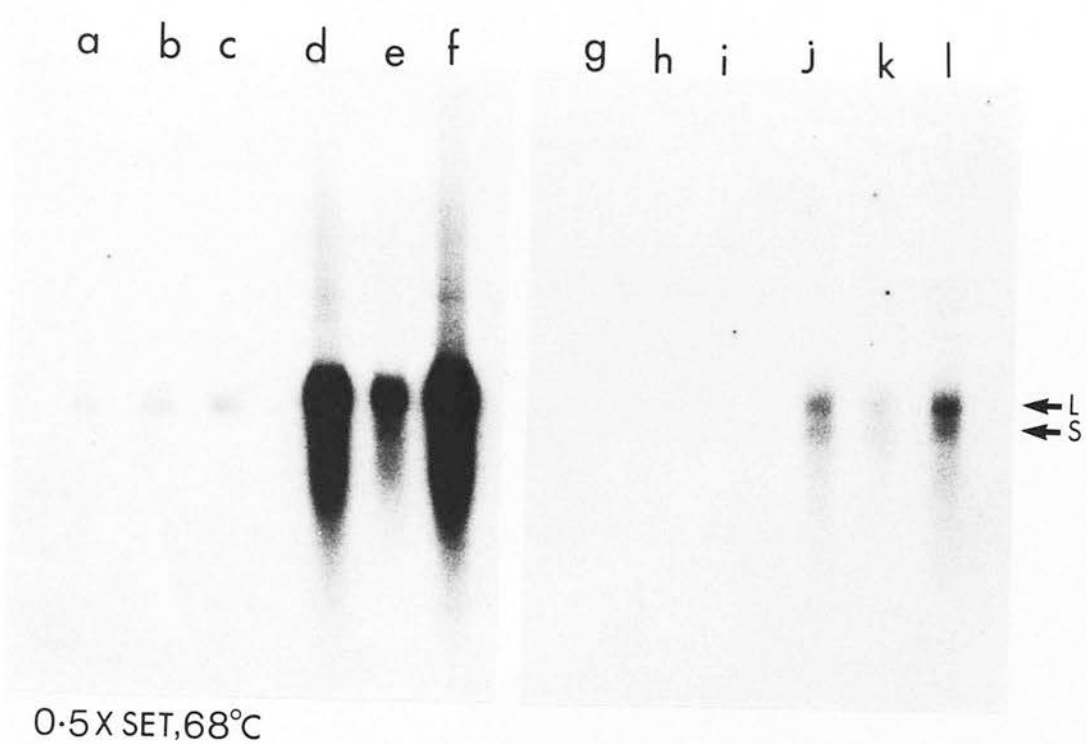
whether the mammary gland and submaxillary gland proteins were coded for by group 1 genes as liver mRNA appeared to be predominantly the product of group 1 gene expression (Clissold and Bishop, 1982 and unpublished results).

Poly(A) mRNA extracted from the lachrymal and submaxillary glands of 7 - 1Ø week old male mice and from the mammary tissue of eight-week old eighteen-day pregnant female mice was electrophoresed against poly(A) mRNA extracted from the livers of 7 - 1Ø week old male mice (gift of A. Duncan). All mRNAs were derived from BALB/c mice. Duplicate filters were hybridized with either nick translated BS-2-2-2 (group 2) or BS-6-5-5 (group 1). The probes were nick translated to the same specific activity and equal amounts of radioactivity were added to the hybridization bags. After the hybridization period, the filters were washed at high stringency (Ø.5 x SET, 68°C). Hybridization signals were detected for the lachrymal gland, mammary gland and submaxillary gland mRNA samples only when they were hybridized with the group 1 probe, indicating that expression in these tissues was predominantly from genes more homologous to the group 1 probe (Figs. R.5.1 and R.5.2).

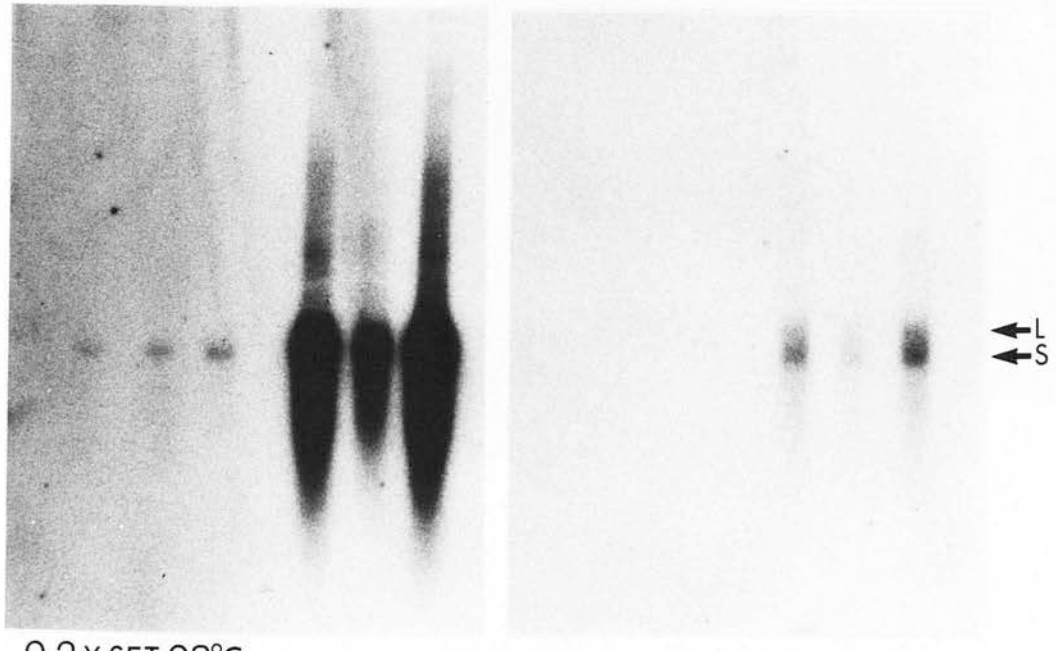
The results do not exclude the presence of low levels of mRNA homologous to group 2 genes in these tissues. In view of the differences between the pIs of the lachrymal and liver MUP proteins, the filters were washed at the higher stringency of Ø.2 x SET, 68°C. It was not possible to determine whether there was a difference in the signals from the mammary gland and submaxillary gland samples, due to the low amounts of MUP mRNA expressed in these tissues and

Figure R.5.1. Northern blots of liver and lachrymal gland poly(A) mRNA probed with either the group 1 or the group 2 probe. Lanes a - f were probed with the group 1 probe, BS-6-5-5; lanes g - l were probed with the group 2 probe, BS-2-2-2. Filters were washed down from 0.5 x SET,68°C to 0.2 x SET,68°C and exposed to X-ray film after each wash. Samples in their respective lanes are given below.

| | | |
|---|---|---|
| a and g : | 2µg | lachrymal gland poly(A) mRNA |
| b and h : | 3µg | " " " " |
| c and i : | 5µg | " " " " |
| d and j : | 1µg | liver poly(A) mRNA |
| e and k : | 0.5µg | " " " |
| f and l : | 2µg | " " " |



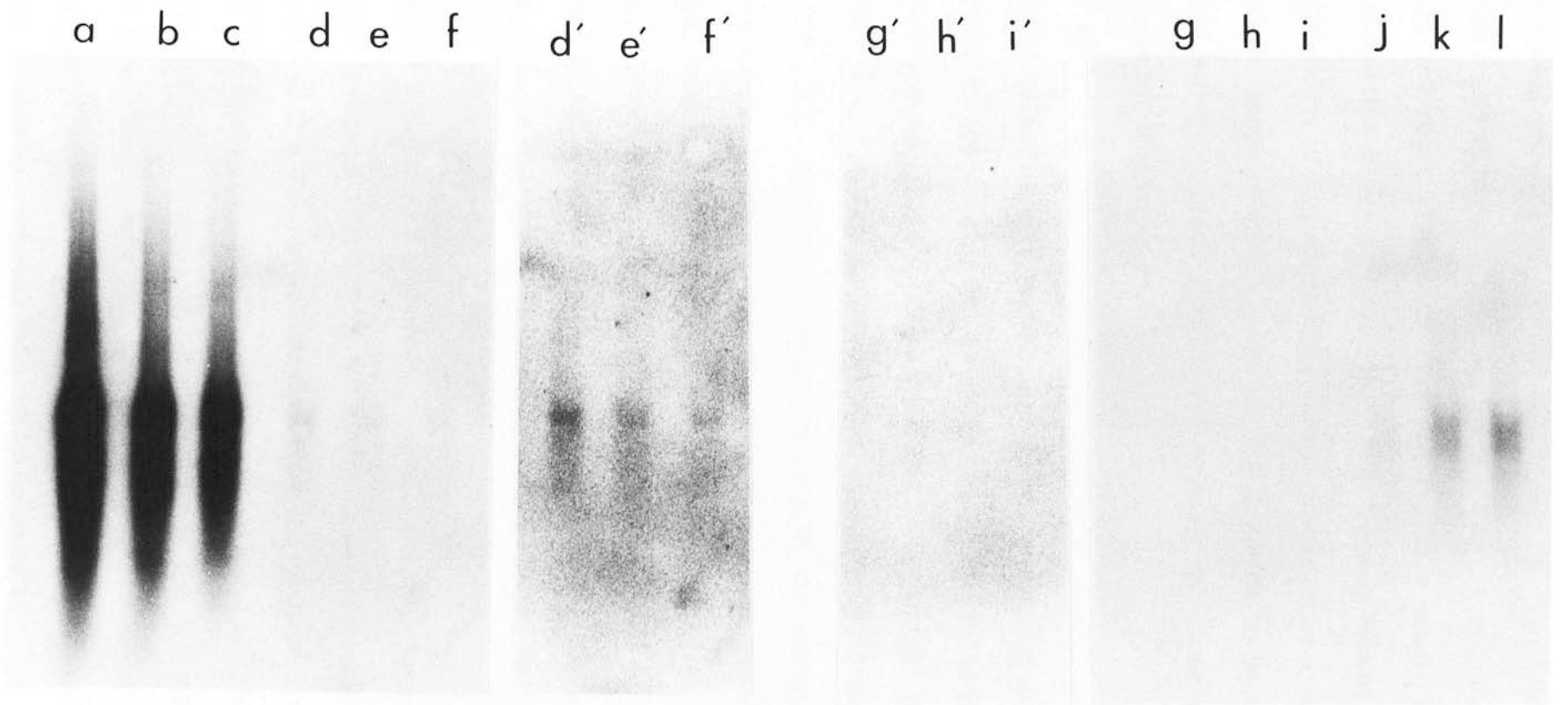

0·5 X SET, 68°C




0·2 X SET, 68°C

Figure R.5.2. Northern blots of liver, mammary gland and submaxillary gland poly(A) mRNA probed with either the group 1 or the group 2 probe. Lanes a - f were probed with the group 1 probe, BS-6-5-5; lanes g - h were probed with the group 2 probe, BS-2-2-2. Samples in their respective lanes are given below.

a and l : 1µg liver poly(A) mRNA (degraded)
b and k : 0.5µg " " " "
c and j : 0.2µg " " " "
d and i : 12µg mammary gland poly(A) mRNA
e and h : 6µg " " " "
f and g : 6µg submaxillary gland poly(A) mRNA
d', e', f', i', h' and g', higher contrast film photographs of the mammary and submaxillary samples d, e, f, i, h and g respectively.


a b c d e f
d′ e′ f′
g′ h′ i′
g h i j k l
0 5 X SET, 68°C

due to under-exposure of the autoradiograph. A significant difference was detected, however, in the hybridization signals of the long and short liver mRNA (Fig.R.5.1). This result indicates that unlike the group 1 probe, the group 2 probe hybridizes preferentially to the short mRNA. Also, judged from a densitometer scan of the autoradiograph, some signal appeared to have been lost from the lachrymal gland samples relative to the liver samples after the higher stringency wash.

Kuhn et al (1984) have reported significant levels of sequences homologous to the male liver derived cDNA, p199, in lachrymal mRNA. Hybridization of lachrymal mRNA and liver mRNA to a 5' subclone of this cDNA (Fig.R.5.3.B) indicated that at a washing stringency of Ø.5 x SET, 68°C, sequences homologous to p199 are comparatively more abundant in the lachrymal gland than in the liver. The size of the lachrymal mRNA was also judged to be in between that of the long and short liver mRNA when probed with this subclone. Signal from the lachrymal gland samples was lost relative to the liver samples on washing down to the higher stringency of Ø.2 x SET, 68°C.

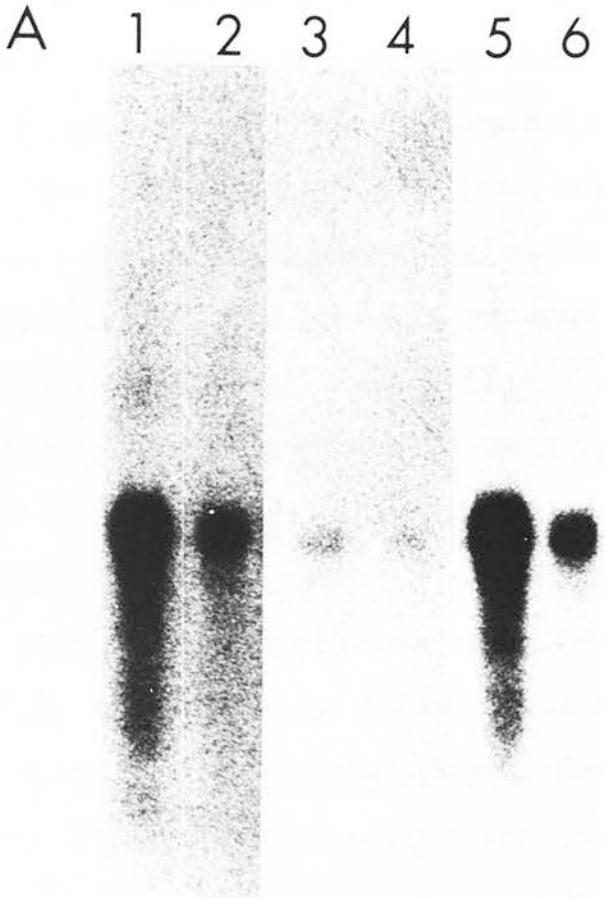
When the 5' subclone of the group 1 cDNA, p499, was hybridized to liver and lachrymal gland mRNA and washed under the higher stringency (Ø.2 x SET, 68°C), signal from the lachrymal gland samples was again lost relative to the liver samples (Fig.R.5.3.B). From its sequence, p499 is known to represent the transcript of a group 1 gene. The 5' subclone of p499 extends from amino acid -1Ø (of the signal peptide) to the PvuII site in exon 4. BS-6-5-5 and the p499 5' subclone overlap over 52 bp in exon 4.

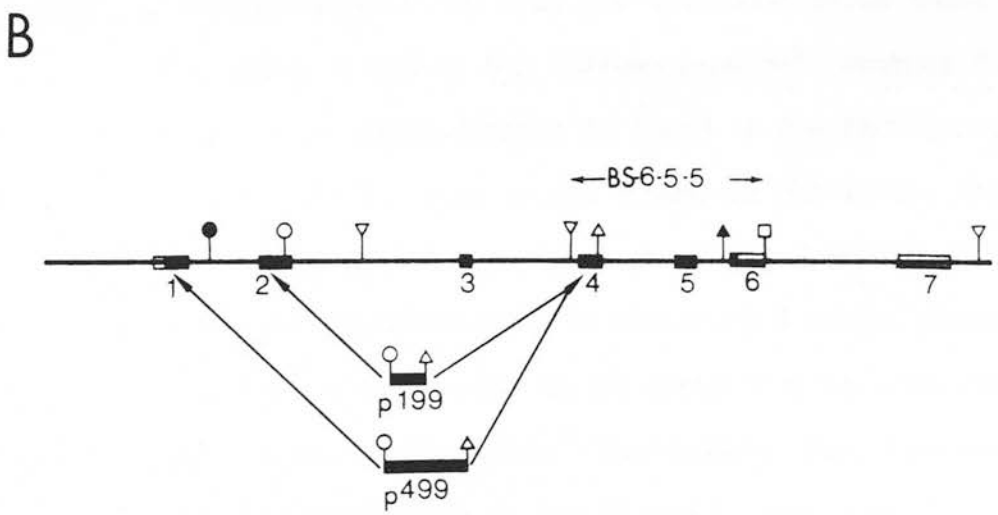
Figure R.5.3.

(A) Northern blot of liver and lachrymal gland poly(A) mRNA probed with the 5' p499 subclone. The hybridized filter was washed down from 0.5 x SET,68°C to 0.2 x SET, 68°C and exposed to X-ray film after each wash. Samples in their respective lanes are given below.

1 and 5 : 1µg liver poly(A) mRNA
2 and 6 : 0.5µg " " "
3 and 7 : 5µg lachrymal gland poly(A) mRNA
4 and 8 : 3µg " " " "

(B) Diagram showing hybridization limits of BS-6-5-5, the 5' p199 subclone and the 5' p499 subclone to the MUP transcription unit.


A
1
2
3
4
5
6
7
8


B
BS-6·5·5
1
2
3
4
5
6
7
p199
p499

The pl99 5' subclone hybridization results suggest that a large proportion of lachrymal mRNA is contributed by sequences that share some homology with this subclone and that have a size in between that of the long and short liver mRNA. The higher stringency wash (Ø.2 x SET, 68°C) suggested that most of the homologous sequences are not transcripts from the same gene origin as pl99. The results of the hybridization with the p499 subclone suggest that the predominant lachrymal gland mRNA is due to group 1 genes that have diverged in exons 1, 2 and 3 from the majority of the cloned group 1 genes. This is because sequence data from the coding region of six cloned group 1 MUP genes indicates that these genes are >= 99% homologous. However, due to the low levels of signal obtained, the hybridization results with the pl99 and p499 5' subclones need further confirmation. These results are compatable with those of Shahan and Derman (1984) as described in detail in the Discussion section. [ The 5' pl99 and 5' p499 subclones were gifts from the Held laboratory.]

In summary, the predominant MUP mRNA was not found to be homologous to the group 2 probe in any of the tissues examined. Whether MUP mRNA homologous to the group 2 probe is found in the lachrymal, submaxillary and mammary glands is not known. In the liver, the short mRNA hybridizes preferentially to the group 2 probe, while the long mRNA hybridizes preferentially to the group 1 probe. Whether the liver mRNA that is homologous to the group 2 probe represents transcription from any of the group 2 pseudogenes characterized or from a gene(s) more homologous at its 3' end to the group 2 probe is

not known. Significant levels of MUP mRNA homologous to the group 1 probe were detected in all tissues at a washing stringency of 0.5 x SET, 68°C.

Finally it should be mentioned that the levels of MUP mRNA detected in the lachrymal, submaxillary and mammary glands were lower than the maximum levels reported (Shaw et al, 1983) judging from the strength of the signal relative to the liver controls. This could be the result of any one or a combination of the following:

1/ Differences in the hybridization stringencies. The most stringent wash used by the Held laboratory in these experiments was 1 x SSC, 65°C. The Tm is ∿80°C at the Na+ concentration of 1 x SSC. The stringent washes used in the experiments described here were 0.2 x SET, 10.02 M tetra-sodium pyrophosphate, 68°C; and 0.5 x SET, 10.02 M tetra-sodium pyrophosphate, 68°C. The Tm values under these Na+ cation concentrations are 70°C and 74°C respectively. [ Tm values were calculated by assuming 40% GC in the sequence and by assuming that DNA-RNA duplexes have a Tm 5°C below that of an equivalent DNA-DNA duplex in aqueous solution.]

2/ Differences in the ages of the animals.

3/ Differences in strains. BALB/c mice were used as opposed to C57BL/Fa mice.

Section 6: Hybridization of the liver cDNA clone pl99 to isolated MUP genes.

Kuhn et al (1984) isolated a cDNA clone, pl99, which was found to be only 85% homologous in its nucleotide sequence to group 1 cDNA clones. It was therefore of interest to determine whether any of the cloned MUP genes were closely related to pl99.

Examination of the nucleotide sequence of pl99 revealed the presence of a SstI site located 3 nucleotides 3' to the PvuII site in exon 4. This suggested that the four group 3 clones BL-8/CL-4, BL-2 and CL-12, were the most likely candidates to share close homology with pl99 since all have a SstI site that maps to the homologous PvuII site. Because no sequence data on group 2 genes was available at the time, it was also thought probable that pl99 could represent a transcript from a group 2 gene.

To investigate this, cloned MUP genes were hybridized to the available 5' pl99 subclone and the hybridization signals of a low stringency wash (1 x SET, 68°C) and a high stringency wash (Ø.2 x SET, 68°C) compared. The autoradiographs showed that none of the isolated clones form stable hybrids with pl99 at high stringency and demonstrated that the group 2 clone BL-25/CL-2, and the group 3 clone CL-12 (and therefore most probably BL-2) were not closely related to pl99. However, due to the small amount of overlap of BL-8/CL-4 and BL-15 with the 5' pl99 subclone it was not possible to draw a conclusion for these genomic clones.

Fortunately MUP15, a cDNA clone containing the entire translated sequences, and which is identical to p199 where the two sequences overlap, was recently isolated from a female BALB/c liver library (A. Chave-Cox, unpublished), Fig.R.6.1(A). The hybridization experiment was therefore repeated with MUP15, and for comparison, a duplicate filter was hybridized with the almost fully cloned group 1 cDNA, MUP11. MUP11, which was also isolated from a BALB/c female liver cDNA library (A.Chave-Cox, unpublished), is a transcript of a group 1 gene, as shown through hybridization studies and confirmed by sequencing. The relative hybridization signals after low stringency and high stringency washes were compared (Fig.R.6.2). Both MUP15 and MUP11 were cloned into M13mp9: this vector contains a 200 bp insertion of pBR322 sequences and so hybridizes to the pCM2 markers. The autoradiographs demonstrated that only BL-8 maintained a stable hybrid with MUP15 under the high stringency wash. However, this clone also maintained a stable hybrid with MUP11 under the high stringency wash. Moreover, with both probes, the same restriction fragments were labelled and in each case stable hybrids were maintained with the 9 kbp HindIII fragment.

The results suggested that BL-8 may represent a clone containing linked MUP genes. For the purposes of further investigation, BL-8 was mapped with the subclone U5. U5 consists of most of the first intron and most of the second exon of the group 1 genomic clone BS-6 (Fig.R.6.1). The results of this latter study suggest that truncated MUP gene sequences are present within the 4.4 kbp SstI fragment that lies 3' to the hybridization limits of pBL-1-4. The exact organisation of these sequences is unclear. They do not include

Figure R.6.1

(A) Extent of hybridization of the cDNA clones 5' p199, 5' p499, MUP11 and MUP15, and a genomic subclone, U5, to BS-6.

(B) Known limits of hybridization of some cDNA clones and genomic subclones to BL-8. For known limits of hybridization of LVA325, LVA132, BL-14, BS-6-2 and BS-6-3 to other MUP clones, see Figure R.2.1.

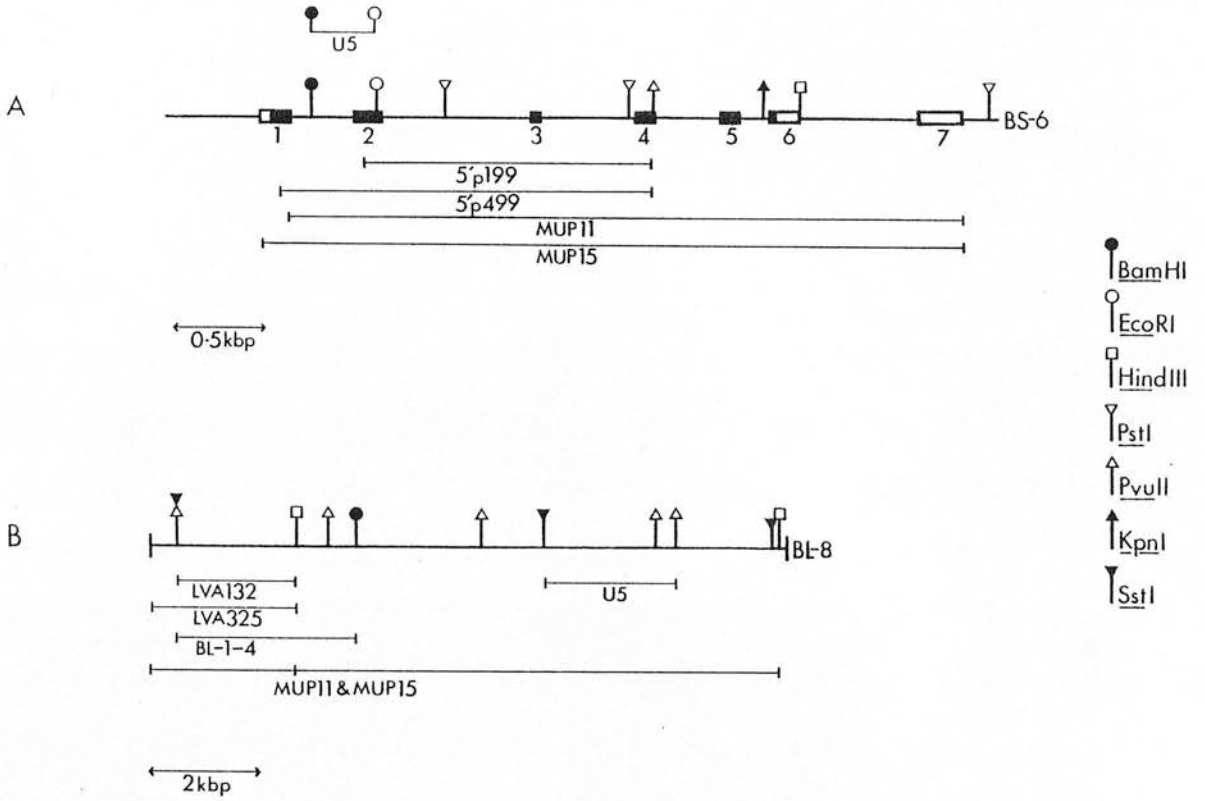
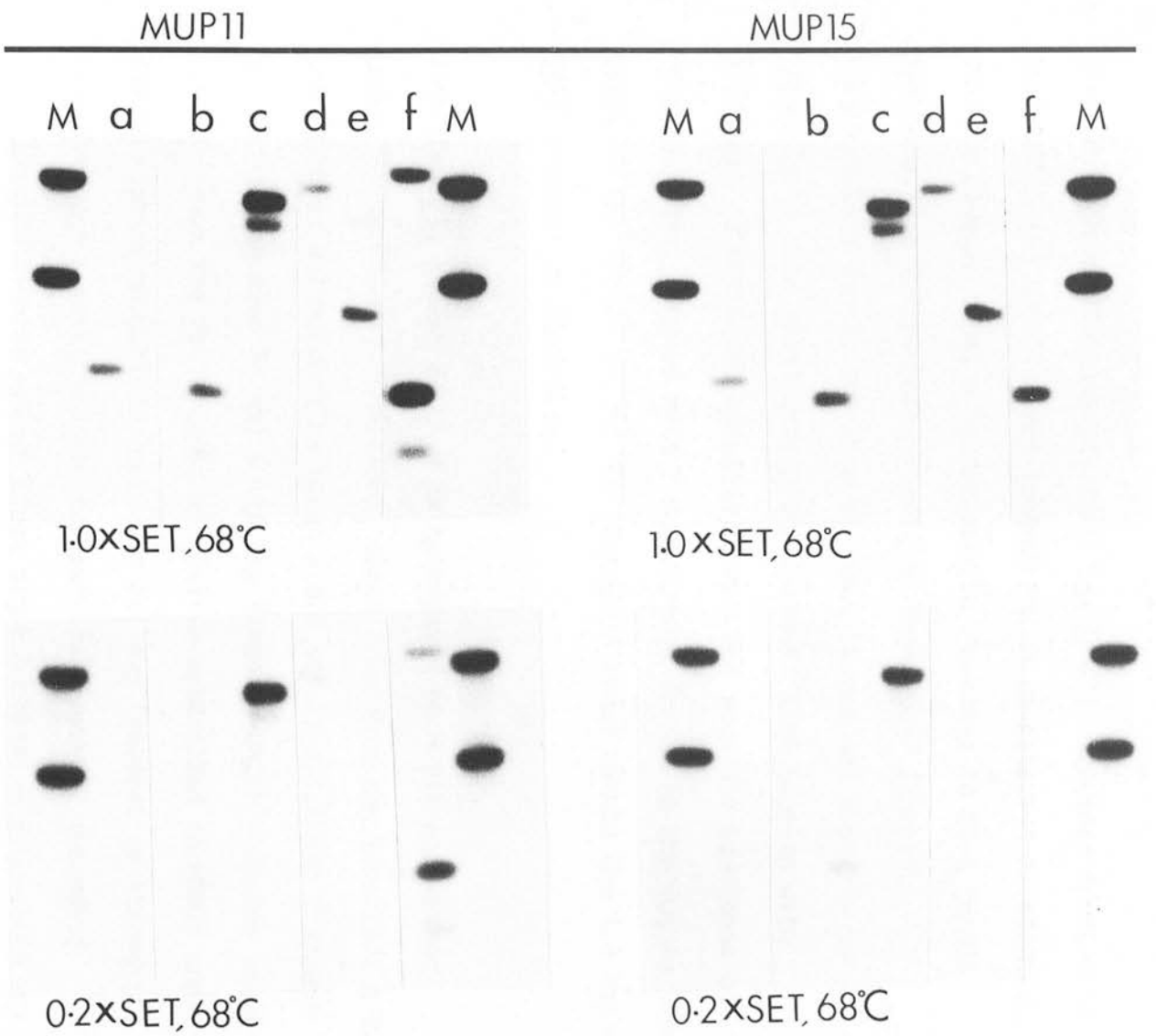

A
U5
1
2
3
4
5
6
7
BS-6
5'p199
5'p499
MUP11
MUP15
0·5kbp
BamHI
EcoRI
HindIII
PstI
PvuII
KpnI
SstI
B
BL-8
LVA132
LVA325
BL-1-4
MUP11 & MUP15
U5
2kbp

Figure R.6.2. Southern blots showing the hybridization reactions of some of the MUP genomic clones with the cDNAs MUP11 and MUP15. 0.5µg DNA samples were restricted, electrophoresed on 0.4% agarose gels and transferred to nitrocellulose. The filters were hybridized with either MUP11 or MUP15 and washed down from 1 x SET (68°C) to 0.2 x SET (68°C). The filters were exposed to X-ray film after each wash. (M) pCM2 digested with EcoRI + BamHI, (a) Bs-105 digested with EcoRI (b) CL-12 digested with PstI, (c) BL-8 digested with HindIII, (d) BL-15 digested with PstI, (e) BL-25 digested with PstI, (f) CL-3 digested with PstI.


MUP11
MUP15
M a b c d e f M
M a b c d e f M
1·0×SET, 68°C
1·0×SET, 68°C
0·2×SET, 68°C
0·2×SET, 68°C

exons 4,5,6 and 7 since the 4.4 kbp SstI fragment does not hybridize with subclones containing these exons only. Fig.R.6.2(B) summarizes the results of the hybridization studies on BL-8.

The strong hybridization of the 9 kbp HindIII fragment of BL-8 to both MUP15 and MUP11 is not thought to be an artifact of cloning. This is because CL-4, a bacteriophage identical to BL-8 in all respects, was independently isolated from a C57 library. Unfortunately due to the lack of time, it was not possible to investigate this intriguing clone further. Fine mapping with subclones of MUP15 and specific exonic and intronic subclones of group 1 genes, coupled with E.M. mapping, should in the future shed light on the nature of the MUP sequences found within the 4.4 kbp SstI fragment of BL-8/CL-4.

When the PstI digest of CL-3 is hybridized to MUP11 and washed under low stringency conditions, three fragments are labelled: a 2.1 kbp fragment which contains exons 4,5,6 and 7; a 1.1 kbp fragment which contains exon 3; and a ∿12 kbp fragment which contains exons 1 and 2. When the PstI digest of CL-3 is hybridized to MUP15 under low stringency conditions, only the 2.1 kbp fragment is strongly labelled, and a faint ∿12 kbp fragment is observed. The MUP15 hybridization results are at first sight surprising since both MUP11 and MUP15 span most of the transcription unit. However, sequence data has revealed that the nucleotide divergence between group 1 and MUP15 is 20% for exons 1 - 3 and 11% for exons 4 - 7. The nucleotide divergence for the 3rd exon, which is contained alone in the 1.1 kbp PstI fragment, is 28%. The hybridization results are therefore

Table R.6.1. Nucleotide divergence between the group 1 consensus sequence and MUP15

| Exon(s) | Size(bp) | Nucleotide divergence(%) |
|---|---|---|
| 1 | 125 | 8.8 |
| 2 | 134 | 24.6 |
| 3 | 74 | 28.3 |
| 4 | 111 | 9.9 |
| 5 | 102 | 6.9 |
| 6 | 44 | 2.3 |
| 7 | 253 | 13.0 |
| 1-3 | 333 | 14.1 |
| 4-7 | 510 | 19.5 |
| 1-7 | 843 | 10.6 |

Sequence data were obtained from A.Chave-Cox and J.Clark

thought to be a consequence of the differences in homology of the 5' and 3' exons to the probes. The 3' exons of MUP15 and BS-2 also share greater homology than do their 5' exons (see Table R.6.1). It is hypothesised that the 3' exons of MUP15 may have been involved in a conversion event with a gene that was ancestral to the group 1 and group 2 genes or with a gene that is equally diverged from group 1 and group 2 genes (A.Chave-Cox, unpublished).

In summary, p199 and MUP15 do not represent transcripts derived from any of the MUP genomic clones isolated, although some truncated MUP sequences that hybridize preferentially to these cDNA clones are present in BL-8/CL-4.

Section 7: Variation in the MUP structural genes of BALB/c and C57BL/Fa mice.

Bennett et al (1982) demonstrated, through probing restriction digests of genomic DNA with MUP cDNA, that despite the overall similarity, some variation in the structural MUP genes is present between inbred mouse strains. However, under the washing conditions used, it was not possible to distinguish whether most of the variation was present in pseudogenes, or whether variation was also found in the abundantly transcribed group 1 genes. Triplicate PstI and EcoRI digests probed with the group 1 probe BS-6-5-5, the group 2 probe BS-2-2-2 and a homologous fragment isolated from BL-2 (Bishop, unpublished), indicated that variation between BALB/c and C57BL/Fa was not restricted to a single group of MUP genes and that variant group 1 genes, variant group 2 genes and variant genes which hybridized preferentially to the BL-2 subclone were all present. It therefore seemed possible to isolate an expressed variant MUP gene by comparing the restriction maps of a large number of characterized genes with the restriction patterns of the genomic DNAs of different inbred mouse strains.

Contamination of the C57 strain. After the bulk of the restriction mapping had been completed, it was found that the C57 strain was contaminated. The C57 mice were alleged to be BC15-albumin congenics and had been originally derived by back-crossing the Petras albumin variant (Petras and MacLaren, 1976) to C57BL/6J mice, for eight generations (T. Roderick, personal communication). There were therefore no reasons to believe that the MUP genotype of the BC15-

albumin variants would be different from that of C57BL/6J mice. The contamination was brought to my attention on discovering that in 1980, the strain had been segregating for coat colour in that white and agouti mice were produced. Isolation of C57 MUP clones from the C57 libraries was initiated in 1981. By this time no variation in coat colour was observed due to the fact that black mice had been selected for breeding purposes. IEF resolution of urinary proteins from 14 mice by S.Hainey revealed that there were differences in both the presence and the intensities of bands between individuals. Also, most of the mice had an extra band, not present in C57BL/Fa, that migrated close to pH 4.6 and none showed a pattern similar to that of C57BL/Fa (C57BL/Fa is the C57BL mouse strain kept at the Edinburgh Animal Breeding House; the male urinary protein IEF pattern of this strain differs from that of C57BL/6J only in the intensity of the most basic band).

To shed light on the nature of the contamination, a comparison of four known genetic markers, by isoelectric focusing of blood samples from 16-18 C57 (BC15) individuals, was kindly carried out by Graham Bullfield. The markers used were Gpi-1 (glucose-6-phosphate isomerase-1, chromosome 7), Alb-1 (serum albumin variant, chromosome 5), Hbb (haemoglobin beta-chain, chromosome 7) and Hba (haemoglobin alpha chain, chromosome 11). The results of this analysis coupled with segregation for c (albino, chromosome 7) and A (agouti, chromosome 2) observed in 1980 suggest that the contamination was by BALB/c.

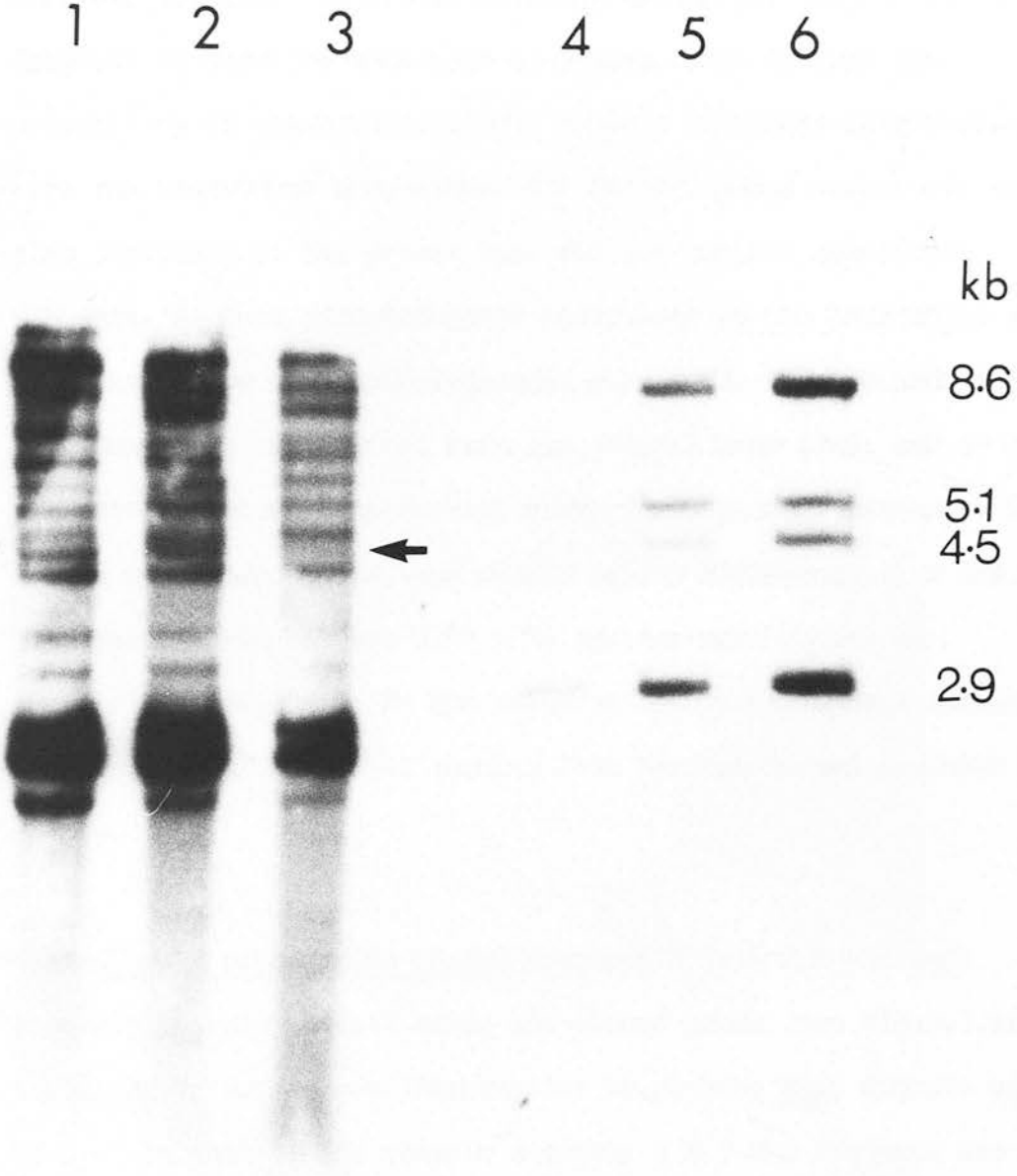
To determine whether the MUP structural locus had been contaminated,

a sample of C57 genomic DNA from the same pool which had been originally used to prepare the C57 library was digested with BamHI and HindIII. Similar digests were prepared from BALB/c and C57BL/Fa DNA. The three samples were electrophoresed in parallel, transferred to nitrocellulose and probed with BS-6-2. The autoradiograph showed that the restriction patterns of the C57 (BC15) and C57BL/Fa samples were identical, except for the presence of a minor ∿4.3 kbp fragment in the latter (Fig.R.7.1). However a minor ∿4.3 kbp fragment was also present in the BALB/c sample. Therefore the absence of the 4.3 kbp fragment in the C57(BC15) sample either represents a difference between C57BL/Fa and C57BL/6J mice or contamination of the C57(BC15) stock by a strain other than BALB/c. In view of the fact that C57BL/Fa and C57BL/6J were separated 50 years ago, the former possibility is quite likely.

The search for variant MUP genes was conducted by comparing the restriction patterns of the cloned BALB/c and C57 MUP genes to the restriction patterns of BALB/c and C57BL/Fa genomic DNAs. C57BL/Fa DNA was used as opposed to C57(BC15) genomic DNA due to the possible contamination of the C57(BC15) MUP structural locus. Restriction digests and probes were chosen that would demonstrate whether a particular fragment unique to one of the cloned MUP genes was represented in the genomic DNAs of both inbred strains. It was primarily of interest to identify fully cloned variant genes, and for this reason genomic blots were designed to identify variant fragments within the regions which hybridized to the subclones BS-6-2, BS-6-5 or BL-1-4.

Figure R.7.1. Southern blot of HindIII + BamHI genomic digests probed with BS-6-2. The digested samples were electrophoresed on a 1% agarose gel prior to transfer. Samples in their respective lanes are given below.

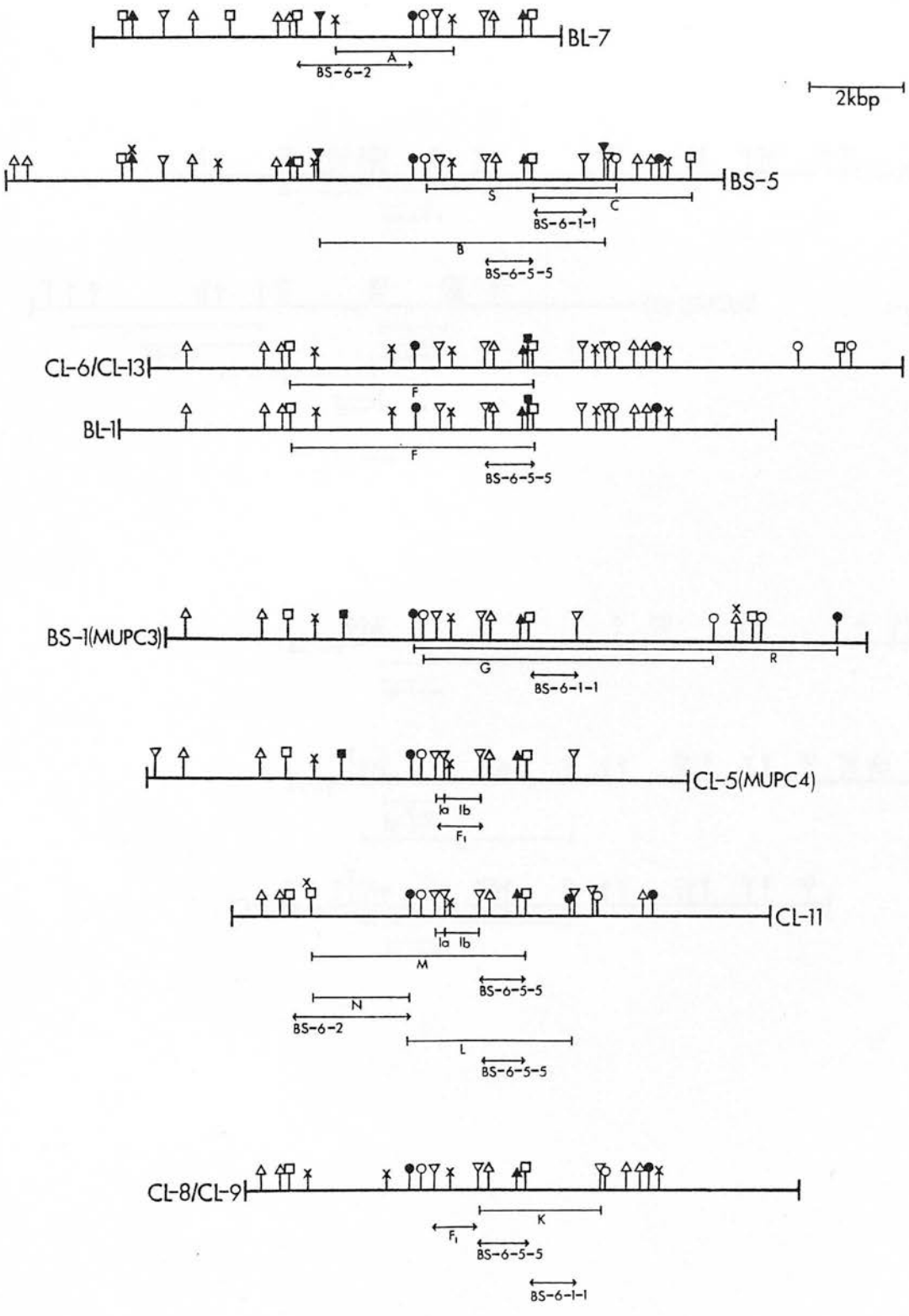
1. 20µg BALB/c genomic DNA digested with HindIII + BamHI.
2. 20µg C57BL/Fa " " " " " " "
3. 20µg C57(BC15) " " " " " " "
4. 30pg of each of the digests CL-1/Kpn 1, CL-1/HindIII, CL-1/PvuII, CL-1/PvuII + EcoRI.
5. 150pg of each of the digests loaded in lane 4.
6. 300pg " " " " " " " " "


1
2
3
4
5
6
kb
8·6
5·1
4·5
2·9

Wherever possible, subclones that lay within the boundaries of the fragment in question were used as probes. This reduced the probability of non-homologous MUP genomic fragments co-migrating with the restricted recombinant MUP bacteriophage marker. It was also important to use probes that did not contain repetitive elements, as such elements would contribute to the background signal and obscure low copy number genomic fragments. The subclones used as probes were all derived from the group 1 gene BS-6, and to allow the identification of potential group 2 and group 3 variants, most of the hybridized blots were washed at low stringency (1 x SET, 68 °C). Agarose gels between Ø.7% - 2% agarose were chosen as appropriate, depending on the sizes of the bacteriophage markers. Brief descriptions of the results from the hybridized Southern blots follow.

MspI digests probed with BS-6-2 Fragment A is a 2.5 kbp MspI fragment, unique to BL-7 among the cloned genes (see Fig.R.7.2). Its strain distribution was investigated by probing MspI digests with BS-6-2. In each of the genomic digests, a 2.5 kbp fragment was found to co-migrate with fragment A (Fig.R.7.3). A densitometer scan of the autoradiograph showed that the 2.5 kbp BALB/c genomic fragment had been labelled 3.5 - 4 times more intensely than the 2.5 kbp C57BL/Fa fragment. Differences in copy number were not found to be restricted to this fragment. Some of the differences in the MspI restriction patterns could result from the methylation of MspI sites at positions which inhibit their cleavage by the enzyme. However, MspI is able to cleave the sequence (CCpGG). Because methylation at CpG represents more than 90% of methylated residues

Figure R.7.2. Unique restriction fragments of isolated MUP clones and the probes used to investigate their strain distribution. For symbols, see Figure R.2.2.


BL-7
A
BS-6-2
2kbp
BS-5
S
C
BS-6-1-1
B
BS-6-5-5
CL-6/CL-13
F
BL-1
F
BS-6-5-5
BS-1(MUPC3)
G
R
BS-6-1-1
CL-5(MUPC4)
Ia Ib
F1
CL-11
Ia Ib
M
BS-6-5-5
N
BS-6-2
L
BS-6-5-5
CL-8/CL-9
K
F1
BS-6-5-5
BS-6-1-1

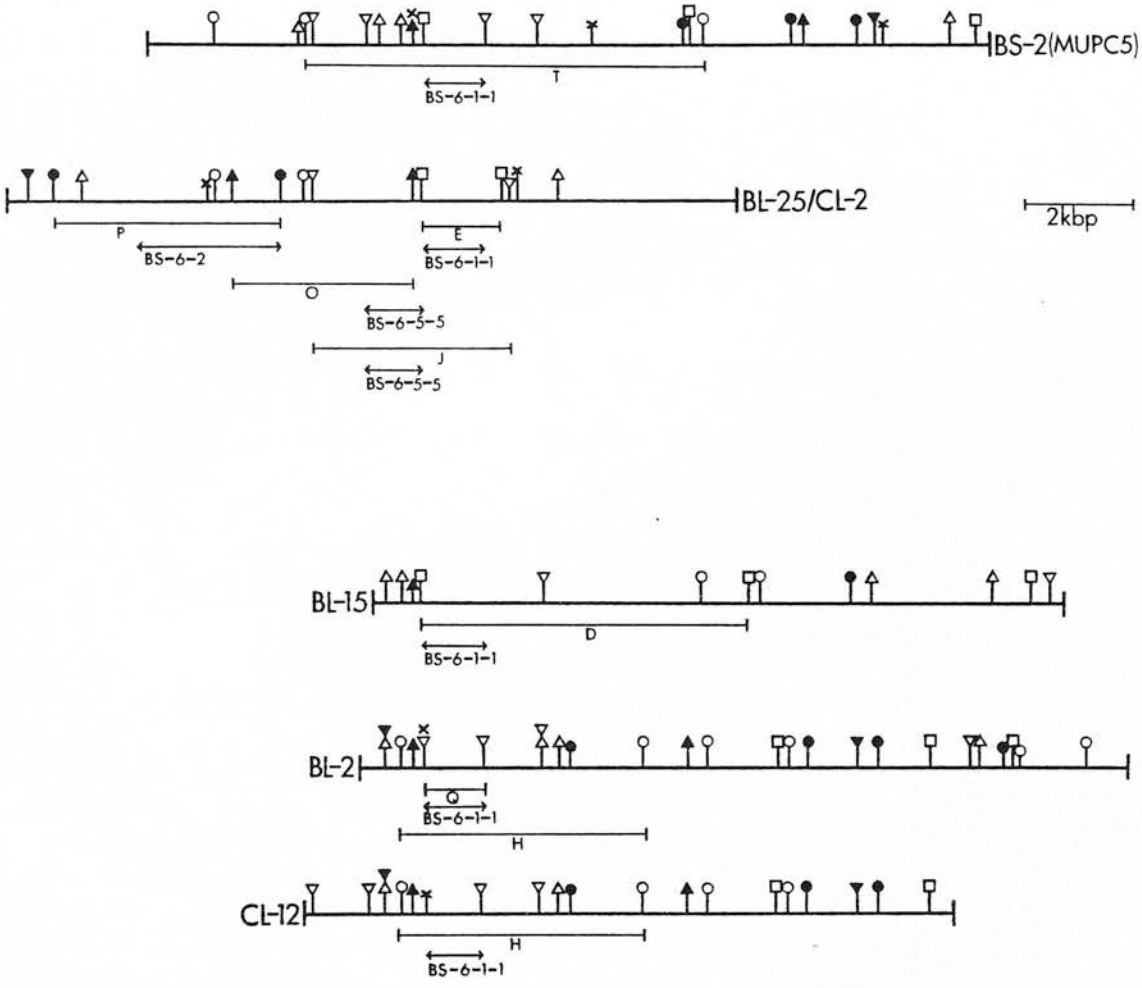
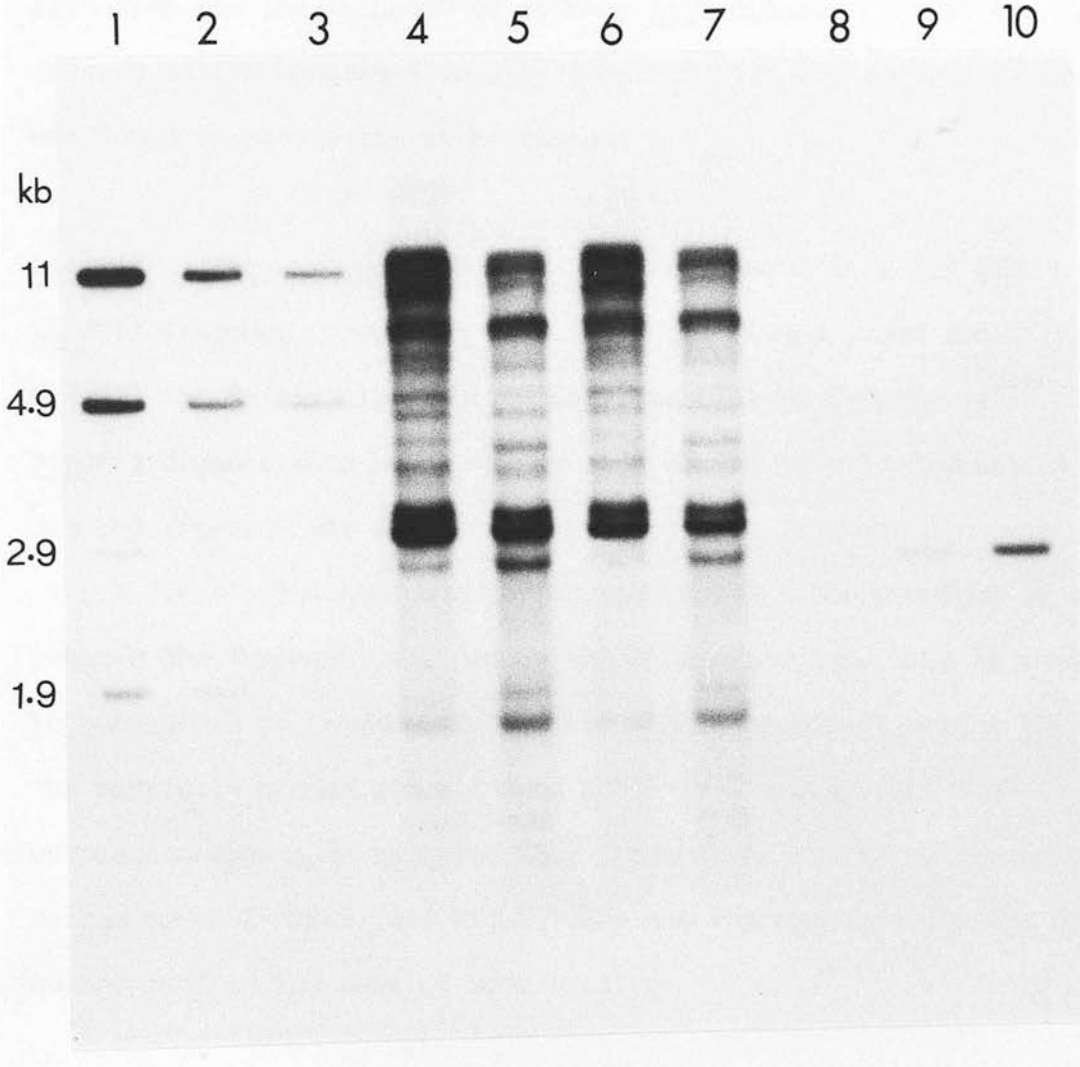

BS-2(MUPC5)
T
BS-6-1-1
BL-25/CL-2
2kbp
P
BS-6-2
E
BS-6-1-1
O
BS-6-5-5
J
BS-6-5-5
BL-15
D
BS-6-1-1
BL-2
O
BS-6-1-1
H
CL-12
H
BS-6-1-1

Figure R.7.3. Southern blot of MspI genomic digests probed with BS-6-2. The digested samples were electrophoresed on a 1% agarose gel prior to transfer. Samples in their respective lanes are given below.

1. 150pg pCM2 marker
2. 60pg " "
3. 30pg " "
4. 17µg C57BL/Fa genomic DNA digested with MspI.
5. 17µg BALB/c " " " " "
6. 15µg C57BL/Fa " " " " "
7. 15µg BALB/c " " " " "
8. 30pg of BL-7 cloned DNA digested with MspI.
9. 150pg " " " " " " "
10. 600pg " " " " " " "


1
2
3
4
5
6
7
8
9
10
kb
11
4.9
2.9
1.9

in eukaryotic genomes (Razin and Riggs 1980; Ehrlich and Wang, 1981) the differences are more likely to be due to sequence polymorphisms.

SstI digests probed with BS-6-5-5. Fragment B is a 6.0 kbp SstI fragment unique to BS-5 (see Fig.R.7.2). Its strain distribution was investigated by probing SstI digests with BS-6-5-5. In each of the samples a heavily labelled ~6.0 kbp genomic fragment was found to co-migrate with fragment B.

HindIII digests probed with BS-6-1-1. Fragment C is a 3.3 kbp HindIII fragment, unique to BS-5 among the cloned genes (see Fig. R.7.2). Its strain distribution was investigated by probing HindIII digests with BS-6-1-1. In each of the genomic digests, a 3.3 kbp fragment was found to co-migrate with fragment C (Fig.R.7.4.A). The same autoradiograph revealed the presence of a minor 6 kbp fragment confined to the BALB/c samples. This is thought to correspond to fragment D, a 6 kbp HindIII fragment unique to the partially cloned group 2 gene BL-15 (see Fig.R.7.2). The autoradiographs also revealed that fragment E, a 1.5 kbp fragment unique to BL-25/CL-2 (see Fig.R.7.2), was represented by a low copy number band in the DNAs of both strains.

EcoRI + HindIII digests probed with BS-6-5-5. To determine whether the absence of an EcoRI site in exon 2 was common to many MUP genes, or unique to BL-1, it was necessary to digest the genomic DNAs with EcoRI and one other enzyme. The digest and probe chosen were EcoRI+HindIII and BS-6-5-5 respectively. Earlier digestion samples not normally included were run onto the gel which was

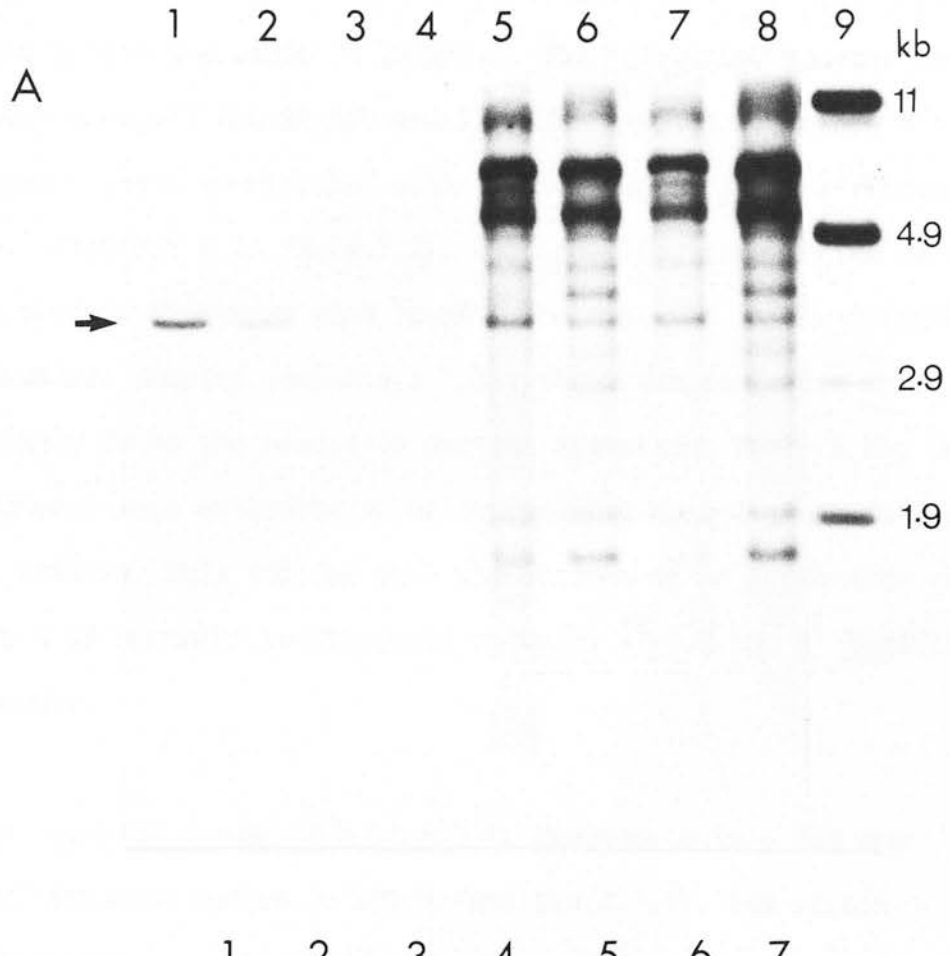
Figure R.7.4.

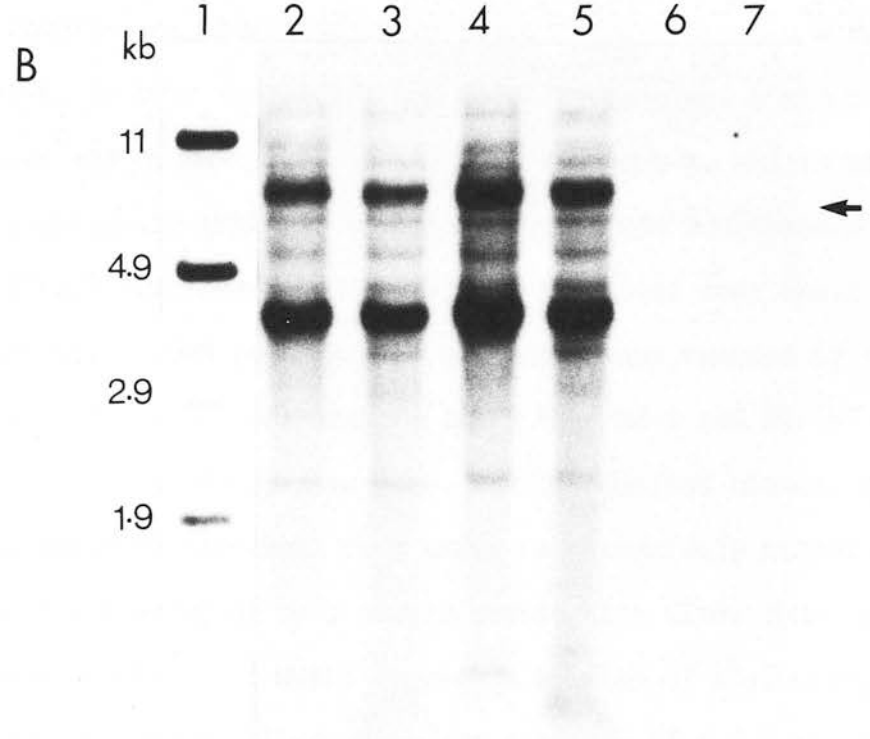
(A) Southern blot of HindIII genomic digests probed with BS-6-1-1. The digested samples were electrophoresed on a 1% agarose gel prior to transfer. Lanes 1-8, ten day exposure of the autoradiograph; lane 9, one day exposure of the autoradiograph. Samples in their respective lanes are given below.

1. 500 pg of BS-5 cloned DNA digested with HindIII.
2. 300 pg " " " " " " " "
3. 60 pg " " " " " " " "
4. 30 pg " " " " " " " "
5. 18 µg of BALB/c genomic DNA digested with HindIII.
6. 12 µg of C57BL/Fa " " " " " "
7. 12 µg of BALB/c " " " " " "
8. 18 µg of C57BL/Fa " " " " " "
9. 0.5 µg of pCM2 marker.

(B) Southern blot of EcoRI digests probed with BS-6-1-1. The digested samples were electrophoresed on a 0.8% agarose gel prior to transfer. Lane 1, one day exposure of the autoradiograph; lanes 2-7, ten day exposure of the autoradiograph. Samples in their respective lanes are given below.

1. 0.5 µg of pCM2 marker
2. 12 µg of C57BL/Fa genomic DNA digested with EcoRI
3. 12 µg of BALB/c " " " " " "
4. 18 µg of C57BL/Fa " " " " " "
5. 18 µg of BALB/c " " " " " "
6. 30 pg of BS-1 digested with EcoRI
7. 60 pg " " " " "


A
1 2 3 4 5 6 7 8 9
kb
11
4·9
2·9
1·9

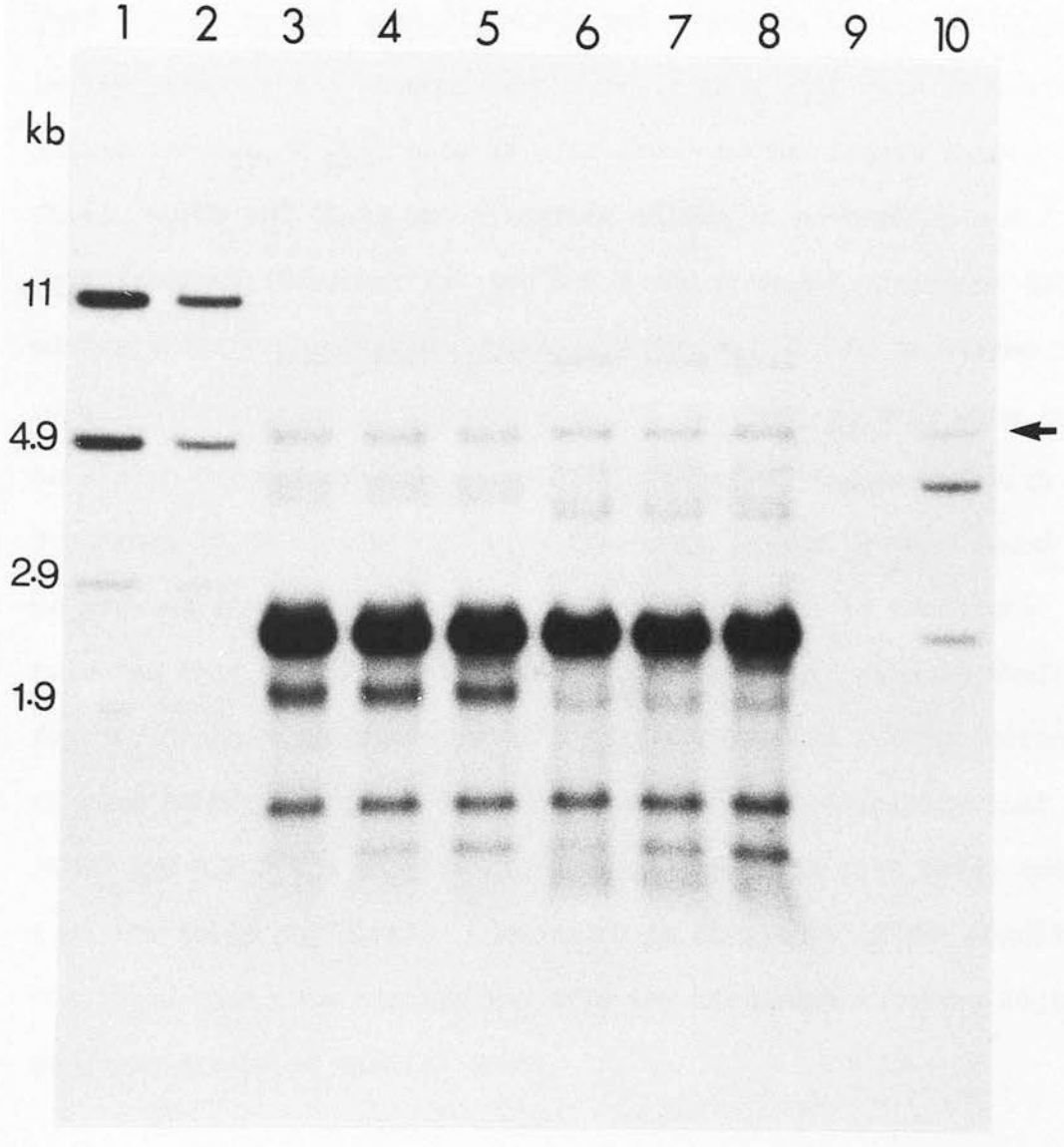

B
kb
1 2 3 4 5 6 7
11
4·9
2·9
1·9

subsequently blotted. This was done in order to determine whether partial digestion was likely to contribute to the hybridization signals from the bands of interest. The hybridized genomic blot showed that all the BALB/c and C57BL/Fa samples contained a minor fragment which co-migrated with the 5.3 kbp 5' HindIII fragment of BL-1 (fragment F in Fig.R.7.2). The labelling intensities of the 5.3 kbp genomic fragments were found to be the same in the different incubation samples (see Fig.R.7.5). These fragments are therefore unlikely to be the result of partial digestion. The 5.3 kbp genomic fragments were estimated to be represented by 1-2 genes in each of the strains. This implies that the absence of an EcoRI site within exon 2 is uncommon in MUP genes carrying a ∿5.3 kbp 5' HindIII fragment.

EcoRI digests probed with BS-6-1-1. Fragment G is a 6.3 kbp EcoRI fragment unique to MUPC3 (see Fig.R.7.2). Its strain distribution was investigated by probing EcoRI digests with BS-6-1-1. In both the BALB/C and C57BL/Fa samples, a minor 6.3 kbp fragment was found to co-migrate with fragment G, and in each case it was estimated that the genomic fragment was represented by a single gene (Fig.R.7.4). BS-1 and BS-1Ø7 are identical over their co-extensively cloned regions. The estimated copy numbers of the 6.5 kbp genomic fragments therefore imply that BS-1 and BS-1Ø7 are probably clones of the same gene. The C57-derived clones, CL-5 and CL-1Ø, are also identical over their co-extensively cloned regions and differ from MUPC3 by a single restriction site. Thus it is likely that MUPC3 and MUPC4 represent alleles of a single gene. The same autoradiograph illustrates the presence of a 4.5 kbp fragment

Figure R.7.5. Southern blot of HindIII + EcoRI genomic digests probed with BS-6-5-5. The digested samples were electrophoresed on a 0.8% agarose gel prior to transfer. Samples in their respective lanes are given below.

1. 60 pg of pCM2 marker
2. 30 pg " " "
3. 15 µg of C57BL/Fa genomic DNA digested with HindIII + EcoRI for 60 min.
4. 15 µg of C57BL/Fa genomic DNA digested with HindIII + EcoRI for 120 min.
5. 15 µg of C57BL/Fa genomic DNA digested with HindIII + EcoRI for 180 min.
6. 15 µg of BALB/c genomic DNA digested with HindIII + EcoRI for 60 min.
7. 15 µg of BALB/c genomic DNA digested with HindIII + EcoRI for 120 min.
8. 15 µg of BALB/c genomic DNA digested with HindIII + EcoRI for 180 min.
9. 30 pg of each of the digests BS-5/EcoRI + HindIII, BS-5/EcoRI, BL-1/HindIII.
10. 150 pg of each of the digests loaded in lane 9.


1
2
3
4
5
6
7
8
9
10
kb
11
4.9
2.9
1.9

in the DNA of each strain. This fragment has been shown (by a different blot) to co-migrate with fragment H, a 4.5 kbp EcoRI fragment unique to BL-2 and CL-12.

PstI digests probed with F1. MUPC3 and MUPC4 are distinguishable by the presence and absence respectively of a PstI site in their fourth introns. A PstI site is also found in the fourth intron of CL-11. MUPC4 and CL-11 are therefore unique in possessing a Ø.3 kbp PstI fragment (fragment Ia) and a Ø.7 kbp fragment (fragment Ib) within their transcription units (see Fig.R.7.2). To determine the strain distribution of fragments Ia and Ib, PstI digests were run on a high percentage agarose gel (2%), blotted, and probed with F1. Fragments which co-migrated with fragments Ia and Ib were found to be present in the digested DNA of each strain. It is therefore presumed that some BALB/c MUP genes contain a PstI site in their fourth intron. The presence of a PstI site in the fourth intron of some BALB/c MUP genes does not discount the possibility that MUPC3 and MUPC4 are allelic clones. In connection with this, note that the Dollo phylogenies (described in Section 4 of the Results) suggested that this restriction site may have been involved in a gene conversion or similar event.

PstI digests probed with BS-6-5-5. Fragment K is a 2.7 PstI fragment unique to CL-8/CL-9 (see Fig.R.7.2). To investigate its strain distribution, BALB/c and C57BL/Fa genomic DNAs were digested with PstI, hybridized with the group 1 probe BS-6-5-5 and washed at high stringency (Ø.2 x SET, 68 °C). The autoradiograph demonstrated that fragment K was represented by a minor band only

in the C57BL/Fa samples (Fig.R.7.6). Fig.R.7.7 shows a different blot (prepared by Melville Richardson) of PstI digests probed with BS-6-5-5 and washed at high stringency. This blot serves to illustrate that the observed strain variation is not an artifact caused by over-loading the C57BL/Fa samples. Fragment K appears to be represented by a single gene in the C57BL/Fa genome which implies that CL-8 and CL-9 are probably clones of the same gene. The PstI restriction site unique to CL-8/CL-9 does not lie within the coding region. This is concluded from comparing the restriction map of CL-8/CL-9 with that of BS-6, a fully sequenced group 1 gene. PstI digests probed with BS-6-1-1 also showed a 2.7 kbp fragment only in the C57BL/Fa samples, while PstI digests probed with F1 showed no such fragment in either the BALB/c or C57BL/Fa samples. These results are consistent with the proposition that the C57BL/Fa 2.7 kbp genomic fragment corresponds to the 2.7 kbp fragment of CL-8/CL-9.

Fragment J, is a 3.8 kbp PstI fragment unique to BL-25/CL-2 (see Fig.R.7.2). The PstI digest probed with BS-6-5-5 showed that a fragment similar in size to fragment J was present in the digested DNA of each strain.

BamHI digests probed with BS-6-5-5. Fragment L is a 3.5 kbp BamHI fragment unique to CL-11 (see Fig.R.7.2). To investigate its strain distribution, BamHI samples were probed with BS-6-5-5. In both the BALB/c and C57BL/Fa samples a 3.5 kbp fragment was found to co-migrate with fragment L (Fig.R.7.8). A densitometer scan of the autoradiograph revealed that the C57BL/Fa 3.5 kbp genomic fragment

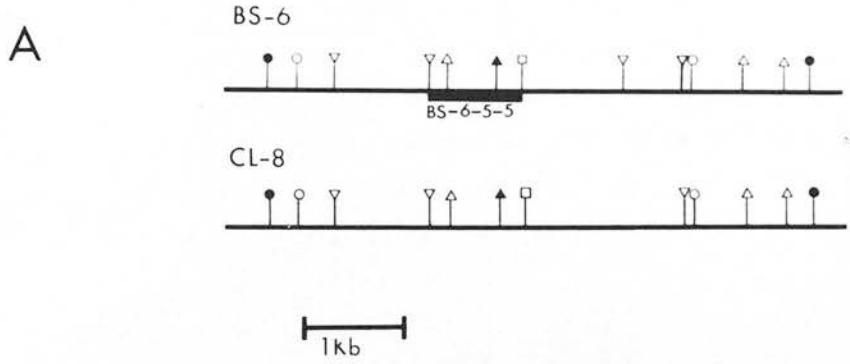
Figure R.7.6.

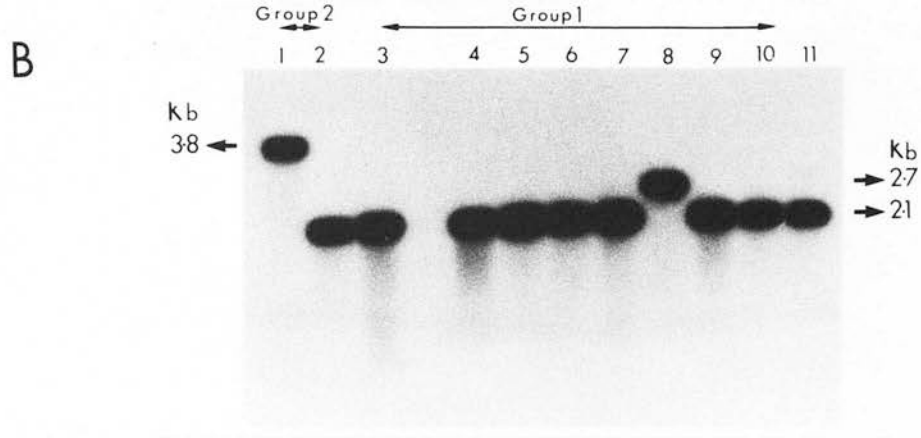
(A) Restriction maps of the transcription units and 3' flanking regions of BS-6 and CL-8. The extent of hybridization of BS-6-5-5 to BS-6 is shown. For symbols, see Figure R.2.2.

(B) Hybridization of some MUP genomic clones to the group 1 probe BS-6-5-5. 0.5µg DNA samples of the Charon 4A clones were digested with PstI and electrophoresed on a 0.4% agarose gel. The Southern transfer was washed under low stringency conditions (1 x SET,68°C) after hybridization. Lanes 1-11, DNA from Charon 4A clones BL-25 (1), BS-2 (2), BS-5 (3), BS-107 (4), CL-3 (5), CL-5 (6), CL-6 (7), CL-10 (9), CL-11 (10) and CL-12 (11).

(C) Southern blot of genomic DNA, digested with PstI and probed with the group 1 probe, BS-6-5-5. Digested samples were electrophoresed on a 0.8% agarose gel prior to transfer. The filter was hybridized to BS-6-5-5 and washed under high stringency conditions (0.2 x SET, 68°C). Samples in their respective lanes are given below.

1. 300 pg of CL-8 cloned DNA digested with PstI.
2. 150 pg " " " " " " "
3. 30 pg " " " " " " "
4. 15 pg " " " " " " "
5. 19 µg of C57BL/Fa genomic DNA digested with PstI.
6. 14 µg " " " " " " "
7. 14 µg of BALB/c genomic DNA digested with PstI.
8. 19 µg " " " " " " "


A
BS-6
BS-6-5-5
CL-8
1kb


B
Group 2
Group 1
1 2 3 4 5 6 7 8 9 10 11
kb
3·8
kb
2·7
2·1


1 2 3 4 5 6 7 8
kb
C
4.4
3·8
3·0
2·7
2.4
2.1

Figure R.7.7. Southern blot of PstI genomic digests probed with BS-6-5-5. Samples were electrophoresed on a 0.8% agarose gel prior to transfer. The filter was washed under high stringency conditions (0.2 x SET, 68°C) after hybridization. Samples in their respective lanes are listed below.

1. 65 pg of pCM2 marker.
2. 32.5 pg " " "
3. 10 µg of JU genomic DNA digested with PstI.
4. 10 µg of BALB/c " " " " " "
5. 10 µg of C57BL/Fa genomic DNA digested with PstI.

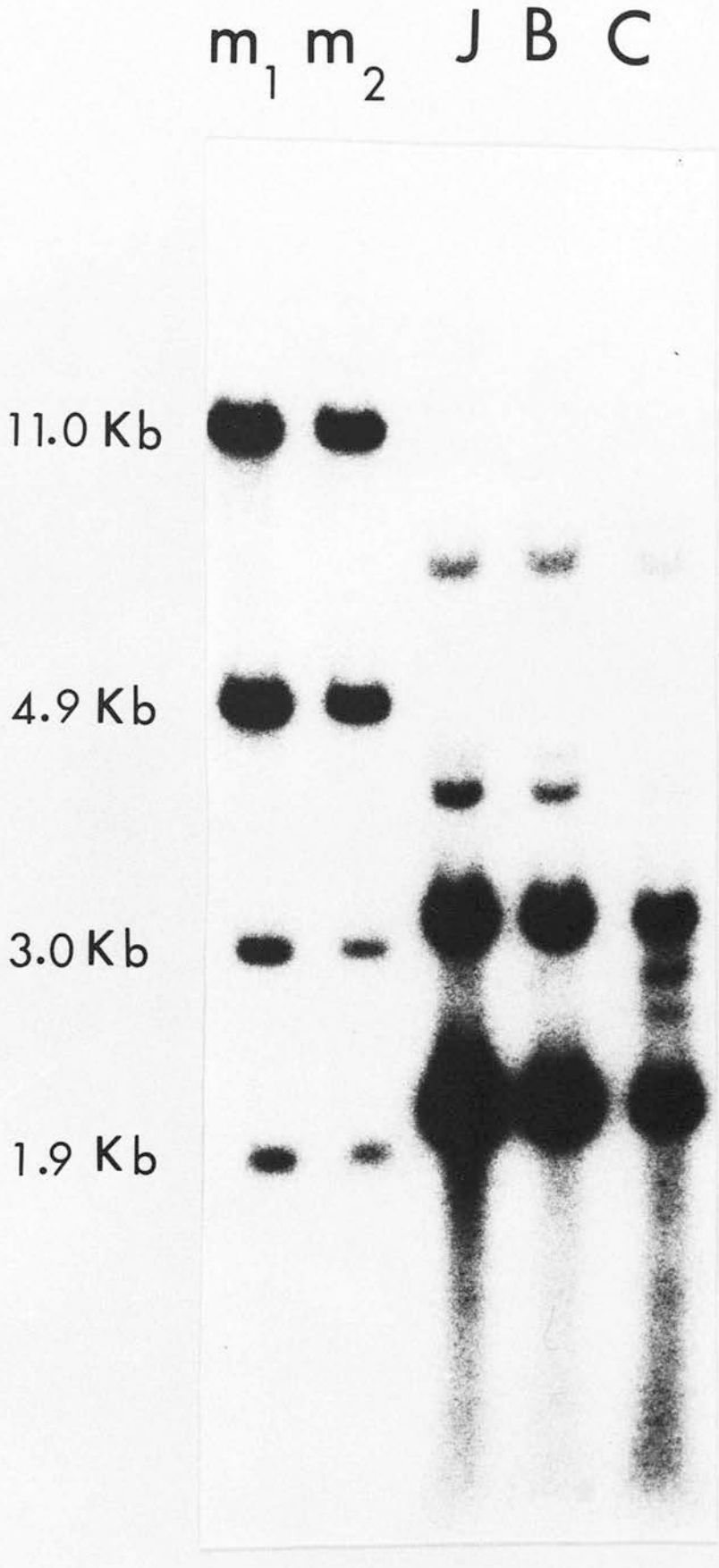
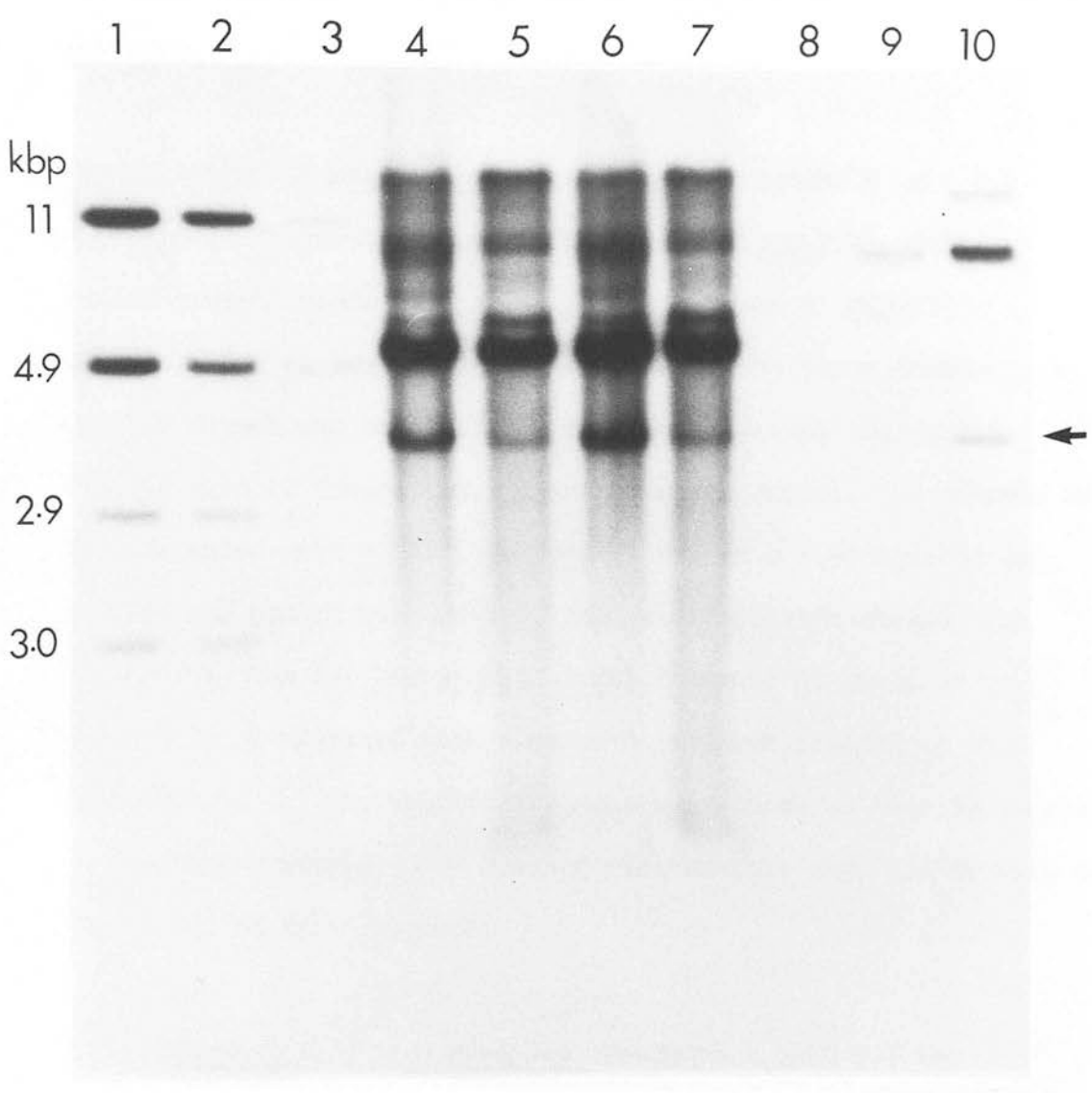

m1 m2 J B C
11.0 Kb
4.9 Kb
3.0 Kb
1.9 Kb

Figure R.7.8. Southern blot of BamHI digests probed with BS-6-5-5. The digested samples were electrophoresed on a 0.8% agarose gel prior to transfer. Samples in their respective lanes are given below.

1. 150 pg of pCM2 marker
2. 30 pg " " "
3. 3 pg " " "
4. 15 µg of C57BL/Fa genomic DNA digested with Bam HI.
5. 15 µg of BALB/c " " " " " "
6. 18 µg of C57BL/Fa " " " " " "
7. 18 µg of BALB/c " " " " " "
8. 15 pg of each of the digests BS-105/Bam HI, CL-11/BamHI, BL-25/BamHI (over loaded).
9. 30 pg of each of the digests loaded in lane 8.
10. 150 pg " " " " " " " " " "


1
2
3
4
5
6
7
8
9
10
kbp
11
4.9
2.9
3.0

had been labelled ∿4 times as heavily as the BALB/c 3.5 kbp genomic fragment.

HindIII+BamHI digests probed with BS-6-2. Fragment M is a 4.8 kbp HindIII fragment unique to CL-11 (see Fig.R.7.2). Other cloned group 1 genes have a 5.1 kbp or 5.3 kbp 5' HindIII fragment. Due to the similarity in size of the three fragments a HindIII digest was not suitable for investigating the strain distribution of fragment M. To maximize separation, the genomic DNAs were digested with HindIII and BamHI, run on a Ø.8% agarose gel, blotted and probed with BS-6-2. The autoradiograph showed that fragment N (the 2.1 kbp HindIII-BamHI fragment of CL-11 in Fig.R.7.2) co-migrated with a genomic fragment present in both strains. The 2.1 kbp genomic fragment was ∿twice as heavily labelled in the BALB/c sample. Once again differences in copy number were not restricted to this fragment.

KpnI digests probed with BS-6-5-5. Fragment O is a 3.4 kbp KpnI fragment unique to BL-25/CL-2 (see Fig.R.7.2). Its strain distribution was investigated by probing KpnI digests with BS-6-5-5. A KpnI fragment similar in size to fragment O was present in the digested DNA of each strain.

BamHI digests probed with BS-6-2. Fragment P is a 4.5 kbp fragment unique to BL-25/CL-2. Its strain distribution was investigated by probing BamHI digests with BS-6-2. In both the BALB/c and C57BL/Fa samples a 4.5 kbp fragment was found to co-migrate with fragment P.

PstI digests probed with BS-6-1-1. PstI digests probed with BS-6-1-1 revealed a weakly hybridizing 1 kbp fragment in the digested DNA of each strain. This fragment is likely to be contributed to by fragment Q, a 1 kbp frament unique to BL-2 (see Fig.R.7.2).

To summarize, a fully cloned variant group 1 MUP gene (CL-8/CL-9) and a partially cloned variant group 2 MUP gene (BL-15) have been identified between the two mouse strains, BALB/c and C57BL/Fa. Some of the other MUP genes isolated may be variants between the two strains but have not been identified as such due to the limitations of the method.

The densitometer scans of the autoradiographs were not suitable for estimating the copy number of individual bands. This was due to the contribution of adjacent bands to the integrated areas of the peaks of interest. However, it was possible to determine whether there are equal numbers of MUP genes in both strains that hybridize to the group 1 probe under low stringency conditions. The EcoRI+HindIII double digest was found to be most suitable for the analysis for the following reasons: (1) the double digest reduced the possibility of obtaining a discrepancy due to poorer transfer efficiency of large genomic DNA, (2) three 15µg samples could be compared and (3) no differences in the sample loadings were observed in the ethidium bromide stain of the gel. A comparison of the areas under the peaks (Fig.R.7.9) was obtained by weighing cut-out shapes of the total scans (Table R.7.1). The mean ratio of the areas of the BALB/c scans

Figure R.7.9. Sketches of the densitometer scans of the HindIII + EcoRI genomic digests probed with BS-6-5-5 (autoradiograph illustrated in Figure R.7.5). 3 - 8 : number of lane scaned.

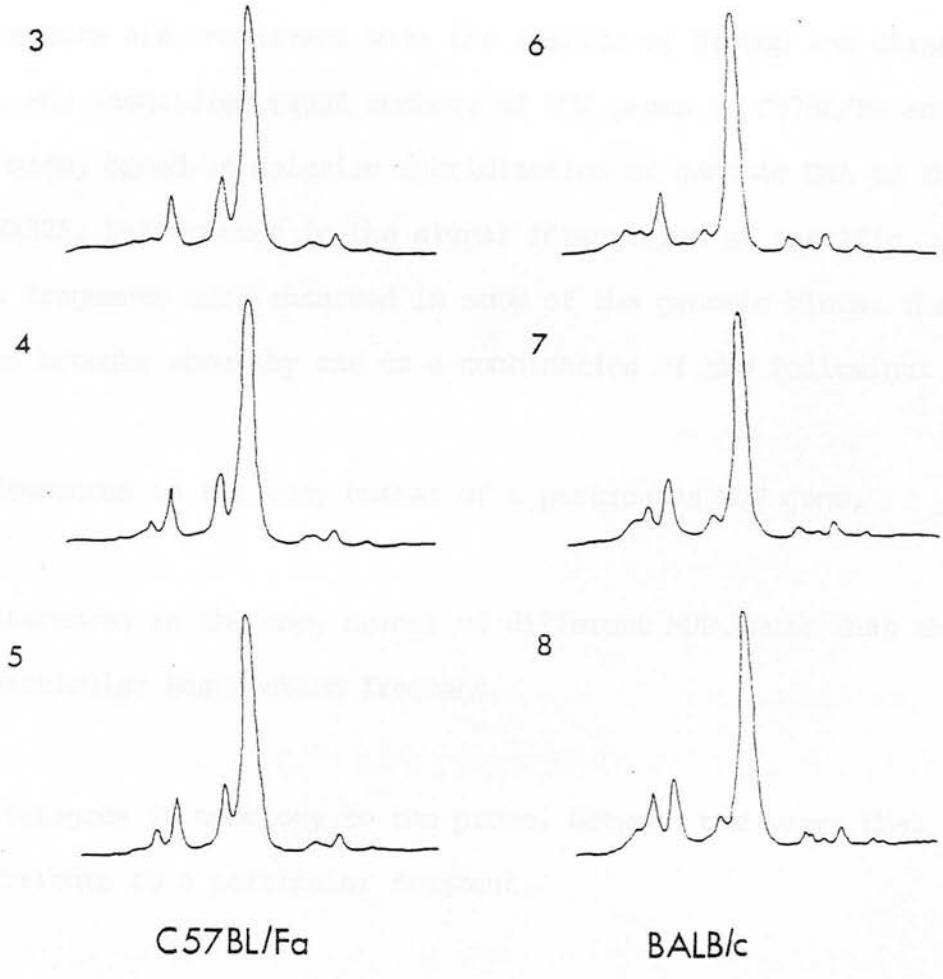



Table R.7.1.

| Incubation time (min) | ratio of BALB/c:C57BL/Fa scan |
|---|---|
| 60 | 0.998 |
| 120 | 0.994 |
| 180 | 1.004 |
| | mean ratio = 0.999 |

to the C57BL/Fa scans was Ø.999, indicating that there are equal numbers of MUP genes in both strains that hybridize to BS-6-5-5 under low stringency conditions.

These results are consistent with the results of Bishop and Clissold (1982), who identified equal numbers of MUP genes in C57BL/Fa and BALB/c mice, based on solution hybridization of genomic DNA to the cDNA LVA325. Differences in the signal intensities of specific genomic fragments were observed in some of the genomic blots. These could be brought about by one or a combination of the following:

(1) differences in the copy number of a particular MUP gene,

(2) differences in the copy number of different MUP genes that share a particular restriction fragment,

(3) differences in homology to the probe, between the genes that contribute to a particular fragment.

The organisation of the 3' flanking region of MUP genes.

The HindIII genomic blot probed with BS-6-1-1 showed two heavily labelled fragments which were ∿6.5 kbp and ∿5.Ø kbp long (Fig. R.7.5). These fragments reflect the common organisation of the 3' flanking region of the MUP genes. The 6.5 kbp fragment is contributed to by genes with a 3' flanking region similar to BS-6. The 5.Ø kbp fragment is contributed to by genes with a 3' flanking

region similar to BS-2 and those with a 3' flanking region similar to BS-1. In the BALB/c samples a minor 6.0 kbp fragment co-migrated with fragment D of BL-15. Bearing in mind the conservation of restriction sites found between the MUP genes, this suggests that the organisation of the 3' flanking region of BL-15 may be unique to this gene. Fragment E of BL-25/CL-2 was represented by a minor band in the digested genomic DNA of each strain. This suggests that the 3' flanking organisation of BL-25/CL-2 is probably uncommon.

The EcoRI genomic digests probed with BS-6-1-1 showed two heavily labelled fragments, which were ∿4.0 kbp and ∿7.0 kbp long. The ∿4.0 kbp fragment is thought to represent the common EcoRI fragment of the group 1 genes, while the ∿7.0 kbp fragment is thought to represent the common EcoRI fragment of the group 2 genes (see Bishop et al, 1982). The genomic fragment which co-migrated with fragment G of BS-1 was estimated to be represented by a single gene in both BALB/c and C57BL/Fa.

BamHI digests probed with BS-6-5-5 also showed that the 9.0 kbp BamHI fragment of BS-1 (fragment R in Fig.R.7.2) does not contribute to a major genomic fragment in either strain. In view of the strong conservation of restriction sites within the regions that hybridize to BS-6-3 (see Fig.R.2.2), the results of the BamHI and EcoRI genomic digests support the conclusion that the organisation of the 3' flanking regions of MUPC3 and MUPC4 is not common among the group 1 genes. The results also support the suggestion that this 3' flanking organisation arose by insertion and/or deletion event(s) that took place within a gene which originally had an organisation

similar to that found in other group 1 genes. The major 5 kbp fragment seen in the HindIII genomic blot is therefore likely to be predominantly contributed to by group 2 genes, which have a 3' organisation similar to BS-2.

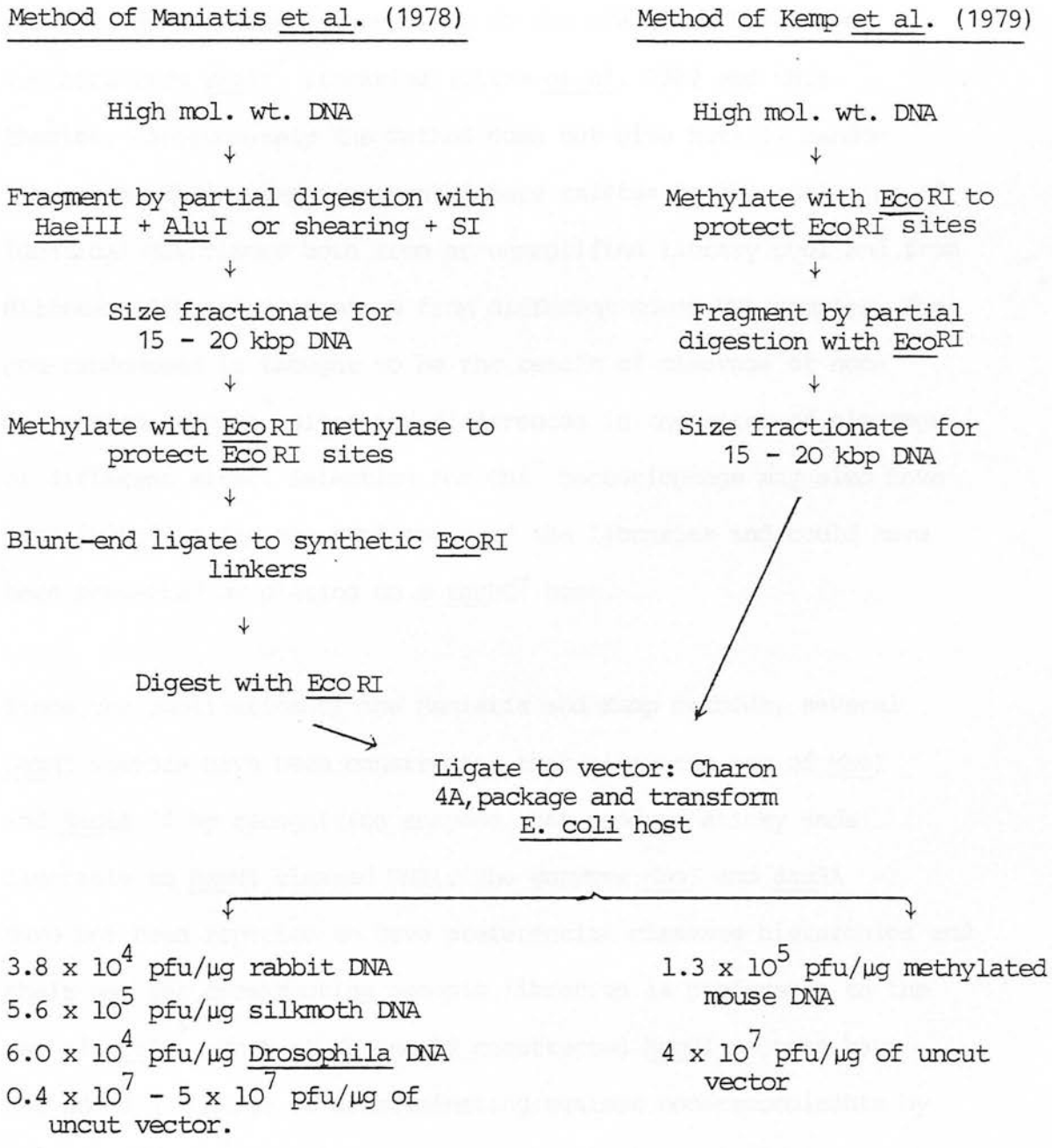
# Discussion

## The cloning of high molecular weight eukaryotic DNA

The cloning of high molecular weight eukaryotic DNA is readily achieved using derivatives of the bacteriophage lambda and *in vitro* packaging extracts (Hohn and Murray, 1977; Grosveld *et al*, 1981). More recently, refinement of cosmid vectors and the methods for screening cosmid libraries have also made these vectors popular, especially where fragments greater than 20 kbp need to be cloned.

Efficient cloning of high molecular weight eukaryotic DNA in a lambda vector was originally achieved by Maniatis *et al* (1978). Although this method is known to give representative libraries, it entails a number of steps that could potentially prove problematical and lead to low cloning efficiency. One of these steps involves blunt end ligation of linkers to the fragmented DNA to be cloned, followed by digestion with the enzyme that recognizes the palindromic sequence in the linker. Kemp *et al* (1979) reported a simpler method for cloning eukaryotic DNA which gave similar cloning efficiencies to the Maniatis method, but eliminated the necessity of using linkers (Fig.D.1). This method takes advantage of the reduced recognition specificity of *Eco*RI from 6 bp to 4 bp under *Eco*RI* conditions, and random and easily ligatable genomic fragments are theoretically generated by partial digestion with *Eco*RI*. Prior methylation of the genomic DNA with *Eco*RI* methylase ensures that the most readily cleaved *Eco*RI* sites, the canonical *Eco*RI sites (GAATTC), are protected; cleavage of these sites would

Figure D.1. Cloning of high molecular weight DNA by the methods of Maniatis et al. (1978) and Kemp et al. (1979).



otherwise result in a non-random population of fragments.

Because of the relative ease with which genomic libraries can be produced by the Kemp method, many of the MUP genes have been isolated from EcoRI* libraries (Clark et al, 1982 and this thesis). Unfortunately the method does not give totally random libraries. The evidence presented here relates to the isolation of identical MUP clones both from an unamplified library pool and from different libraries prepared from different mouse DNA samples. The non-randomness is thought to be the result of cleavage of non-conventional EcoRI* sites and differences in the rates of cleavage of different sites. Selection for $Chi^+$ bacteriophage may also have contributed to the non-randomness of the libraries and could have been prevented by plating on a recBC$^-$ host.

Since the publication of the Maniatis and Kemp methods, several BamHI vectors have been constructed that allow the use of MboI and Sau3A (4 bp recognition enzymes that produce sticky ends ligatable to BamHI cleaved DNA). The enzymes MboI and Sau3A have not been reported to have preferential cleavage hierarchies and their use for constructing genomic libraries is preferable to the use of EcoRI*. Most of the newly constructed BamHI vectors have the added advantage of discriminating against non-recombinants by Spi selection. When using this type of selection, it is important to avoid the preferential growth of $Spi^-$ $Chi^+$ recombinants over $Spi^+$ $Chi^-$ recombinants, and to this end Chi sites have been engineered into the arms of some of these vectors (Loenen and Brammar, 1980).

Many of the new vectors also allow cloning into other sites besides BamHI. [This is convenient when a gene that is to be cloned is known to be contained within a particular restriction fragment.] For example λ L47 can be used as a vector for BamHI, EcoRI, HindIII, BglII, XhoI and SalI fragments (Loenen and Brammar, 1980) and EMBL3 and EMBL4 can be used as vectors for BamHI, BglII, EcoRI and SalI fragments (Frischauf et al, 1983). In the case of the latter two vectors, the BamHI, EcoRI and SalI sites are present in linkers. These facilitate cloning without the removal of the internal fragment and may allow the recovery of the cloned fragments. A new derivative, EMBL3A, that has amber mutations in the A and B genes, can also be used in conjunction with the microplasmid πVX for the selection of specific sequences (Seed, 1983). In this method of screening, a portion of the sequence of interest is cloned into the microplasmid which in turn is used to infect bacteria with amber mutations in their selective marker genes. Since the plasmid is supF, bacteria carrying it will be able to grow under the appropriate selection. These bacteria are then infected with an EMBL3A recombinant library. Because EMBL3A carries amber mutations in the genes A and B, only those bacteriophage which have recombined with the microplasmid will be able to multiply after a second plating on suppressor-free bacteria.

In conclusion, the development of a versatile range of BamHI vectors that accept fragments generated by Sau3A and MboII cleavage has led to considerable simplification in the methodology for constructing random genomic libraries and the isolation of

specific sequences of interest.

Restriction site homologies between cloned MUP genes

MUP clones isolated from the C57 libraries were found to show extensive homology to each other and to the MUP clones isolated from the BALB/c libraries (Clark et al, 1982), implying that Mus musculus musculus MUP genes share a common structure. Exceptions were those genes thought to have undergone re-arrangements.

MUP genes fall into three groups based on their homology to the group 1 and group 2 probes (Bishop et al, 1982). Non-allelic group 1 genes (from the BALB/c libraries) share most of their restriction sites in common. However they can be divided into two sub-groups based on the presence of an insertion and/or deletion at the 5' flanking region of the gene. Alignment of restriction sites suggests that the insertion and/or deletion is located 1.9 kbp or 2.1 kbp 5' to the cap site. It is possible that the insertion and/or deletion is the result of more than one event that took place in the ancestral group 1 genes leading to the current sub-groups. Clarification of this awaits further characterization of the genes.

The insertion and/or deletion is not the only common feature that distinguishes the two sub-groups. Genes with a small 5' HindIII fragment all have a SstI site not present in genes with a large 5' HindIII fragment (this may be the result of the insertion and/or deletion event since it lies within the region to which the

event has been mapped), and all have a KpnI site located ∿7 kbp from the beginning of the transcription unit. No restriction sites that are common to all members of one sub-group were found immediately 5' to the transcription unit, within the transcription unit or 3' to the transcription unit, and most restriction site polymorphisms were found to be confined to a single gene. The exceptions are: (1) the HindIII fragment located 3.8 kbp 5' to the cap sites of BL-7 and MUPC1 (group 1 genes with a small 5' HindIII fragment); (2) a MspI site located ∿300 bp 5' to the cap sites of CL-8/CL-9 and MUPC2 (group 1 genes with a large 5' HindIII fragment); (3) a PstI site lying within the transcription unit of CL-11 and MUPC4 (group 1 genes with a large 5' HindIII fragment).

MUPC2, CL-8/CL-9 and CL-11 are nearly identical to MUPC1 over a range of 4.5 to 7.0 kbp downstream from the end of the transcription unit. The 3' sequence organization found in BS-6 is believed to be representative of that which is present in most other group 1 genes. This conclusion is based on (a) the isolation of C57 group 1 genes with 3' flanking sequences similar to that of BS-6 and (b) the results of the EcoRI and BamHI genomic digests probed with BS-6-1-1 and BS-6-5-5 respectively. The 3' sequence organization of MUPC3 and MUPC4 does not appear to be common and must have arisen after the duplication that led to the division of the group 1 genes.

The Dollo phylogenies, which were obtained using restriction site data, suggest that the group 1 genes with a small 5' HindIII fragment are more homogeneous than those with a large 5' HindIII fragment. This is also borne out by the limited amount of sequencing

data around the TATA box, where BS-1 is found to have a unique A-rich region (P.Ghazal, personal communication). BL-1, CL-8 and CL-11 all share an uninterrupted stretch of A residues, ∿17 bp long, that is located 16 - 18 bp 5' to the TATA box. BS-6, BS-5 and BL-7 have a similarly positioned stretch of uninterrupted A residues that is ∿12 bp long. In these latter three clones the A residues are preceded by a ∿30 bp long A-rich stretch consisting of A residues interrupted occasionally by single C residues. BS-1 appears to have an A-rich region similar to, but longer than, genes with the small 5' HindIII fragment.

The differences in length of the A-rich regions may be the result of 'polymerase slippage' (Efstratiadis et al, 1980) within the mouse genome. They are not thought to be an artifact of replication within the M13 vector since different subclones of the same original clone give identical sequences. It would be interesting to determine whether the A-rich structure, which maps ∿18 bp 5' to the cap sites is of transcriptional importance, since deletion-mutation studies on several genes have identified sequences located at a similar position that influence transcriptional efficiency. Examples are the 1st and 2nd distal signals located at ∿-50 bp and ∿-80 bp from the HSV-tk gene (McKnight, 1982; McKnight et al 1984); the sequence conferring heat shock response located at ∿-50 bp from the Hsp70 heat shock-gene (Pelham, 1982; Pelham and Bienz, 1982); the 'CAAT box' and sequences located at -80 bp from the rabbit β-globin gene (Grosveld et al, 1982) and the regions lying within 100 bp 5' of the α and β interferon genes (Zinn et al, 1983; Ragg and Weissmann, 1983).

Most group 2 genes isolated share extensive restriction site homology. An exception is BL-25/CL-2 which has diverged away from other members in the group, as illustrated by the Dollo phylogenies and confirmed by a limited amount of sequencing. Clark et al (unpublished) have recently completed sequencing the 'transcription unit' of BS-2 (a group 2 gene) and have confirmed the structural similarity of group 2 genes to group 1 genes. From the available sequencing data it appears that group 1 and group 2 genes have diverged in their nucleotide sequences by 10%. BS-2 has a short 5' A residue stretch of 16 bp, similar to that of most group 1 genes with a large 5' HindIII fragment. The precise significance of this is unknown, although it may imply that the ancestral A rich region had such a form. This issue could be settled by more sequence data from the 5' flanking regions of group 2 genes.

The cDNA MUP15 and genomic clones BL-8/CL-4, BL-2 and CL-12 all share a common SstI site present in exon four, and none form stable hybrids with the group 1 and group 2 probes at high stringency. Based on their shared characteristics these clones form a separate group of MUP genes, group 3. CL-12 and the 5' halves of BL-8/CL-4 do not form stable hybrids with MUP15 at 0.2 x SET, 68°C and BL-2 and CL-12 share only very limited homology with BL-8/CL-4 in the region where the latter are not thought to have undergone re-arrangement. [No homology is found in the 3' halves of the clones.] From these observations it is evident that group 3 is a diverged set of genes.

Homologies between genes from the three groups appear to extend into the 5' and 3' flanking regions. Homology between group 1 and group 2 genes was found to extend for at least 3.5 kbp 5' to the cap site. Whether group 3 genes are homologous at their 5' flanking sequences to other members of the gene family is not known, since these sequences were not cloned. A slight bias towards the cloning of the 3' halves of MUP genes is probably due to (a) specific sequences within the regions cloned that resulted in their preferential selection and (b) the use of an incomplete cDNA clone, LVA325, to probe the libraries (this was done because the structure of the MUP genes was not known at that time). Homologies in the 3' flanking regions were found between all three groups to extend up to $\sim$7.5 kbp 3' from the end of the transcription unit. These homologies were often found to be interrupted by postulated insertions and/or deletions of up to $\sim$2 kbp.

The homologies in the flanking sequences, which extend for a few kilobases, suggested that the MUP genes were part of a large duplication unit. Although MUP clones contain up to 8 or 9 kbp of 5' and 3' flanking DNA sequences, linkage was not observed in any of the bacteriophages isolated from the libraries using the probe LVA325. In order to determine whether the MUPs were closely linked on chromosome 4, Clark et al (1984b) isolated a number of over-lapping clones by screening the BALB/c sperm genomic library with 5' and 3' flanking probes. From the screen using the 5' flanking probe, two clones each containing a group 1 gene linked to a group 2 gene were isolated. Restriction mapping of these clones showed that the group 1 and group 2 genes were divergently

orientated and that their 5' ends were separated by ∿15 kbp of DNA. Blots of BALB/c mouse genomic DNA, digested with appropriately chosen restriction enzymes, established that this is the predominant arrangement of group 1 and group 2 genes (Fig.D.2.A).

Divergent orientation of linked members has been found in many other gene families. Examples are the Drosophila heat shock genes at 87A and 87C (Leigh Brown and Ish-Horowitz, 1981), some Drosophila yolk protein genes (Garabedian et al, 1985), the Class II MHC α and β genes of the mouse (Steinmetz and Hood, 1983), the Chorion genes of the silkmoth (Jones and Kafatos, 1980a; 1980b) and some histone genes of Notophthalmus, Xenopus, Chicken and Man (Hentschel and Birnstiel, 1981). The divergent orientations of the genes in these families were most likely brought about by unequal crossing-over, or transposition events.

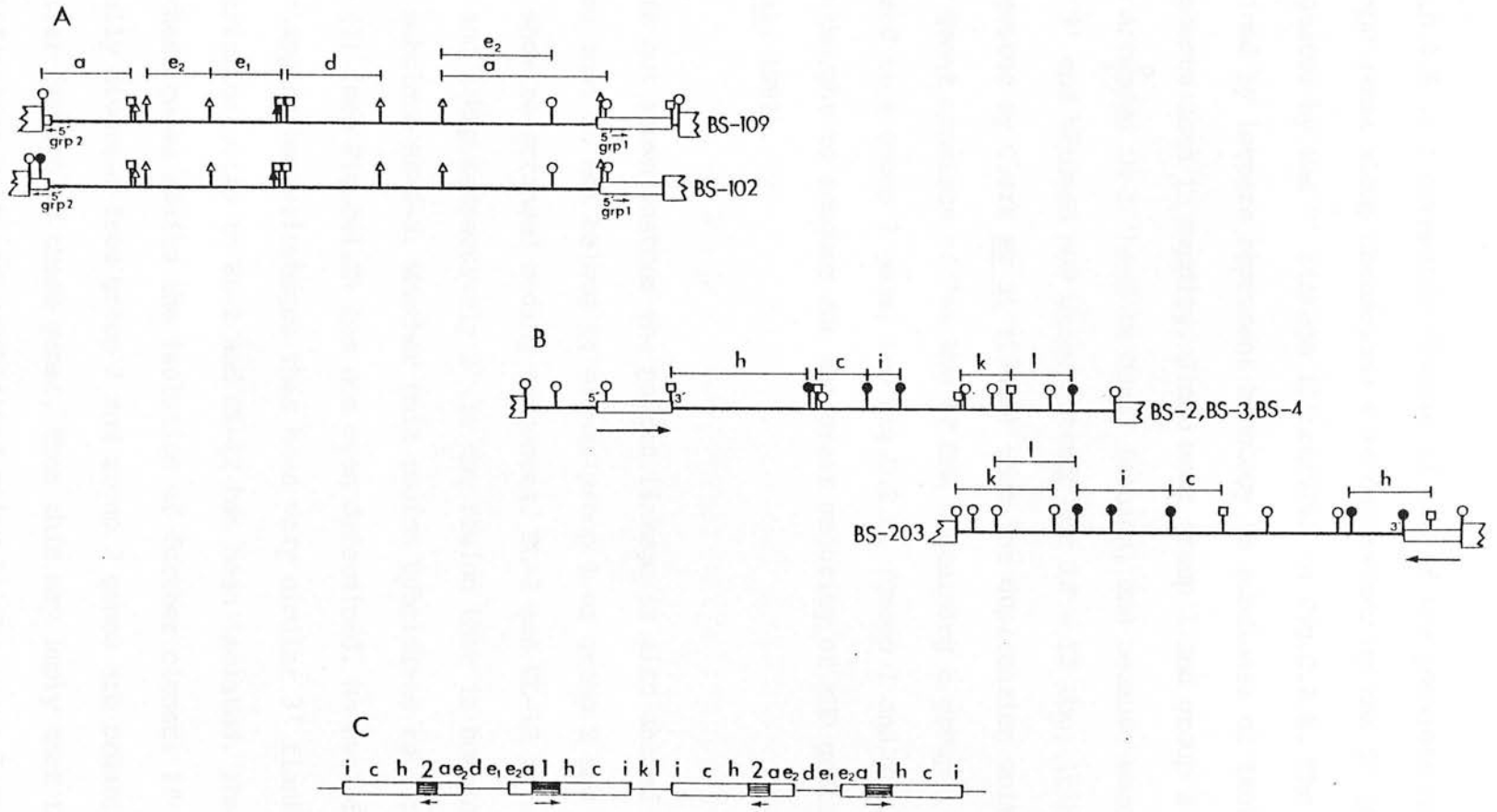
The relative orientation of various subclones in bacteriophages isolated using the 3' flanking probe, suggests that the 3' ends of MUP genes are also linked and are separated by ∿26 kbp or more of flanking DNA. Fig.D.2.B shows two bacteriophages from a much larger set of overlapping clones that are thought to represent 3' linkage (see Clark et al, 1984b). From the bacteriophages isolated it is not possible to determine the exact arrangement of this linkage and we do not know if the 3' ends of group 1 genes are linked to each other or to those of group 2 genes or both. It is possible that more than one arrangement is present as has been found for the Hc (high cysteine) chorion genes (Eickbush and Kafatos, 1982). Isolation of MUP genes from cosmid libraries should give the answer to this

Figure D.2.

(A) Restriction maps of BS-109 and BS-102, clones which show head to head linkage of group 1 and group 2 genes.

(B) Overlapping restriction maps and probe homologies at the 3' flanking region of a set of MUP clones.

(C) A schematic diagram of one possible arrangement of the 45 kbp MUP gene pairs. a, c, d, $e_1$, $e_2$, h, i, k and l, represent known limits of hybridization of the genomic clones to the probes used (see Clark et al., 1984b). Narrow box, MUP transcription unit; broad box, Charon 4A arms. For other symbols see Figure R.2.2.


A
a
e2
e1
d
e2
a
BS-109
BS-102
grp2
grp1
B
h
c
i
k
l
BS-2,BS-3,BS-4
BS-203
C
i c h 2 a e2 d e1 e2 a 1 h c i k l i c h 2 a e2 d e1 e2 a 1 h c i

question.

Fig.D.2.B is a schematic diagram of one of the possible arrangements of MUP genes along chromosome 4 as determined by the 5' linkage and suggested by the 3' linkage illustrated in Fig.D.2.B. The regions defined by letters represent homology to subclones or isolated fragments used in mapping. Since most group 1 and group 2 MUP genes are arranged in a 'head to head' fashion, and because homology at the 3' end between MUP genes extends for 10 - 12 kbp, it has been suggested by Clark et al (1984b) that the duplication unit of the MUP genes consists of ∿45 kbp of DNA containing a group 1 gene paired to a group 2 gene, see Fig.D.2.C. (group 1 and group 2 genes are thought to account for the great majority of MUP genes, Bishop et al, 1982).

It is not known whether the paired linkage is also shared by MUP genes that do not belong to either group 1 or group 2 and which do not show re-arranged coding sequences. BL-2 and CL-12 extend ∿4.5 kbp and 1 kbp respectively 3' to the region that is homologous to the subclone BS-6-3. Whether this region hybridizes to (i) or to (k) and (l) (see Fig.D.1.C) has not been determined. No set of overlapping bacteriophages that have very similar 3' flanking restriction sites to BL-2 and CL-12 has been isolated. The linkage of these genes awaits the isolation of further clones. If MUP genes equally diverged from group 1 and group 2 genes are organized in a similar fashion to these genes, then this may imply that the amplification unit was established prior to the group 1 and group 2 gene differences.

In the chorion gene family, adjacent gene pairs constituting the duplication unit are more closely related to each other and appear to be expressed during a similar developmental period. Overlapping bacteriophages have revealed that the Hc (high cysteine) chorion gene pairs, expressed late in chorionogenesis, are clustered in a stretch of ∿130 kbp of DNA, and that this cluster is flanked by ∿100 kbp and ∿40 kbp clusters of chorion genes expressed in the middle stage of chorionogenesis (Eickbush and Kafatos, 1982). More recently it was shown that the divergently orientated and closely linked Drosophila yolk protein genes, yp1 and yp2, share cis acting regulatory sequences that are necessary for expression of the genes in the ovaries and fat bodies (Garabedian et al, 1985). It is therefore possible that closely linked MUP pairs may be under similar hormonal and/or tissue specific control. However, no universal pattern for the organization of genes within a family and their regulation has been established. This is demonstrated quite strikingly by the different linkage orders found in the globin gene clusters of vertebrates (Dudgson et al, 1979; Hoshbach et al, 1983), and by the different organization of histone genes within and between species (Hentschel and Birnstiel, 1982).

Hybridization, restriction mapping and sequencing studies on the group 1 and group 2 genes have shown that there is greater homology within the groups than between the groups. It is therefore proposed that homogenization events, if they occur, are more common within than between groups (Clark et al, 1984b).

Gene conversion between the divergently orientated hsp7Ø heat-shock genes has been suggested by Leigh Brown and Ish-Horowitz (1981). In their conversion model for divergently orientated genes, these authors proposed that the rate of conversion would be inversely proportional to the distance between the genes. A relatively low rate of homogenization (if it occurs) between genes within a MUP duplication unit, could partly result from the relative orientation of the group 1 and group 2 genes coupled with the large distances between them (∿15 kb compared with 1.7 kb between Drosophila Hsp7Ø divergently orientated genes). It has been speculated by Ollo and Rougen (1983) that the frequency of gene conversion in sequences arranged in tandem may be inversely dependent on the distance between the sequences. If this is true, then unequal crossing-over could be the predominant method of homogenization within the MUP gene family. However, at present we know little about the sequences that influence gene conversion. Evidence for gene homogenization events within the MUP gene family comes from the nucleotide sequence of the cDNA clone MUP15 (discussed in section 6 of the Results). In a later section I will present further evidence for gene homogenization events between the members of the MUP gene family.

Finally, within the proposed 45 kbp duplication unit, certain regions are more conserved than others. The conserved regions include the group 1 and group 2 genes and the region homologous to BS-6-3, [(c) in Fig.D.2.C]. Regions (h) and (i), which flank (c), appear to show differences in the extent of their homology with the subclones used to map the 3' region of the MUP genes. The differences in region (h) have been discussed in the Results

section. While conservation of the expressed group 1 genes may be due to functional selection, it is difficult to explain conservation between the group 2 pseudogenes, given that other regions within the duplication unit are diverging. The divergence of regions (i) and (l) may have been due to the presence of sequences within these regions that made them particularly susceptible to events such as deletion, insertion, and base substitution. The conservation of region (c) on the other hand may be due to some functional importance or may be fortuitous.

Truncated MUP genes

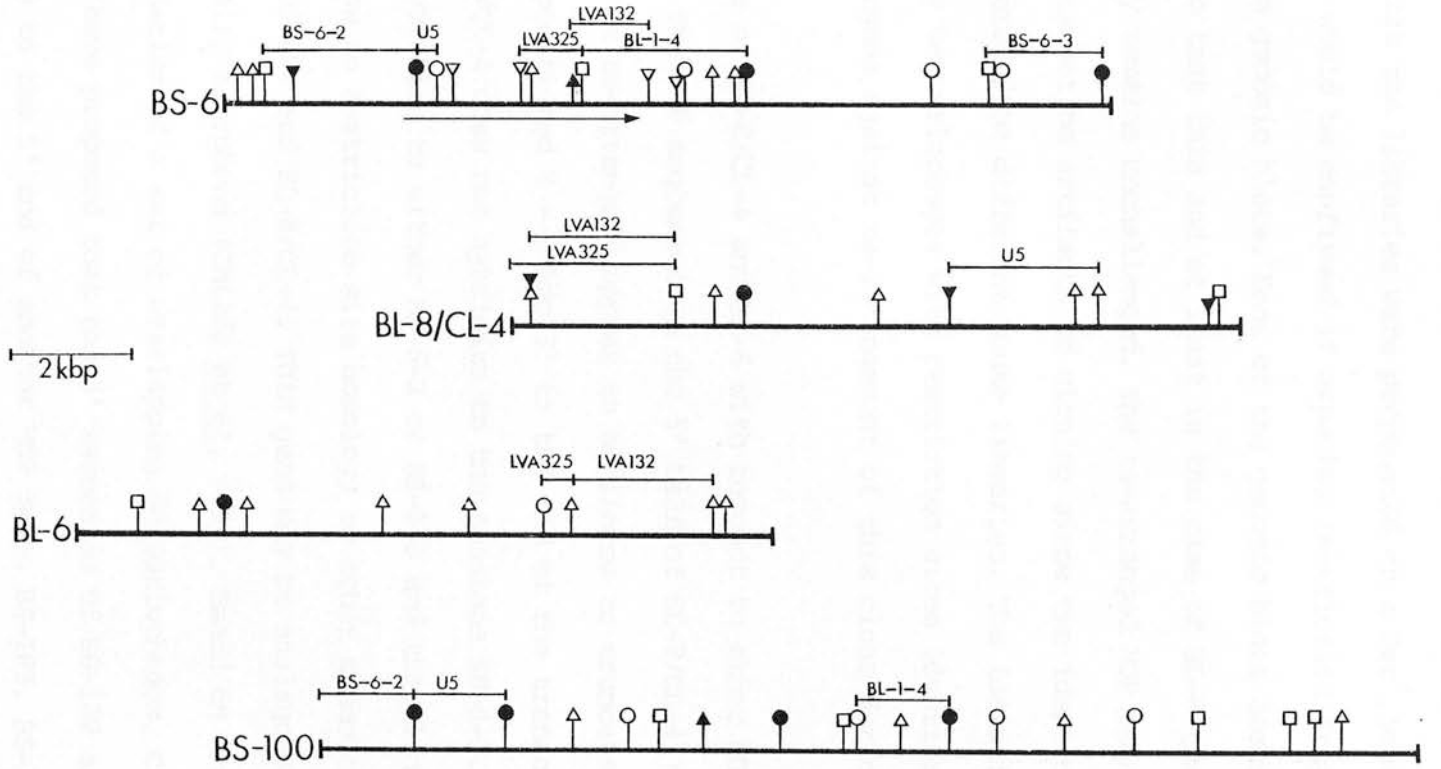
Three truncated MUP genes have been isolated from the genomic libraries. These are BL-6, BS-1ØØ and BL-8/CL-4 (Fig.D.3). BL-6 (Clark et al, 1982) does not hybridize to the subclone BS-6-2 and therefore lacks exon 1. It does not share restriction site homology with other isolated genomic clones in the region that hybridizes to the cDNA LVA325, and unlike other genomic clones, the regions that hybridize to LVA325 and LVAl32 (cDNA clones that share the 5' half of exon 6) do not overlap.

BL-8/CL-4 appears to contain truncated MUP sequences within the 4.4 kbp SstI fragment, as discussed in section 6 of the Results. The exact arrangement of these sequences is not known.

BS-1ØØ contains a MUP pseudogene that presumably resulted from a re-arrangement which caused the loss of exons 3 to 7 (Clark et al,

Figure D.3.

Truncated MUP genes. BL-6, BL-8/CL-4 and BS-100 contain truncated MUP sequences. The restriction map of BS-6 is included for a direct comparison. For symbols, see Figure R.2.2.


BS-6-2
U5
LVA132
LVA325
BL-1-4
BS-6-3
BS-6
LVA132
LVA325
U5
BL-8/CL-4
2 kbp
LVA325
LVA132
BL-6
BS-6-2
U5
BL-1-4
BS-100

1984) and led to the insertion of a 6 - 7 kbp fragment between the 5' half of the gene and its 3' flanking sequences.

Some of these re-arrangements may be cloning artifacts in view of the fact that the libraries were propogated on a $Rec^+$ host. This suggestion would be confirmed if expected restriction fragments are absent from genomic blots. None of the genomic blots described are suitable to test this and at least in the case of BL-6 the possibility remains unchallenged. The re-arranged MUP sequences of BL-8/CL-4 cannot be artifacts of cloning since two identical clones were recovered from different mouse libraries. The isolation of overlapping bacteriophages with restriction sites identical to BS-1Ø0 also argues against re-arrangement of this clone during cloning.

The linkage of BL-8/CL-4 and BL-6 with respect to other MUP genes is not known. The MUP sequences in the 5' half of BL-8/CL-4 (not thought to be re-arranged) appear to be linked to truncated sequences positioned 8 - 9 kbp 3' to the end of the transcription unit. BL-8/CL-4 does not hybridize to the subclone BS-6-3. BL-6 does not hybridize to either BS-6-2 or BS-6-3 and the flanking regions show no restriction site homology to other genomic clones including BS-1Ø0 and BL-8/CL-4. This gene may be analagous to the histone family's orphons (Childs et al, 1981). Based on the characterization of a set of overlapping bacteriophages, Clark et al (1984b) have proposed that the 3' sequences of BS-1Ø0 are linked via ∿23 kbp to the 5' end of another MUP gene, BS-1Ø5. BS-1Ø5 hybridizes poorly to probes derived from group 1 and group 2 genes although it shares some homology in its 5' flanking sequences to

group 1 genes. This homology breaks down after ∿9 kbp. In summary, the MUP pseudogene BS-1ØØ may be linked in a head to tail fashion to a group 3 MUP gene.

Processed pseudogenes have been found in many large gene families. Examples are the mouse α-globin pseudogene (Nishioka et al, 1980; Vanin et al, 1980), the human β-tubulin pseudogene (Wilde et al, 1982; Lee et al, 1983), a human immunoglobulin gene (Hollis et al, 1982) and the mouse ribosomal protein L32 (Dudov and Perry, 1984). Based on restriction mapping none of the isolated MUP clones contain processed pseudogenes.

Homologies between cloned MUP genes and cloned $\alpha_{2u}$globulin genes

The sequences of 5 incomplete and 1 complete liver $\alpha_{2u}$globulin cDNA clones (Unterman et al, 1981; Dolan et al, 1982) as well as the sequence of one $\alpha_{2u}$globulin submaxillary cDNA clone (Laperche et al, 1983), have been reported. The sequence of the exonic regions of an $\alpha_{2u}$globulin gene, 2Ø7, has also been reported (Dolan et al, 1982). Homology between all liver $\alpha_{2u}$globulin coding sequences is greater than 98% at the nucleotide level. Homology between $\alpha_{2u}$globulin salivary and liver coding sequences is ∿95% at the nucleotide level. Genomic blots probed with either the liver cDNA α7, or with a region of intron 6 from 2Ø7, were found to show a similar complexity in the banding pattern when washed at high stringency (Ø.1 x SSC, 65°C), Dolan et al (1982). It therefore seems that, as in the case of MUPs, the $\alpha_{2u}$globulin liver

Table D.1. Nucleotide divergence between MUP and $\alpha_{2u}$ globulin sequences.

| Sequences compared | nucleotide divergence |
|---|---|
| Liver $\alpha_{2u}$ globulin cDNA x submaxillary gland $\alpha_{2u}$ globulin cDNA | 4.7% |
| Liver $\alpha_{2u}$ globulin cDNA x BS-6 transcription unit | 19.0% |
| Liver $\alpha_{2u}$ globulin cDNA x MUP15 | 19.6% |
| Submaxillary gland $\alpha_{2u}$ globulin cDNA x BS-6 transcription unit | 18.9% |
| Submaxillary gland $\alpha_{2u}$ globulin cDNA x MUP15 | 19.2% |
| MUP15 x BS-6 transcription unit | 13.6% |

transcripts are mainly the result of the expression of a very homologous set of genes.

Sequence comparisons of the coding regions, excluding the signal peptide, of BS-6 (a group 1 MUP gene), BS-2 (a group 2 MUP gene) and the rat gene 2Ø7 (Ghazal et al, 1985 ), demonstrated that 2Ø7 is equally diverged from both group 1 and group 2 genes (replacement site divergence= ∿2Ø%) and that group 1 and group 2 genes are more homologous to each other (replacement site divergence= ∿1Ø%) than they are to the $\alpha_{2u}$globulin gene. If the $\alpha_{2u}$globulin cDNA clones isolated are representative of the highly transcribed $\alpha_{2u}$genes in rat liver, then it would appear that these are equally divergent from the highly transcribed MUP genes in mouse liver as they are from the group 2 pseudogenes. A comparison of the nucleotide sequence of the $\alpha_{2u}$globulin liver mRNA, the $\alpha_{2u}$globulin submaxillary cDNA, the group 1 long messenger RNA and the cDNA MUP15 (a group 3 gene) showed that the two MUP sequences were more closely related to each other than they are to the $\alpha_{2u}$globulin sequences, and that the two $\alpha_{2u}$globulin sequences are more closely related to each other than are MUP15 and group 1 mRNA (Table D.1).

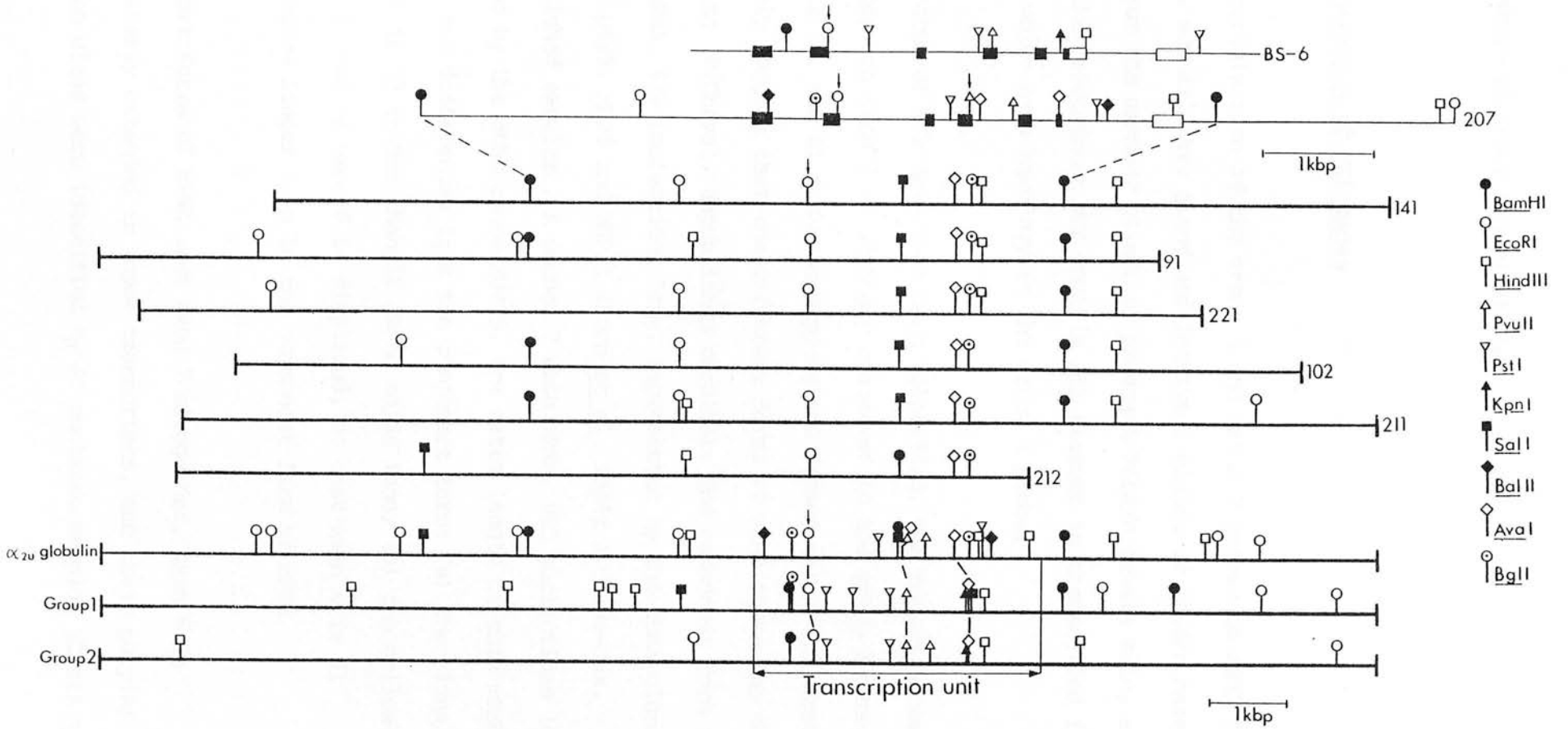
It was possible to compare restriction maps of the $\alpha_{2u}$globulin cDNA clones and the $\alpha_{2u}$globulin gene 2Ø7, isolated by Unterman et al (1981) and Dolan et al (1982), with the $\alpha_{2u}$globulin genes isolated by Kurtz (1981), since four common restriction enzymes were used (EcoRI, HindIII, BamHI and AvaI). It appears that all the $\alpha_{2u}$globulin cloned genes are closely related to each other. Most sites are shared within the coding region. Homology in

restriction sites at the 5' flanking regions appears to extend for at least 3 kbp from the beginning of the transcription unit, while homology at the 3' flanking regions appears to extend for at least 1 kbp from the end of the transcription unit. All available data so far indicate that the $\alpha_{2u}$globulin genes may form a more closely related gene family than the MUPs.

The restriction maps of the isolated $\alpha_{2u}$globulin genes are more similar to each other than they are to the restriction maps of any of the isolated MUP genes (Fig.D.4). A notable difference between the transcription units of the MUP genes and the reported transcription unit of an $\alpha_{2u}$globulin gene (2Ø7) lies in the lengths of the introns (Fig.D.4). Such differences have been noted for many other gene families (see Breathnach and Chambon, 1982). Based on restriction site homologies, it is likely that the transcription unit of 2Ø7 is representative of that which is found in other $\alpha_{2u}$-globulin genes. A common difference in the structure of the genes between the two species could be brought about by different unequal crossing-over events (Smith, 1976).

It will be interesting to determine whether the rat $\alpha_{2u}$globulin genes have a duplication unit similar to that proposed for the group 1 and group 2 genes. This would give an insight into the evolution of the rat and mouse gene families and possibly into their functional importance, if they turn out to have been independently amplified. So far all published $\alpha_{2u}$globulin recombinant bacteriophage extend for a maximum of 9 kb 5' to the EcoRI site in exon 2. It is therefore not possible to speculate on

Figure D.4. Restriction maps of cloned $\alpha_{2u}$ globulin genes. The transcription unit of an $\alpha_{2u}$ globulin gene (from Dolan et al., 1982) is drawn below that of BS-6 for direct comparison. 141, 91, 221, 102, 211 and 212 represent restriction maps of the $\alpha_{2u}$ globulin genes as published by Kurtz, 1981. The three lower lines represent composite maps of the characterized $\alpha_{2u}$ globulin genes, the MUP group 1 genes and the MUP group 2 genes.


BS-6
207
1kbp
141
91
221
102
211
212
α2u globulin
Group1
Group2
Transcription unit
1kbp
BamHI
EcoRI
HindIII
PvuII
PstI
KpnI
SalI
BalII
AvaI
BglI

the nature of their organization.

Expression of MUP genes

The hybridization of the group 1 and group 2 probes to mRNA from the liver, submaxillary gland and lachrymal glands of BALB/c male mice and from the mammary glands of pregnant BALB/c female mice, shows that the predominant MUP mRNA in all tissues is transcribed from genes which have homology to the group 1 probe.

Four forms of MUP mRNA have been identified. The region between the TATA box and exon 7 is strongly conserved in the group 1 genes (see Fig.R.2.2), and SI nuclease mapping and primer extension experiments strongly suggest that the different forms of mRNA share the same cap site (P.Ghazal, unpublished results). The commonest form is a 7 exon mRNA, 838 nucleotides long, represented by the cDNA clones pl057, p499, MUP8 and MUP11 (Kuhn et al, 1984; A.Chave-Cox, unpublished results). A second 7 exon mRNA, 921 nucleotides long, is defined by the cDNA clone MUP15. The extra length of this mRNA is due to two differences from the commonest form: (a) the signal peptide is 22 rather than 18 amino acids long, (b) the splice point at the 3' end of exon 6 is displaced, so that exon 6 is 31 nucleotides longer than in the commonest form of mRNA.

Two short forms of mRNA lack exon 7 altogether. Exon 6 is considerably extended in these transcripts, and their poly(A) addition sites were identified by SI nuclease mapping (Clark et

al, 1984a). The transcripts define mRNAs 741 and 759 nucleotides long, based on the assumption that exons 1 - 5 are identical in structure to the same region of the longer forms. It is possible that some MUP genes give rise to the differently spliced mRNAs. Sequencing data suggest that a short mRNA cDNA clone (LVA325) may represent a transcript of BL-1, a gene believed to be present as a single copy in the BALB/c genome (see Fig.R.7.5). However, the cDNA clone LVA325 may represent a truncated and incompletely processed mRNA for two reasons: (a) the SI nuclease protection studies indicate that the short forms of mRNA terminate at nucleotides 158 - 16Ø and 176 - 178 of exon 6; in contrast, LVA325 terminates at nucleotide 166 of exon 6. (b) LVA325 contains intron 5.

Alternative splicing in the 3' untranslated region, is also found in the transcripts of the $\alpha_{2u}$globulin genes (Laperche et al, 1983). The mRNA that is represented by the liver cDNA clones is spliced similarly to the predominant MUP liver mRNA. A submaxillary cDNA clone, however, represents an mRNA which contains an extra 121 nucleotides of 3' untranslated sequence derived from the 3' region of intron 6.

Unlike the long liver mRNA, the short liver mRNA was found to hybridize preferentially to the group 2 probe. Whether the transcripts homologous to the group 2 probe represent transcripts of group 2 pseudogenes, transcripts of group 2 functional genes or transcripts of genes having a 3' structure homologous to group 2 genes, is not known. The first suggestion is possible because all the splicing donor and acceptor sites follow the GT-AG rule

Table D.2. Comparison of the TATA boxes of group 1 and group 2 MUP genes with the consensus TATA box.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | G | - | G | T | A | T | A | A/T | A | A/T | - | G | - | - | G | consensus from 60 genes (Breathnach & Chambon, 1981). |
| | A | 10 | | 8 | 4 | 58 | 4 | 51 | 38 | 53 | 30 | 20 | 11 | | | 14 | |
| Base | T | 10 | | 6 | 49 | 1 | 56 | 6 | 22 | 6 | 20 | 7 | 9 | | | 10 | |
| Frequency | G | 30 | | 32 | 1 | 1 | 0 | 0 | 0 | 0 | 8 | 23 | 29 | | | 26 | |
| | C | 9 | | 14 | 6 | 0 | 0 | 3 | 0 | 1 | 2 | 10 | 11 | | | 10 | |
| | | G | A | G | T | A | T | A | T | A | A | G | G | A | C | A | Group 1 genes |
| | | G | A | G | T | A | T | A | T | G | A | G | G | A | C | A | BS-2 |

(Breathnach and Chambon, 1981) and in view of the fact that pseudogene transcripts have been represented in cDNA libraries (e.g. the leukocyte interferon cDNA clone LeIFN6, Goeddel et al, 1981).

The sequence TATATGA in BS-2 is located at an identical position to the TATA box in group 1 genes. This sequence is not identical to that of the group 1 TATA box, TATATAA, or to the consensus sequence TATA(AT)A(AT) drawn from 60 eukaryotic genes (Breathnach and Chambon, 1981; Table D.2). None of the 60 TATA boxes compared by Breathnach and Chambon (1981) contain G residues in their 6th positions, arguing for strong selection against a G residue at this site (see Table D.2). Point mutations within the TATA box have been found to result in a marked reduction in transcription efficiency (Dierks et al, 1983). Therefore, if BS-2 is transcribed at all, its rate of transcription initiation is likely to be low. The TATA boxes of three other partially sequenced group 2 genes (P.Ghazal, unpublished results) do not contain G residues. Thus the A to G transition found in the TATA box of BS-2 must have been acquired after the redundancy of an ancestral group 2 gene.

The possibility that most of the mRNA is contributed by transcripts of functional genes homologous to the group 2 probe could be tested by selecting liver mRNA with a synthetic oligonucleotide probe that does not cross-hybridize with group 1 transcripts, translating the selected mRNA in vitro and challenging the translation products with MUP antibody. Alternatively a cDNA library enriched in sequences corresponding to the short MUP mRNA may be screened with sequences specific to

group 2 MUP genes and the isolated clones then characterized.

Kuhn et al, (1984) have prepared 5' subclones of the cDNA clones p199 and p499 and have found that male lachrymal RNA and male and female liver RNA of C57BL/6J mice contain significant amounts of MUP mRNA homologous to p199. Translation products of liver mRNA selected with the 5' p199 subclone gives a set of proteins that are slightly more acidic than those selected by the 5' p499 subclone (Kuhn et al, 1984). The full-length p499 cDNA selects all the liver MUP mRNA, i.e. both those species selected by the 5' p199 sequence and those selected by the 5' p499 sequence. When full-length p499 was used to select lachrymal sequences, the translation products were considerably more basic than other MUP proteins, and in particular did not coincide with the translation products of the liver mRNA selected by 5' p199. Thus the lachrymal mRNA sequences that hybridize preferentially with 5' p199 are similar but not identical to p199. This observation is consistent with the observed loss of hybridization signal when 5' p199 probed lachrymal mRNA was washed at high stringency.

It is useful to establish a phylogenetic relationship between members of a gene family since there is often a correlation between similarity in sequence and similarity in function. For example, leukocyte interferons are more homologous to each other than they are to a fibroblast interferon, although the interferon genes are thought to have originated from a common ancestor (Taniguchi et al, 1980 ). Thus approximately 60% of the amino acids are common

between leukocyte interferons compared with approximately 23% between leukocyte and fibroblast interferons. The globin gene family provide another example. Members that give rise to the α-like polypeptide chains are more homologous to each other than those that give rise to the β-like polypeptide chains and vice versa. An extreme example within the same gene family are the human $\alpha_1$ and $\alpha_2$ genes that code for identical α-polypeptide chains even though they are thought to be the products of a duplication that took place prior to the mammalian divergence (Liebhaber et al, 1981; Zimmer et al, 1980). The rat $\alpha_{2u}$globulin genes provide yet another example. The liver and submaxillary $\alpha_{2u}$globulin translation products can be distinguished by isoelectric focusing. The sequence of a submaxillary cDNA clone differs from isolated liver cDNAs by ~5% in its nucleotide sequence, while different liver transcripts only differ by ~1% (Laperche et al, 1983).

That such a phenomenon is also true for the MUP genes has recently been suggested by the work of Shaw et al (1983) and Shahan and Derman (1984). Shaw et al (1983) found that processed lachrymal MUPs were considerably more basic than MUPs from other tissues. Shahan and Derman (1984) studied the restriction enzyme patterns of cDNA prepared from different MUP expressing tissues and found different HaeIII or MboII restriction patterns for each of the tissues they examined: liver, lachrymal, submaxillary and sublingual glands. They interpreted the results to show that different MUP genes are expressed in different tissues. Some caution must be exercised in the interpretation of these results since the cDNA pools used in the restriction analyses were not sized. However, the

restriction patterns obtained in the liver agree with what would be expected from MboII and HaeIII restriction digests of the sequenced group 1 cDNAs and genes (BL-1, BS-1, BS-6, BS-5, pl057, p499, MUP11, MUP8) and from the sequence of MUP15, if we assume that cDNA pools used were 2/3 - 3/4 the length of the full messenger RNA and were lacking in 5' sequences. Shahan and Derman (1984) noted that the differences are unlikely to be the result of alternative splicing since similar EcoRI+AvaI restriction patterns were obtained in the different tissues.

The combined patterns of MboII and HaeIII restriction digests suggest that the predominant mRNA species in the submaxillary, lachrymal and sublingual glands of Swiss white mice (NCS) are not homologous to the sequenced group 1 genes and cDNAs, or to MUP15. BALB/c MUP mRNA of all tissues that have been examined, hybridizes preferentially to the group 1 probe. It is possible that Swiss white alleles of those isolated group 1 genes that have not been sequenced, CL-8/CL-9, CL-11, BL-7 and MUPC4, contribute to the different cDNA populations in tissues other than liver. However, CL-8/CL-9, CL-11, BL-7 and MUPC4 are closely related to the sequenced group 1 genes (BS-6, BS-1, BL-1, and BS-5) and are therefore more likely to contribute to the liver mRNA population.

MUP expression in the lachrymal and submaxillary glands appears to be the result of transcription from genes that hybridize preferentially to the group 1 probe, but which have diverged away from the group 1 genes isolated. BL-2 and CL-12 and the un-rearranged MUP sequences of BL-8/CL-4 are unlikely to contribute to

the predominant mRNA species in these tissues as they hybridize ∿equivalently to group 1 and group 2 probes at high stringency.

Shahan and Derman (1984) also analyzed the liver, lachrymal gland, submaxillary gland and sublingual gland precursors of MUPs by *in vitro* translation of hybrid selected mRNA in a rabbit reticulocyte lysate, followed by immunoprecipitation and IEF or Laemmli gel electrophoresis. Like Shaw *et al* (1983) they obtained ∿12 distinguishable liver components. However, unlike Shaw *et al* (1983) who obtained a single distinguishable component for the submaxillary gland and a single distinguishable component for the sublingual gland, Shahan and Derman (1984) obtained ∿4 distinguishable components for each of these tissues. The lachrymal products were found to be substantially more basic than the MUP components from other tissues, in agreement with Shaw *et al* (1983), although only 3 components were distinguished. The differences in the number of components observed in the lachrymal, submaxillary and sublingual glands between the two groups of workers may be due to one or a combination of the following:

(1) Differences in the *in vitro* translation systems. Shaw *et al* (1983) included dog pancreas membranes to process the protein products, unlike Shahan and Derman (1984) who used a rabbit reticulocyte lysate system devoid of membranes. Consequently, differences due to the signal peptide would not have been observed by Shaw *et al* (1983) while differences due to glycosylation and other such post-translational modifications would not have been observed by Shahan and Derman (1984).

Table D.3. N-terminal sequence differences between MUP1, MUP2, MUP3 and the cDNA, MUP15.

| Comparison | Difference |
|---|---|
| MUP1 - MUP2 | 0 |
| MUP3 - MUP1 and MUP2 | 8/36 |
| MUP15 - MUP1 and MUP2 | 11/36 |
| MUP3 - MUP15 | 3/36 |

(2) Shaw et al (1983) used two dimensional electrophoresis while Shahan and Derman (1984) separately used one dimensional electrophoresis and IEF.

(3) Differences in mouse strains. Shaw et al (1983) used C57BL/6J mice; Shahan and Derman used Swiss white mice.

Finlayson et al (1974) determined the amino-terminal sequences of the major urinary components MUP1, MUP2 and MUP3 by isolating MUP1 and MUP3 from the urine of BALB/c mice and MUP2 from the urine of C57BL mice. The region sequenced covers parts of exons 1 and 2: ∿ fourteen C-terminal amino acids of exon 1 and ∿ twenty-two N-terminal amino acids of exon 2. MUP1 and MUP2 were found to be homologous if not identical, while MUP3 differed from MUP1 and MUP2 by 8/36 amino acids. Comparison of the amino acid sequences of MUP1 and MUP2 to those of the products of the sequenced group 1 genes and cDNA clones, shows that they are encoded by transcripts of group 1 genes. MUP3 appears to be more homologous to the cDNA MUP15 than it is to the group 1 genes (Table D.3). Since more than 3 components can be resolved by isoelectric focusing of the urinary MUPs of BALB/c and C57BL mice, the amino acid sequences of Finlayson et al (1974) for each of the components MUP1, MUP2 and MUP3 are likely to represent an average of more than one protein.

Two dimensional gel electrophoresis of the processed translation products of hybrid selected MUP mRNA from the livers of male

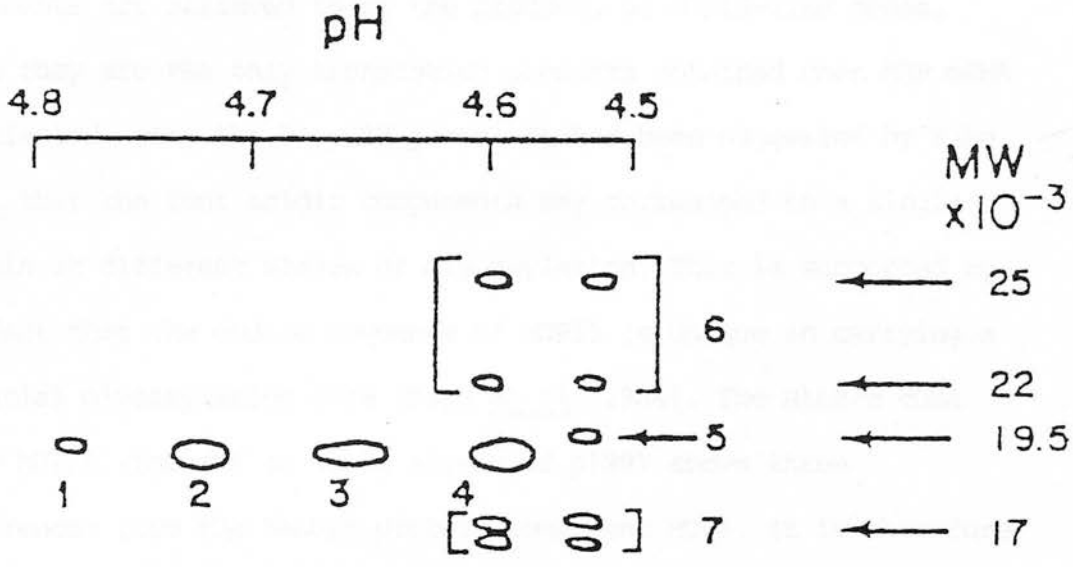



Figure D.5. Sketch representing the processed translation products of hybrid-selected MUP mRNA after two dimensional gel electrophoresis. From Kuhn et al (1984).

C57BL/6J mice and of the urinary proteins of male C57BL/6J mice shows that there are up to ∿12 distinct components (Kuhn et al, 1984). Fig.D.5 is a sketch representing the patterns observed. The four most acidic components show charge and size heterogeneity between the urinary products and translation products. These components are believed to be the products of MUP15-like genes, since they are the only translation products obtained when MUP mRNA is selected using the 5' p199 probe. It has been suggested by Kuhn et al that the four acidic components may correspond to a single protein at different stages of glycosylation. This is supported by the fact that the coding sequence of MUP15 is unique in carrying a potential glycosylation site (Kuhn et al, 1984). The BALB/c cDNA clone MUP15 (thought to be an allele of p199) shows three differences from the BALB/c protein component MUP3. It is therefore likely that at least in BALB/c mice more than one MUP15-like gene is expressed in the liver.

The processed translation products of hybrid selected MUP mRNA contain a set of low molecular weight components (17,ØØØ) whose mRNAs are not selected by the 5' p199 subclone or the 5' p499 subclone, although they are selected by the full p499 clone (the low molecular weight components are shown as component 7 in Fig.D.5). These components are not the result of transcription from any of the cloned group 1 genes since sequence comparisons show that all the sequenced group 1 genes are very similar to p499 and since all cloned group 1 genes form stable hybrids with the 5' p499 subclone. Due to the hybridization of their mRNAs, it is possible that the low molecular weight components represent transcripts of BL-2 and CL-12

Table D.4 Nucleotide and amino acid differences between group 1 transcription units.

| Clone origin | Clone | Number of base pairs sequenced | Boundaries | Clone compared with: BL-1 | BS-1 | BS-5 | MUP11 | MUP8 | p499 | p1057 | LVA325 | LVA325 | BL-1 'short mRNA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Genomic; sperm BALB/c | BS-6 | 878 | Exons 1 → 7 | 4 + + | 5 + | 3 + | 4 + | 4 + + | 7 + + | 8 + + | 1 | 3 | 5 |
| Genomic; liver, male BALB/c | BL-1 | 878 | Exons 1 → 7 | | 5 +++ | 3 +++ | 6 +++ | 0 | 5 + + | 8 ++++ | 1 | 0 | |
| Genomic sperm BALB/c | BS-1 | 878 | Exons 1 → 7 | | | 4 + + | 7 + + | 5 +++ | 6 +++ | 5 + | 1 | - | - |
| Genomic:sperm BALB/c | BS-5 | 878 | Exons 1 → 7 | | | | 5 + + | 3 +++ | 6 +++ | 7 +++ | 1 | - | - |
| cDNA:liver, female BALB/c | MUP11 | 726 | 3' 9bp of exon 1 → 269 bp 3' of TGA stop codon in exon 6 | | | | | 4 + + | 9 +++ | 10 +++ | 1 | - | - |
| cDNA:liver, female BALB/c | MUP8 | 682 | 3' 99 bp of exon 2 + exons 3 → 7 | | | | | | 3 + | 6 ++++ | 1 | - | - |
| cDNA:liver,male C57BL/6J | p499 | 789 | 3' 72 bp of exon 1 + exons 2 → 7 | | | | | | | 5 ++++ | 1 | - | - |
| cDNA:liver, male BALB/c | p1057 | 801 | 3' 84 bp of exon 1 + exons 2 → 7 | | | | | | | | 1 | - | - |
| cDNA:liver, male BALB/c | LVA132 | 215 | Exon 6 + 169 bp of exon 7 | | | | | | | | | - | - |
| cDNA: liver, male BALB/c | LVA325 | 170 | Exon 6 + 124 bp of intron 6 | | | | | | | | | | 0 |
| Genomic:liver, male BALB/c | BL-1 'short mRNA | 797 | Exons 1 → 6 + 169 bp of intron 6 | | | | | | | | | | |

KEY: number: base substitutions
+: amino acid change

or MUP genes homologous to these clones. It is also possible that these components represent products of group 2 genes. However, the latter suggestion appears to be less likely since mRNA homologous to the group 2 probe represents a much smaller proportion of the total MUP mRNA than component 7 appears to represent of the total MUP mRNA translation products.

Most sequenced group 1 genes and cDNA clones potentially give rise to different mature proteins of the same size. The cDNA clone p499 may represent an allele of BL-1, since both lack the EcoRI site in exon 2 due to a T to A transversion. However Table D.4 indicates that the cDNA p499 is as divergent from BL-1 as some non-allelic genes are from one another. It is therefore more likely to represent a transcript of an allele of an uncloned BALB/c group 1 MUP gene. If all the sequenced genes are transcribed in the liver then at least seven different group 1 protein products may contribute to the urinary MUP proteins and these would probably be represented by components 1,2,3,4 and 5. Tables D.5 and D.6 give the differences in basic and acidic amino acids of the sequenced group 1 genes and cDNAs from their consensus protein sequence. These results suggest that BS-6 and BS-5 probably contribute to the more basic MUPs, p499 and MUP11 to the intermediately charged MUPs and BL-1, pl057 and BS-1 to the more acidic MUPs of components 1,2,3,4 and 5. It is of course possible for different proteins of the same charge (and size) to contribute to one component.

Some of the more closely related group 1 genes that contribute to the same liver MUP mRNA translation products are likely to be under

Table D.5. Amino acid differences from the consensus group 1 sequence.

| Amino acid change | Charge change | Cloned gene/cDNA |
|---|---|---|
| TTG → GTG<br>Leu → Val | nonpolar → nonpolar | p499 |
| GTA → AGA<br>Val → Arg | nonpolar → + | MUP11 |
| AGA → AAA<br>Arg → Lys | + → + | BS-5 |
| AAT → AAA<br>Asn → Lys | 0 → + | BL-1, p499 |
| TTC → GTC<br>Phe → Val | nonpolar → nonpolar | p1057 |
| CAA → AAA<br>Gln → Lys | 0 → + | BL-1 |
| GAG → AAG<br>Glu → Lys | - → + | BS-1, p1057 |

Table D.6. Charge differences of MUP protein sequences relative to the consensus group 1 protein sequence.

| Cloned gene/cDNA | Difference in charge | Net difference in charge |
|---|---|---|
| BS-6 | 0 | 0 |
| BS-5 | 0 | 0 |
| p499 | 0 → + | +1 |
| MUP11 | 0 → + | +1 |
| p1057 | - → + | +2 |
| BS-1 | - → + | +2 |
| BL-1 | 0 → +, 0 → + | +2 |

similar hormonal control. The hormonal regulation of different liver MUP mRNAs has been investigated by Knopf et al (1984), by treating C57BL/Fa thyroidectomized female mice with different hormonal regimes. The total level of MUP mRNA was assayed by Northern blot analysis and the induced MUP mRNAs were identified by two-dimensional gel electrophoresis of the translation products.

The MUP mRNA level in the livers of thyroidectomized female mice was found to be ∿1/10th of that found in the livers of normal female mice. Treatment of the surgically altered mice with growth hormone, resulted in a four fold increase in MUP mRNA. Two-dimensional electrophoresis of the translation products of growth hormone treated thyroidectomized mice showed a very similar pattern to that of normal mice: only component 3 and low levels of components 6 and 7 were present (see Fig.D.5). Treatment of the thyroidectomized mice with testosterone resulted in a ten-fold increase in MUP mRNA; components 6 and 7 were slightly induced, while components 2 and 4 were substantially induced. As in the growth hormone treated mice and normal mice, component 3 was found to represent a major product. Treatment of the thyroidectomized mice with thyroxine resulted in a substantial increase in components 1,4 and 6 and once again component 3 represented a major product. These results illustrate that there is substantial variation in the hormonal regulation of different liver MUP components. To understand the hormonal regulation of the MUP genes, it will be necessary to identify the genes that contribute to the different components. As mentioned above, MUP15 (pl99) like sequences are thought to give rise to component 6.

Shaw et al (1983) found that the major translation product(s) of the mammary gland co-migrated with component 3 of the liver, and that the translation product(s) of the submaxillary gland co-migrated with component 5 of the liver. MUP expression in the mammary gland is only detected after the first pregnancy. This implies that the regulation of component 3 must differ in the liver and mammary glands. MUP mRNA in the submaxillary gland is not modulated by hormones. Therefore regulation of component 5 must also differ in the liver and submaxillary gland. The more basic lachrymal gland MUPs appear to be regulated by testosterone. In hypophysectomized female mice, lachrymal gland MUP mRNA can be induced by testosterone; testosterone induction of liver MUP mRNA is not observed in hypophysectomized female mice (Shaw et al, 1983). These differences in hormonal regulation could relate to tissue specific differences or could be brought about by differences in the regulatory regions of different MUP genes. The results of Shahan and Derman (1984) and Shaw et al (1983) argue that at least in the submaxillary and lachrymal glands (where different sets of MUPs are expressed to the set which is expressed in the liver) some of the differences are likely to be gene specific.

Variation in the major urinary protein genes between inbred strains

Variation in major urinary proteins between inbred strains has long been established (Finlayson et al, 1963). This variation is of both a qualitative and quantitative nature, and could be brought

about by transcriptional and post-transcriptional control (e.g. relating to mRNA stability) or post-translational events and/or differences in the MUP structural genes.

The processed translation products of p499-selected mRNA are almost identical in pattern to the urinary MUPs (Knopf et al, 1983). It is therefore unlikely that the differences in urinary proteins of inbred strains (Mus musculus musculus) are due to differences in the expression of MUP genes that are not detected by our probes under low stringency conditions. Further evidence comes from protein sequence data on the urinary proteins. Finlayson et al (1974) sequenced the 36 N-terminal amino acids of the classically identified major urinary protein components MUP1, MUP2 and MUP3. These components were purified by passing non-dialyzable material from the urine of BALB/c and C57BL mice through a sephadex G100 column and then subjecting the pooled fractions of interest to chromatographic separation on DEAE-cellulose. The partial sequences of the pooled fractions representing MUP1, MUP2 or MUP3 show that at least the majority of the proteins constituting these components have 5' sequences very similar to those of cloned MUP genes and MUP cDNAs.

The in vivo rate of MUP synthesis in the liver and the rate of MUP excretion both vary in different mouse strains. The relationship between the two is not constant, and this led Berger and Szoka (1981) to propose the existence of unknown post-translational effects.

The *in vivo* rate of MUP synthesis in the livers of C57BL/6J female mice is twice as high as it is in C3H/HeJ female mice, unlike the total concentrations of MUP mRNA, which are the same. Berger (1983) interpreted this to show that post-transcriptional control contributes to the differences in the urinary MUPs of some inbred strains. It is not known whether this type of control acts on all or some species of MUP mRNA.

Differences in the transcription of MUP genes between inbred strains have been examined by comparing the products of MUP-selected mRNA translated in a membrane-free rabbit reticulocyte lysate (Clissold and Bishop, 1982). The differences observed between C57BL/Fa and BALB/c mice using this type of analysis are again both quantitative and qualitative. Transcriptional differences could be due to physiological differences between the inbred strains (e.g. in the concentrations of circulating hormones or in the abundance of hormone receptors; see Mainwaring, 1983) and/or variation in the MUP genes.

That some of the differences are due to variation in MUP genes is strongly suggested by the fact that restriction enzyme differences in the MUP structural locus are observed between inbred mouse strains, and by the fact that the suggested MUP regulatory locus *MUP-1* (Szoka and Paigen, 1978) is linked to the MUP structural locus (Bennett *et al*, 1982). Some of the variation appears to be due to restriction enzyme differences in the group 2 pseudogenes as suggested by the isolation of BL-15. That all the restriction enzyme differences are unlikely to be the result of pseudogenes is

demonstrated by the isolation of CL-8/CL-9, a MUP gene with a unique restriction fragment present in C57BL/Fa but not BALB/c mice. CL-8/CL-9 shows no detectable re-arrangements compared to other group 1 genes with a large 5' HindIII fragment and the sequence of the first exon and 5' flanking region containing the TATA box of CL-8 is very similar to that of other group 1 genes. Also all sequenced group 1 genes with a small or large 5' HindIII fragment have been found to have correct open reading frames that could potentially code for a MUP protein.

The variation in genomic DNA between BALB/c and C57BL/Fa mice consisted of: (1) the presence of variant restriction fragments, (2) differences in the intensity of hybridization of individual restriction fragments.

Differences between the hybridization signals of co-migrating fragments of BALB/c and C57BL/Fa genomic digests could arise in a number of ways. Bearing in mind the conservation of restriction sites in the MUP gene family, most of the differences probably represent differences in homology to the probe and/or copy number of homologous restriction fragments. BALB/c and C57BL/Fa mice have the same total number of MUP genes. I therefore propose that differences in the hybridization signals of co-migrating genomic fragments between BALB/c and C57BL/Fa are due to different homogenization events that took place in the two mouse lineages from which BALB/c and C57BL/Fa mice were derived.

Variation, between C57BL/Fa and BALB/c, in the concentrations of

different MUP mRNAs, has been illustrated by isoelectric focusing of unprocessed MUP mRNA translation products (Clissold and Bishop, 1982). This variation may correlate with the differences between some of the co-migrating fragments of BALB/c and C57BL/Fa genomic digests. DNA coding sequences of group 1 genes are very homologous ( >= 99% ) and it is therefore possible that, within a strain, some MUP genes give rise to the same protein. Divergence of non-allelic or allelic genes between strains could thus give rise to differences in the number of genes coding for the same protein without affecting the total number of genes. To date, however, all group 1 genes that have been sequenced and most of the sequenced cDNAs give rise to different proteins (Table D.4).

Bennett et al (1982) did not observe any variation in the hybridization signals of specific restriction fragments between C57BL/6By, C57BL/6J and C57L/J on the one hand and BALB/cBy, C3H/HeJ, AKR/J and DBA/2J on the other, when the genomic DNAs of these strains were digested with HindIII. Such differences were also not observed in EcoRl digests of the genomic DNAs of AKR/J and C57L/J. The results of the BALB/c and C57BL/Fa HindIII and EcoRI genomic digests probed with BS-6-1-1 are in agreement with the results of Bennett et al (1982). These show no striking variation between the strains in the hybridization signals of individual fragments although a 1.5 kbp HindIII fragment is ∿ twice as heavily labelled in the C57BL/Fa samples (see Fig.R.7.4). Therefore what initially appeared to be a discrepancy between the results described in this thesis and those of Bennett et al (1982) can now be explained by differences in the choice of enzymes and

probes.

Gene polymorphisms are often associated with particular restriction enzyme polymorphisms within or flanking the transcription unit. Examples are provided by the human β-globin gene (Orkin et al, 1982) and genes within the mouse major histocompatability complex (Steinmetz et al, 1982; Steinmetz et al, 1984). Therefore variant restriction fragments between strains could represent genes that contribute to qualitative differences in the mRNA populations of C57BL/Fa and BALB/c mice.

Relationship of the C57 MUP clones to the BALB/c MUP clones.

Using eight 6 base-pair restriction enzymes, five group 1 genes (CL-5, CL-1Ø, CL-8, CL-9, CL-11) and a group 3 gene (CL-12) were identified which have different restriction maps from those of the cloned BALB/c genes. Mapping with MspI also revealed that BL-7 had a unique restriction map. Due to contamination of the C57 mouse strain it is possible that some BALB/c genes have been isolated from the C57 library. However, this appears to be unlikely due to the following:

(1) a genomic blot comparing BALB/c DNA with the C57 DNA used to construct the library, does not support extensive contamination, if any, of the C57 MUP structural locus with the BALB/c MUP structural locus;

(2) a clone (CL-8/CL-9) having a restriction fragment unique to

C57BL/Fa genomic DNA was isolated from the C57 library;

(3) clones with restriction fragments not represented in the C57BL/Fa genomic DNA were not isolated.

C57 and BALB/c clones that are identical to each other probably represent alleles or very closely related genes. It is possible that some of the identical clones have sequence differences which could be distinguished by finer mapping. However, sequencing may be necessary to identify further variation, bearing in mind that homology between non-allelic group 1 genes is >=99%.

Some of the clones with restriction enzyme differences may represent alleles of MUP genes previously isolated. The Dollo Parsimony phylogenies suggest that CL-8 is most closely related to BL-1. These two clones differ from one another by three 6-base restriction sites and by one MspI site. This is comparable to the differences detected between other group 1 genes which are known not to be alleles: e.g. over the same cloned region, BS-6 and BS-5 (BALB/c clones that share a similar 3' flanking region structure to BL-1 and CL-8) differ from each other by two 6-base restriction sites and two MspI sites. This coupled with the facts that (1) the identical clones CL-6 and CL-13 show no differences in their commonly cloned regions from BL-1, and (2) the fragment unique to MUPC2 is represented ∿equivalently in C57BL/Fa and BALB/c genomic DNAs, strongly suggests that CL-8 is not an allele of BL-1.

BL-7 is most closely related to CL-1 and CL-3 and differs from these

clones by two MspI sites. CL-1 and CL-3 show no restriction site differences from BS-6 and BL-14 over their commonly cloned regions, suggesting that they are allelic clones of BS-6 and BL-14. However the possibility exists that BL-7 is an allele of CL-3 and/or CL-1. This is because the MspI fragment unique to BL-7 among the cloned genes is labelled 3 to 4 times more heavily in the BALB/c genomic DNA than in the C57BL/Fa genomic DNA, and because MspI site polymorphisms are known to be high in mammalian DNA (Barker et al 1984). CL-11 is unlikely to be an allele of any of the MUP genes previously isolated because its relationship to other group 1 genes was variable in the phylogenies drawn by the Dollo Parsimony method and because it differs from these genes by a relatively large number of restriction sites. Also the restriction fragments unique to CL-11 were present in both BALB/c and C57BL/Fa genomic DNAs.

MUPC3 and MUPC4 represent BALB/c and C57 clones respectively that differ in their restriction maps by a single 6-base restriction site. MUPC3 and MUPC4 share a common 3' flanking sequence arrangement not present in other group 1 genes isolated. MUPC3 appears to be represented by a single gene in BALB/c (see Section 7 of the results), and it is therefore likely that MUPC3 and MUPC4 are allelic clones. The Dollo Parsimony phylogenies suggest that the restriction polymorphism between MUPC3 and MUPC4 may have resulted from gene conversion of MUPC4 by CL-11 or another gene that shares the polymorphic PstI site, although other explanations are equally likely.

The considerable divergence of the group 3 genes would suggest that

BL-2 and CL-12 may be alleles. However a 1 kbp PstI restriction fragment present in BL-2 but not CL-12 was observed in both BALB/c and C57BL/Fa genomic DNAs. It is therefore impossible at this point to determine whether BL-2 and CL-12 represent allelic genes.

The partially cloned group 1 genes BS-1Ø9/1 and BS-1Ø2/1 show minor restriction site differences from each other and also from other BALB/c group 1 genes. Other differences are found in the sequenced region 5' to the BamHI site in intron 1 (P.Ghazal, unpublished results). BS-1Ø9/1 and BS-1Ø2/1 are not alleles of CL-8/CL-9, MUPC4 or CL-11 since their restriction maps show that they both belong to the sub-group with a large 5' HindIII fragment.

The sequences of most MUP cDNAs are found to differ from each other and from those of sequenced MUP genes. Reverse transcriptase misincorporates dCMP residues at a frequency of ∿Ø.Ø7% when Poly d[A-T] is used as a template (Sirover and Loebb, 1977). Because of the high infidelity of reverse transcriptase in vitro, it is possible that some of the differences observed in the sequences of MUP cDNAs are due to errors introduced during cloning. The sequence divergence between group 1 cDNAs is the same as that found between group 1 genes. Therefore, given the multigene nature of the MUPs and the presence of ∿2Ø different MUP mRNAs in liver cells it is quite likely that the differences are genuine.

Of the cDNA clones isolated only two (MUP8 and LVA325) are identical to a group 1 gene (BL-1) where their sequences overlap. The cDNA clones p1Ø57 and MUP11 however could represent transcripts

**Table .D.7.** Postulated relationships between BALB/c and C57 cloned MUP sequences

| | BALB/c allelic clone | C57 allelic clone |
|---|---|---|
| (1) | BS-6,BL-14 | CL-1,CL-3 |
| (2) | BL-7 | CL-1(?),CL-3(?) |
| (3) | BS-5 | not cloned |
| (4) | BS-1,BS-1Ø7 | CL-5,CL-1Ø |
| (5) | BL-1 | CL6/CL-13 |
| (6) | BS-1Ø9/1 | CL-1(?),CL-3(?) |
| (7) | BS-1Ø2/1 | CL-1(?),CL-3(?) |
| (8) | not cloned | CL-8 |
| (9) | not cloned | CL-11 |
| (1Ø) | BS-2,BS-3,BS-4 | not cloned |
| (11) | BS-1Ø2/2 | not cloned |
| (12) | BS-1Ø2/2 | not cloned |
| (13) | BL-15 | not cloned |
| (14) | BL-25 | CL-2 |
| (15) | BL-8 | CL-4 |
| (16) | BL-2 | CL-12(?) |
| (17) | MUP8(cDNA) | CL-6/CL-13(?) |
| (18) | LVA325(cDNA) | CL-6/CL-13(?) |
| (19) | MUP11(cDNA) | ? |
| (2Ø) | p1Ø57(cDNA) | ? |
| (21) | not cloned | p499(cDNA) |
| (22) | MUP15(cDNA) | p199(cDNA) |

of unsequenced group 1 genes. The C57BL/6J liver cDNA clone p499 is not a transcript of CL-8/CL-9, CL-5, CL-1Ø or CL-11 since it lacks the EcoRI site in exon 2. Due to the contamination of the C57(BC15) mouse strain, the possibility that p499 represents the transcript of an allele of one of these cloned genes cannot be completely excluded however. It is unlikely to represent a transcript of MUPC2 since these are probably alleles of BL-1 and as mentioned previously p499 is no more homologous to BL-1 than non-allelic group 1 BALB/c genes are to each other. It therefore appears that p499 represents a transcript of an uncloned gene.

The postulated relationships between the strains of the isolated MUP genomic clones and cDNA clones are summarised in Table D.7. If we assume that the relationships are correct, then it appears that a minimum of 1Ø different group 1 sequences have been identified. This corresponds to a large proportion of the estimated number of group 1 genes (∿ 15) and implies that the majority of Mus musculus musculus group 1 genes are very similar in structure.

Finlayson et al (1958; 1963) originally resolved the urinary proteins of inbred mouse strains into three components: MUP1, MUP2 and MUP3. Different inbred strains were found to have one of two patterns: MUP1 + MUP3 (e.g. BALB/c mice) or MUP2 + MUP3 (e.g. C57BL mice). To a large extent mouse strains that are more closely related to each other have a similar MUP pattern. An exception is C57BR which has the urinary pattern MUP1 + MUP3 unlike other C57 mice. Szoka and Paigen (1978) followed up the work of Finlayson et al (1958; 1963) and showed that traces of MUP1 and MUP2 are present in

all strains studied. The MUP1 and MUP2 components are known to represent the products of more than one gene since two dimensional isoelectric focusing of the urinary MUPs of C57BL/6J mice gives at least 7 distinguishable gene products (Kuhn et al, 1984) and 20 MUP precursor proteins are translated from liver mRNA of both BALB/c and C57BL/Fa mice (Clissold and Bishop, 1982).

Comparison of partial amino acid sequences of the N-terminal regions of components MUP1 and MUP2 showed that these are very similar and that they are transcripts of group 1 genes. Although it is conceivable that MUP1 and MUP2 have diverged in their C-termini, this seems unlikely for the following reasons: (1) clones which are homologous to group 1 genes in their N-termini and which diverge from group 1 genes in their C-termini have not been isolated; (2) all cDNA clones isolated from BALB/c and C57BL/6J liver libraries either represent group 1 genes or the MUP15 gene; (3) no differences have been found in the relative hybridizations of BALB/c and C57BL/Fa mRNAs to 3' exonic probes (LVA325, BS-6-5-5) when washed at low and high stringency conditions (P.Clissold, unpublished results).

With possibly one exception (JU), all inbred strains examined express MUP3 (Hudson et al, 1967; Szoka and Paigen 1978, 1979), and are therefore likely to carry active MUP15 or MUP15-like sequences in their genomes. Polyacrylamide gel electrophoresis of the urinary proteins of JU (which based on restriction enzyme analysis has the genotype $Mup\text{-}1^{c}$; Bishop et al, 1982) only reveals components that contribute to MUP1 (see Hainey and

Bishop, 1982). Iso-electric focusing, however, reveals minor bands that co-migrate with the MUP3 components of BALB/c (genotype Mup-$1^c$). In view of the observed difference in MUP3 expression, it would be interesting to determine whether MUP15-like sequences in JU differ from those of other inbred mouse strains.

To summarize, the major differences in the urinary proteins of inbred mouse strains appear to be due to differences in the expression of closely related group 1 genes. Future studies on the differences in expression between inbred strains will require the identification of the genes that contribute to MUP1 and MUP2.

Finally, it is hoped that expression of the isolated MUP genes in tissue culture will identify variant MUP genes between BALB/c and C57BL/Fa. The introduction of variant MUP genes and mutated versions of these genes into mice may help us identify some of the sequences that contribute to the hormonal regulation and tissue-specific expression of mammalian genes. That such an approach is feasible has now been successfully demonstrated by the tissue-specific expression of the rat pancreatic elastase 1 gene in transgenic mice (Swift et al, 1984) and by the expression of a functionally re-arranged heavy chain immunoglobulin gene in lymphoid tissues of transgenic mice (Grosschdel et al, 1984).

REFERENCES

Alt, F.W., Kellems, R.E., Bertino, J.R. and Schimke, R.T. (1978). Selective multiplication of dihydrofolate reductase genes in methotrexate-resistant variants of cultured murine cells. J.Biol.Chem. 253, 1357-1370.

Alt, F.W., Rosenberg, N., Casanova, P.J., Thomas, E. and Baltimore, D. (1982). Immunoglobulin heavy-chain expression and class switching in a murine leukaemia cell line. Nature 296, 325-331.

Anson, D.S. (1983). Ph.D. Thesis, University of Edinburgh.

Antoine, M. and Niessing, J. (1984). Intron-less globin genes in the insect Chironomus thummi thummi. Nature 310, 795-798.

Aviv, H. and Leder, P. (1972). Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid cellulose. Proc.Natl.Acad.Sci. USA, 69, 1408-1412.

Baltimore, D. (1981). Gene Conversion: Some implications for immuno-globulin genes. Cell 24, 592-594.

Banerji, J., Rusconi, S. and Schaffner, W. (1981). Expression of a β- globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299-308.

Banerji, J., Olson, L. and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell 33, 729- 740.

Barker, D., Schafer, M. and White, R. (1984). Restriction sites containing CpG show a higher frequency of polymorphism in human DNA. Cell 36, 131-138.

Barnett, T., Pachl, C., Gergen, J.P. and Wensink, P.C. (1980). The isolation and characterization of Drosophila yolk protein genes. Cell 21, 729-738.

Barth, R., Gross, K., Gremke, L. and Hastie, N. (1982). Developmentally regulated mRNAs in mouse liver. Proc.Natl.Acad.Sci. USA 79, 500-504.

Bennett, K., Lalley, P., Barth, R. and Hastie, N. (1982). Mapping the structural genes coding for the major urinary proteins in the mouse: combined used of recombinant inbred strains and somatic cell hybrids. Proc.Natl.Acad.Sci. USA 79, 1220-1224.

Benoist, C. and Chambon, P. (1980). In vivo sequence requirements of the SV40 early promoter region. Nature 290, 304-310.

Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980). The ovalbumin gene - sequence of putative control regions. Nucleic Acids Res. 8, 127-142.

Bentley, D.L. and Rabbitts, T.H. (1983). Evolution of immunoglobulin V genes: Evidence indicating that recently duplicated human $V_K$ sequences have diverged by gene conversion. Cell 32, 181-189.

Benton, W.D. and Davis, R.W. (1977). Screening λgt recombinant clones by hybridization to single plaques in situ. Science 196, 180-182.

Berger, F.G. and Szoka, P. (1981). Biosynthesis of the major urinary proteins in mouse liver: a biochemical genetic study. Biochem. Genet. 19, 1261-1273.

Berger, F.G. (1983). Studies on genetic variation in major urinary protein synthesis in mouse liver. Biochem.Genet. 21, 15-

23.

Berget, S.M. (1984). Are U4 small nuclear ribonucleoproteins involved in polyadenylation Nature 309, 179-181.

Birchmeier, C., Grosschedl, R. and Birnstiel, M.L. (1982). Generation of authentic 3' termini of an H2A mRNA in vivo is dependent on a short inverted DNA repeat and on spacer sequences. Cell 28, 739-745.

Bishop, J.O. (1979). A DNA sequence cleaved by restriction endonuclease R. EcoRI in only one strand. J.Mol.Biol. 128, 545-559.

Bishop, J.O. and Davies, J.A. (1980). Plasmid cloning vectors that can be nicked at a unique site. Molec.Gen.Genet. 179, 573-580.

Bishop, J.O., Clark, A.J., Clissold, P.M., Hainey, S. and Francke, U. (1982). Two main groups of mouse major urinary protein genes, both largely located on chromosome 4. EMBO J. 1, 615- 620.

Blake, C.C.F. (1979). Exons encode protein functional units. Nature 277, 598.

Brady, J., Bolen, J.B., Radonovich, M., Salzman, N. and Khoury, G. (1984). Stimulation of simian virus 40 late gene expression by simian virus 40 tumor antigen. Proc.Natl.Acad.Sci. USA 81, 2040-2044.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978). Ovalbumin gene: Evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc.Natl.Acad. Sci. 75, 4853-4857.

Breathnach, R. and Chambon, P. (1981). Organization and expression

of eucaryotic split genes coding for proteins. Annual Review of Biochemistry 50, 349-383.

Bunick, D., Zandomeni, R., Ackeman, S. and Weinmann, R. (1982). Mechanism of RNA polymerase II-specific initiation of transcription in vitro: ATP requirement and uncapped runoff transcripts. Cell 29, 877-886.

Burch, J.B.E. and Weintraub, H. (1983). Temporal order of chromatin structural changes associated with activation of the major chicken vitellogenin gene. Cell 33, 65-76.

Busslinger, M., Moschonas, M. and Flavell, P.A. (1981). $\beta^{+}$ thalassemia: Aberrant splicing results from a single point mutation in an intron. Cell 27, 289-298.

Busslinger, M., Hurst, J. and Flavell, R.A. (1983). DNA methylation and the regulation of globin gene expression. Cell 34, 197-206.

Cameron, J.R., Philippsen, P. and Davis, R.W. (1977). Analysis of chromosomal integration and deletions of yeast plasmids. Nucleic Acids Res. 4, 1429-1448.

Canaani, D. and Berg, P. (1982). Regulated expression of human interferon $\beta_1$ gene after transduction into cultured mouse and rabbit cells. Proc.Natl.Acad.Sci. USA 79, 5166-5170.

Chandler, V.L., Maler, B.A. and Yamamoto, K.R. (1983). DNA sequences bound specifically by glucocorticoid receptor in vitro render a heterologous promoter hormone responsive in vivo. Cell 33, 489-499.

Chao, M.V., Mellon, P., Charnay, P., Maniatis, T. and Axel, R. (1983). The regulated expression of β-globin genes introduced into mouse erythroleukemia cells. Cell 32, 483-493.

Charnay, P., Treisman, R., Mellon, P., Chao, M., Axel, R. and Maniatis, T. (1984). Differences in human α- and β- globin gene expression in in mouse erythroleukemia cells: the role of intragenic sequences. Cell 38, 251-263.

Childs, G., Maxon, R., Cohn, R.H. and Kedes, L. (1981). Orphons: dispersed genetic elements derived from tandem repetitive genes of eucaryotes. Cell 23, 651-663.

Chu, M.L., de Wet, W., Bernard, M., Ding, J.F., Morabito, M., Myers, J. Williams, C. and Ramirez, F. (1984). Human pro α1(I) collagen gene structure reveals evolutionary conservation of a pattern of introns and exons. Nature 310, 337-340.

Church, G.M. and Gilbert, W. (1984). Genomic Sequencing. Proc.Natl.Acad. Sci. USA 81, 1991-1995.

Clark, J.L. and Steiner, D. (1969). Insulin biosynthesis in the rat: demonstration of two proinsulins. Proc.Natl.Acad.Sci. USA 62, 278-285.

Clark, A.J., Clissold, P.M. and Bishop, J.O. (1982). Variation between mouse major urinary protein genes isolated from a single inbred line. Gene 18, 221-230.

Clark, A.J., Clissold, P.M., Al-Shawi, R., Beattie, P. and Bishop, J. (1984a). Structure of mouse major urinary protein genes: different splicing configurations in the 3'- non-coding region. EMBO J. 3, 1045-1052.

Clark, A.J., Hickman, J. and Bishop, J. (1984b). A 45kb DNA domain with two divergently orientated genes is the unit of organisation of the murine major urinary protein genes. EMBO J. 3, 2055-2064.

Clarke, L. and Carbon, J. (1976). A colony bank containing synthetic Col E1 hybrid plasmids representative of the entire E. coli genome. Cell 9, 91-99.

Clissold, P.M. and Bishop, J.O. (1981). Molecular cloning of cDNA sequences transcribed from mouse liver endoplasmic reticulum mRNA. Gene 15, 225-235.

Clissold, P.M. and Bishop, J.O. (1982). Variation in mouse major urinary protein (MUP) genes and the MUP gene products within and between inbred lines. Gene 18, 211-220.

Clissold, P.M., Hainey, S. and Bishop, J.O. (1984). Messenger RNAs coding for mouse major urinary proteins are differentially induced by testosterone. Biochem.Genet. 22, 379-387.

Compere, S.J. and Palmiter, R.D. (1981). DNA methylation controls the inducibility of the mouse metallothionein-1 gene in lymphoid cells. Cell 25, 233-240.

Coulson, A. and Winter, G. (1982). Chain terminator sequencing course manual. MRC Centre, Cambridge.

Cowan, N.J., Wilde, C.D., Chow, L.T. and Wefald, F.C. (1981). Structural variation among human β-tubulin genes. Proc.Natl.Acad.Sci.USA 78, 4877-4881.

Craik, C.S., Buchman, S.R. and Beychok, S. (1980). Characterization of globin domains: heme binding to the central exon product. Proc.Natl.Acad.Sci. USA 77, 1384- 1388.

Derman, E. (1981). Isolation of a cDNA clone for mouse urinary proteins: Age- and sex-related expression of mouse urinary protein genes is transcriptionally controlled. Proc.Natl.Acad.Sci. USA 78, 5425-5429.

de Villiers, J., Olson, L., Tyndall, C. and Schaffner, W. (1982).

Transcriptional "enhancers" from SV40 and polyoma virus show a cell type preference. Nucleic Acids Res. 10, 7965-7976.

Dierks, P., van Voyen, A., Cochran, M.D., Dobkin, C., Reiser, J. and Weissmann, C. (1983). Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit β- globin gene in mouse 3T6 cells. Cell 32, 695-706.

Di Maio, D. and Maniatis, T. (1982). Intact bovine papilloma virus -human DNA recombinant plasmids that propogate as episomes in mouse cells and bacteria. In Eukaryotic Viral Vectors, ed. Gluzman, Y. Cold Spring Harbor Laboratory.

Dodgson, J.B., Strommer, J. and Engel, J.D. (1979). Isolation of the chicken β-globin gene and a linked embryonic β-like globin gene from a chicken DNA recombinant library. Cell 17, 879-887.

Dodgson, J.B., McCune, K.C., Rusling, D.J., Krust, A. and Engel, J.D. (1981). Adult chicken α-globin genes $\alpha^{A}$ and $\alpha^{D}$: no anemic shock α-globin exists in domestic chickens. Proc.Natl.Acad.Sci. USA 78, 5998-6002.

Dolan, M., Sugarman, B.J., Dodgson, J.B. and Engel, J.D. (1981). Chromosomal arrangement of the chicken β-type globin genes. Cell 24, 669-677.

Dolan, K.P., Unterman, R., McLaughin, M., Nakhasi, H.L., Lynch, K.R. and Feigelson, P. (1982). The structure and expression of very closely related members of the $\alpha_{2u}$ globulin gene family. J. Biol. Chem. 257, 13527-13534.

Dudov, K.P. and Perry, R.P. (1984). The gene family encoding the mouse ribosomal protein L32 contains a uniquely expressed intron- containing gene and an unmutated unprocessed gene. Cell 37, 457-468.

Dynan, W.S. and Tjian, R. (1983). Isolation of transcription factors that discriminate between different promoters recognized by RNA polymerase II. Cell 32, 669-680.

Early, P.W., Davis, M.M., Kaback, D.B., Davidson, N. and Hood, L.(1979). Immunoglobulin heavy chain gene organization in mice: analysis of a myeloma genomic clone containing variable and α constant regions. Proc.Natl.Acad.Sci. USA 76, 857-861.

Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., De Riel, J.K., Forget, B.G., Weissman, S.M., Slighton, J.L., Blechl, A.E., Smithies, O., Baralle, F.E.., Shoulders, C.C. and Proudfoot, N.J. (1980). The structure and evolution of the human β-globin gene family. Cell 21, 653-668.

Eickbush, T.H. and Kafatos, F.C. (1982). A walk in the chorion locus of Bombyx mori. Cell 29, 633-643.

Engel, J.D., Sugarman, B.J. and Dodgson, J.B. (1982). A chicken histone H3 gene contains intervening sequences. Nature 297, 434-436.

Farris, J.S. (1977). Phylogenetic analysis under Dollo's Law. Syst.Zool. 26, 77-88.

Faulds, D., Dower, N., Stahl, M.M. and Stahl, F.W. (1979). Orientation-dependent recombination hotspot activity in bacteriophage lambda. J.Mol.Biol. 131, 681-695.

Felsenstein, J. (1982). Numerical methods for inferring evolutionary trees. The Quarterly Review of Biology 57, 379-404.

Felsenstein, J. (1983). Parsimony in systematics: Biological and

Statistical issues. Ann.Rev.Syst. 14, 313-333.

Fink, G.R. and Styles, C.A. (1974). Gene conversion of deletions in the HIS4 region of yeast. Genetics 77, 231-244.

Finlayson, J.S. and Baumann, C.A. (1958). Mouse proteinuria. American Journal of Physiology 192, 69-72.

Finlayson, J.S., Asofsky, R., Potter, M. and Runner, C.C. (1965). Major urinary protein complex of normal mice: origin. Science 149, 981-982.

Finlayson, J.S., Hudson, D.M. and Armstrong, B.L. (1969). Location of the Mup-a locus on mouse linkage group VIII. Genet.Res.,Camb. 14, 329-331.

Finlayson, J.S., Potter, M., Shinnick, C.C. and Smithies, O. (1974). Components of the major urinary complex of inbred mice: determination of $NH_2$-terminal sequences and comparison with homologous components from wild mice. Biochemical Genetics 11, 325-334.

Fitzgerald, M. and Shenk, T. (1981). The sequence 5'- AAUAAA-3' forms part of the recognition site for polyadenylation of late SV40 mRNAs. Cell 24, 251-260.

Fradin, A., Manley, J.L. and Prires, C.L. (1982). Methylation of simian virus 40 HpaII-site affects late, but not early viral gene expression. Proc.Natl.Acad.Sci. USA 79, 5142-5146.

Frischauf, A.-M., Lehrach, H., Poustka, A. and Murray, N. (1983). Lambda replacement vectors carrying polylinker sequences. J.Mol.Biol. 170, 827-842.

Garabedian, M.J., Hung, M-C. and Wensink, P.C. (1985). Independent control elements that determine yolk protein gene expression

in alternative Drosophila tissues. Proc.Natl.Acad.Sci. USA, 82, 1396-1400.

Gardner, R.C., Howarth, A.J., Messing, J. and Shepherd, R.J. (1982). Cloning and sequencing of restriction fragments generated by EcoRI*. DNA 1, 109-115.

Geisse, S., Scheidereit, C., Westphal, H.M., Hynes, N.E., Groner, B. and Beato, M. (1982). Glucocorticoid receptors recognize DNA sequences in and around murine mammary tumor virus DNA. EMBO J. 1, 1613-1619.

Gerber-Hyber, S., May, F.E.B., Westley, B.R., Felber, B.K., Hosbach, H.A., Andres, A-C. and Ryffel, G.U. (1983). In contrast to other Xenopus genes the estrogen-inducible vitellogenin genes are expressed when totally methylated. Cell 33, 43-51.

Ghazal, P., Clark, A.J. and Bishop, J.O. (1985). Evolutionary amplification of a pseudogene. Proc.Natl.Acad.Sci. USA. In press.

Gidoni, D., Dynan, W.S. and Tjian, R. (1984). Multiple specific contacts between a mammalian transcription factor and its cognate promoters. Nature 312, 409-413.

Gil, A. and Proudfoot, N.J. (1984). A sequence downstream of AAUAAA is required for rabbit β-globin mRNA 3'-end formation. Nature 312, 473-474.

Gilbert, W. (1978). Why genes in pieces. Nature 271, 501.

Gluzman, Y. Editor, (1982). Eukaryotic viral vectors. Cold Spring Harbor Laboratory.

Gluzman, Y. and Shenk, T. Editors, (1983). Enhancers and eukaryotic gene expression. Cold Spring Harbor Laboratory.

Goeddel, D.V., Leung, D.W., Dull, T.J., Gross, M., Lawn, R.M., McCandliss, R., Seeburg, P.H., Ullrich, A., Yelverton, E. and Gray, P.W. (1981). The structure of eight distinct cloned human leukocyte interferon cDNAs. Nature 290, 20-26.

Goodman, H.M., Greene, P.J., Garfin, D.E. and Boyer, H.W. (1977). In Nucleic Acid-Protein Recognition. H.J. Vogel, Ed. (Academic Press, New York) 239-259.

Grabowski, P.J., Padgett, R.A., Sharp, P.A. (1984). Messenger RNA splicing in vitro: an excised intervening sequence and a potential intermediate. Cell 37, 415-425.

Grosschedl, R. and Birnstiel, M.L. (1980). Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. Proc.Natl. Acad.Sci. USA 77, 1432-1436.

Grosschedl, R., Weaver, D., Baltimore, D. and Costantini, F. (1984). Introduction of a μ immunoglobulin gene into the mouse germ line: specific expression in lymphoid cells and synthesis of functional antibody. Cell 38, 647-658.

Grosveld, F.G., Dahl, H-H.M., de Boer, E. and Flavell, R.A. (1981). Isolation of β-globin related genes from a human cosmid library. Gene 13, 227-237.

Grosveld, G.C., de Boer, E., Shewmaker, C.K. and Flavell, R.A. (1982). DNA sequences necessary for transcription of the rabbit β- globin gene in vivo. Nature 295, 120-126.

Groudine, M. and Weintraub, H. (1982). Propogation of globin DNase I hypersensitive sites in absence of factors required for induction: a possible mechanism for determination. Cell 30, 131-139.

Günzburg, W.H. and Groner, B. (1984). The chromosomal integration site determines the tissue-specific methylation of mouse mammary tumour virus proviral genes. EMBO J. 3, 1129-1135.

Hagenbüchle, O., Tosi, M., Schibler, U., Bovey, R., Wellaner, P.K. and Young, R.A. (1981). Mouse liver and salivary gland α- amylase mRNAs differ only in 5' non-translated sequences. Nature 289, 643-646.

Hagenbüchle, O., Wellaner, P.K., Cribbs, D.L. and Schibler, U. (1984). Termination of transcription in the mouse α-amylase gene Amy-2$^{a}$ occurs at multiple sites downstream of the polyadenylation site. Cell 38, 737-744.

Haigh, L.S., Owens, B.B., Hellewell, S. and Ingram, V.M. (1982). DNA methylation in chicken α-globin gene expression. Proc.Natl.Acad.Sci. USA 79, 5332-5336.

Hainey, S. and Bishop, J.O. (1982). Allelic variation at several different genetic loci determines the major urinary protein phenotype of inbred mouse strains. Genet.Res.Camb. 39, 31-39.

Harvey, R.P., Whiting, J.A., Coles, L.S., Krieg, P.A. and Wells, J.R.E. (1983). H2A.F: an extremely variant histone H2A sequence expressed in the chicken embryo. Proc.Natl.Acad.Sci. USA 80, 2819- 2823.

Hastie, N.D., Held, W.A. and Toole, J.J. (1979). Multiple genes coding for the androgen-regulated major urinary proteins of the mouse. Cell 17, 449-457.

Heilig, R., Perrin, F., Gannon, F., Mandel, J.L. and Chambon, P. (1980). The ovalbumin gene family: structure of the X gene and evolution of duplicated split genes. Cell 20, 625-637.

Henikoff, S., Kelly, J.D. and Cohen, E.H. (1983). Transcription terminates in yeast distal to a control sequence. Cell 33, 607- 614.

Hentschel, C.C. and Birnstiel, M.L. (1981). The organization and expression of Histone Gene families. Cell 25, 301-313.

Hofer, E., Hofer-Warbinek, R. and Darnell, J.E.,Jr. (1982). Globin RNA transcription: a possible termination site and demonstration of transcriptional control correlated with altered chromatin structure. Cell 29, 887-893.

Hoffman, H.A. (1970). Starch-gel electrophoresis of murine major urinary protein. Proceedings of the Society for Experimental Biology and Medicine. 135, 81-83.

Hohn, B. and Murray, K. (1977). Packaging recombinant DNA molecules into bacteriophage particles in vitro. Proc.Natl.Acad.Sci. USA 74, 3259-3263.

Hollis, G.F., Hieter, P.A., McBride, O.W., Swan, D. and Leder, P. (1982). Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. Nature 296, 321- 325.

Hosbach, H.A., Wyler, T. and Weber, R. (1983). The Xenopus laevis globin gene family: chromosomal arrangement and gene structure. Cell 32, 45-53.

Hudson, D.M., Finlayson, J.S. and Potter, M. (1967). Linkage of one component of the major urinary protein complex of mice to the brown coat colour locus. Genet.Res.,Camb. 10, 195-198.

Hynes, N., van Ooyen, A.J.J., Kennedy, N., Herrlich, P., Ponta, H. and Groner, B. (1983). Subfragments of the large terminal repeat cause glucocorticoid-responsive expression of mouse

mammary tumor virus and of an adjacent gene. Proc.Natl.Acad. Sci. USA 80, 3637-3641.

Iatrou, K. and Tsitilou, S.G. (1983). Coordinately expressed chorion genes in Bombyx mori: is developmental specificity determined by secondary structure recognition EMBO J. 2, 1431- 1440.

Jackson, J.A. and Fink, G.R. (1981). Gene conversion between duplicated genetic elements in yeast. Nature 292, 306-311.

Jeffreys, A.J., Wilson, V., Wood, D., Simons, J.P., Kay, R.M. and Williams, J.G. (1980). Linkage of adult α- and β- globin genes in Xenopus laevis and gene duplication by tetraploidization. Cell 21, 555-564.

Jeffreys, A.J. (1982). Evolution of globin genes. In Genome Evolution, Dover, G.A. and Flavell, R.D. Eds. (Academic Press) 157-176.

Jones, C.W. and Kafatos, F.C. (1980). Coordinately expressed members of two chorion multi-gene families are clustered, alternating and divergently orientated. Nature 284, 635-638.

Jones, C.W. and Kafatos, F.C. (1980). Structure, organization and evolution of developmentally regulated chorion genes in a silkmoth. Cell 22, 855-867.

Karin, M. and Richards, R.I. (1982). Human metallothionein genes - primary structure of the metallothionein II gene and a related processed gene. Nature 299, 797-802.

Karin, M., Haslinger, A., Holtgreve, H., Richards, R.I., Krauter, P., Westphal, H.M. and Beato, M. (1984). Characterization of DNA sequences through which cadmium and glucocorticoid hormones induce human metallothionein-$II_A$ gene. Nature

308, 513-519.

Keller, E.B. and Noon, W.A. (1984). Intron splicing: A conserved internal signal in introns of animal pre-mRNAs. Proc.Natl.Acad. Sci. USA 81, 7417-7420.

Kemp, D.J., Cory, S. and Adams, J.M. (1979). Cloned pairs of variable region genes for immunoglobulin heavy chains isolated from a clone library of the entire mouse genome. Proc.Natl.Acad.Sci., USA 76, 4627-4631.

Killary, A.M. and Faurnier, R.E.K. (1984). A genetic analysis of extinction: trans-dominant loci regulate expression of liver-specific traits in hepatoma hybrid cells. Cell 38, 523-534.

King, C.R. and Piatigorsky, J. (1983). Alternative RNA splicing of the murine αA-crystallin gene: Protein coding information within an intron. Cell 32, 707-712.

Klar, A.J.S. and Strathern, J.N. (1984). Resolution of recombination intermediates generated during yeast mating type switching. Nature 310, 744-748.

Klein, H.L. (1984). Lack of association between intrachromosomal gene conversion and reciprocal exchange. Nature 310, 748-753.

Klein, H.L. and Petes, T.D. (1981). Intrachromosomal gene conversion in yeast. Nature 289, 144-148.

Knopf, J.L., Gallagher, J.F. and Held, W.A. (1983). Differential multi-hormonal regulation of the mouse urinary protein gene family in the liver. Mol.Cell Biol. 3, 2232-2240.

Kobayashi, I., Murialdo, H., Crasemann, J.M., Stahl, M.M. and Stahl, F.W. (1982). Orientation of cohesive end site cos

determines the active orientation of χ sequence in stimulating recA recBC-mediated recombination in phage λ lytic infections. Proc.Natl.Acad.Sci. USA 79, 5981-5985.

Konarska, M.M., Padgett, R.A. and Sharp, P.A. (1984). Recognition of cap structure in splicing in vitro of mRNA precursors. Cell 38, 731-736.

Konkel, D.A., Maizel, J.V. and Leder, P. (1979). The evolution and sequence comparison of two recently diverged mouse chromosomal β- globin genes. Cell 18, 865-873.

Krainer, A.R., Maniatis, T., Ruskin, B. and Green, M. (1984). Normal and mutant human β-globin pre-mRNAs are faithfully and efficiently spliced in vitro. Cell 36, 993-1005.

Krauter, K., Leinwand, L., D'Eustachio, P., Ruddle, F. and Darnell, J.E.Jr. (1982). Structural genes of the mouse major urinary protein are on chromosome 4. J.Cell.Biol. 94, 414-417.

Kuhn, N.J., Woodworth-Gutai, M., Gross, K.W. and Held, W.A. (1984). Subfamilies of the mouse major urinary protein (MUP) multigene family: sequence analysis of cDNA clones and differential regulation in the liver. Nucleic Acids Res. 12, 6073-6090.

Kurtz, D.T. (1981a). Rat $\alpha_{2u}$ -globulin is encoded by a multigene family. Journal of Molecular and Applied Genetics 1, 29-38.

Kurtz, D.T. (1981b). Hormonal inducibility of rat $\alpha_{2u}$ - globulin genes in transfected mouse cells. Nature 291, 629- 631.

Lacy, E. and Maniatis, T. (1980). The nucleotide sequence of a rabbit β-globin pseudogene. Cell 21, 545-553.

Laimins, L.A., Khoury, G., Gorman, C., Howard, B. and Gruss, P.

(1982). Host-specific activation of transcription by tandem repeats from simian virus 40 and Moloney murine sarcoma virus. Proc. Natl.Acad.Sci. USA 79, 6453-6457.

Laimins, L.A., Kessel, M., Rosenthal, N. and Khoury, G. (1983). Viral and cellular enhancer elements. In Enhancers and Eukaryotic Gene Expression. Gluzman, Y. and Shenk, T. Eds. (Cold Spring Harbor Laboratory) 28-37.

Langford, C.J., Klinz, F-J., Donath, C. and Gallwitz, D. (1984). Point mutations identify the conserved, intron-contained TACTAAC box as an essential splicing signal sequence in yeast. Cell 36, 645-653.

Laperche, Y., Lynch, K.R., Dolan, K.P. and Feigelson, P. (1983). Tissue-specific control of $\alpha_{2u}$-globulin gene expression: constitutive synthesis in the submaxillary gland. Cell 32, 453-460.

Lawson, G.M., Knoll, B.J., March, C.J., Woo, S.L.V., Tsai, M- J. and O'Malley, B.W. (1982). Definition of 5' and 3' structural boundaries of the chromatin domain containing the ovalbumin multigene family. J.Biol.Chem. 257, 1501-1507.

Leder, P., Hanson, N.J., Konkel, D., Leder, A., Nishioka, Y.N. and Talkington, C. (1980). Mouse globin system: a functional and evolutionary analysis. Science 209, 1336-1342.

Lee, F., Mulligan, R., Berg, P. and Ringold, G. (1981). Glucocorticoids regulate expression of dihydrofolate reductase cDNA in mouse mammary tumor virus chimeric plasmids. Nature 294, 228-232.

Lee, M.G-S., Lewis, S.A., Wilde, C.D. and Cowan, N.J. (1983). Evolutionary history of a multigene family: an expressed human β - tubulin gene and three processed pseudogenes. Cell 33, 477-487.

Leigh Brown, A.J. and Ish-Horowicz, D. (1981). Evolution of the 87A and 87C heat-shock loci in Drosophila. Nature 290, 677-682.

Le Meur, M.A. Galliot, B. and Gerlinger, P. (1984). Termination of the ovalbumin gene transcription. EMBO J. 3, 2779-2786.

Lerner, M.R. and Steitz, J.A. (1979). Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. Proc.Natl.Acad.Sci. USA 76, 5495-5499.

Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L. and Steitz, J. (1980). Are snRNPs involved in splicing Nature 283, 220-224.

Lewin, B. (1983). Genes. John Wiley and Sons Inc.

Liebhaber, S.A., Goossens, M. and Kan, Y.W. (1981). Homology and concerted evolution at the α1 and α2 loci of human α-globin. Nature 290, 26-29.

Loenen, W.A.M. and Brammar, W.J. (1980). A bacteriophage lambda vector for cloning large DNA fragments made with several restriction enzymes. Gene 10, 249-259.

Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. and Tizard, R. (1979). The structure and evolution of the two nonallelic rat preproinsulin genes. Cell 18, 545-558.

Lynch, K.R., Dolan, K.P., Nakhasi, H.L., Unterman, R. and Feigelson, P. (1982). The role of growth hormone in $\alpha_{2u}$ globulin synthesis: a reexamination. Cell 28, 185-189.

Mainwaring, W.I.P. (1980). Steroid receptors. In Cellular

Receptors for Hormones and Neurotransmitters, ed. Schulster, D. and Levitzki, A. John Wiley and Sons Ltd.

Maki, R., Traunecker, A., Sakano, H., Roeder, W. and Tonegawa, S. (1980). Exon shuffling generates an immunoglobulin heavy chain gene. Proc.Natl.Acad.Sci. USA 77, 2138-2142.

Maki, R., Roeder, W., Traunecker, A., Sidman, C., Wabl, M., Raschke, W. and Tonegawa, S. (1981). The role of DNA rearrangement and alternative RNA processing in the expression of immunoglobulin δ genes. Cell 24, 353-365.

Maniatis, T., Hardison, R.C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G.K. and Efstratiadis, A. (1978). The isolation of structural genes from libraries of eukaryotic DNA. Cell 15, 687- 701.

Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982). Molecular cloning: A laboratory manual. Cold Spring Harbor Laboratory.

Manley, J.L., Fire, A., Cano, A., Sharp, P.A. and Gefter, M.L. (1980). DNA dependent transcription of adenovirus genes in a soluble whole-cell extract. Proc.Natl.Acad.Sci. USA 77, 3855- 3859.

Manley, J.L. (1983). Accurate and specific polyadenlyation of mRNA precursors in a soluble whole cell lysate. Cell 33, 595-603.

Mathis, D.J. and Chambon, P. (1981). The SV40 early region TATA box is required for accurate *in vitro* initiation of transcription. Nature 290, 310-315.

Matthis, P.D., Bernard, H.U., Scott, A., Brady, G., Hushimoto-Gotoh, T. and Schütz, G. (1983). A bovine papilloma virus vector with a dominant resistance marker replicates extrachromosomally in mouse and *E. coli* cells. EMBO J.

2, 1487- 1492.

McGhee, J.D., Wood, W.I., Dolan, M., Engel, J.D. and Felsenfeld, G. (1981). A 200 base pair region at the 5' end of the chicken adult β-globin gene is accessible to nuclease digestion. Cell 27, 45-55.

McIntyre, K.R. and Seidman, J.G. (1984). Nucleotide sequence of mutant I-Aβ$^{bm12}$ gene is evidence for genetic exchange between mouse immune response genes. Nature 308, 551-553.

McKnight, S.L. (1982). Functional relationships between transcriptional control signals of the thymidine kinase gene of Herpes simplex virus. Cell 31, 355-365.

McKnight, S.L. and Kingsbury, R. (1982). Transcriptional control signals of a eukaryotic protein-coding gene. Science 217, 316-325.

McKnight, S.L., Kingsbury, R.C., Spence, A. and Smith, M. (1984). The distal transcription signals of the Herpes virus tk gene share a common hexanucleotide control sequence. Cell 37, 253- 262.

Melton, D.W., Konecki, D.S., Brennand, J. and Caskey, T.C. (1984). Structure, expression and mutation of the hypoxanthine phosphoribosyl-transferase gene. Proc.Natl.Acad.Sci. USA 81, 2147- 2151.

Messing, J. and Vieira, J. (1982). A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. Gene 19, 269-276.

Montell, C., Fisher, E.F., Caruthers, M.H. and Berk, A.J. (1983). Inhibition of RNA cleavage but not polyadenylation by a point mutation in mRNA 3' consensus sequence AAUAAA. Nature 305, 600- 605.

Motwani, N.M., Unakar, N.J. and Roy, A.K. (1980). Multiple hormone requirement for the synthesis of $\alpha_{2u}$ globulin by monolayers of rat hepatocytes in long term primary culture. Endocrinology 107, 1606-1613.

Mount, S.M. (1982). A catalogue of splice junction sequences. Nucleic Acids Res. 10, 459-472.

Mount, M.S., Pettersson, I., Hinterberger, M., Karmas, A. and Steitz, J.A. (1983). The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. Cell 33, 509-518.

Murray, N.E., Brammar, W.J. and Murray, K. (1977). Lambdoid phages that simplify the recovery of in vitro recombinants. Mol. Gen.Genet. 150, 53-61.

Nabeshima, Y., Fujii-Kuriayama, Y., Muramatsu, M. and Ogata, K. (1984). Alternative transcription and two modes of splicing result in two myosin light chains from one gene. Nature 308, 333- 338.

Nagata, S., Mantei, N. and Weissman, C. (1980). The structure of one of the eight or more distinct chromosomal genes for human interferon-α. Nature 287, 401-408.

Nishioka, Y., Leder, A. and Leder, P. (1980). Unusual α- globin-like gene that has clearly lost both globin intervening sequences. Proc.Natl.Acad.Sci. USA 77, 2806-2809.

Ollo, R. and Rougeon, F. (1983). Gene conversion and polymorphism: Generation of mouse immunoglobulin α2a chain alleles by differential gene conversion by γ2b chain gene. Cell 32, 515-523.

Orkin, S.H., Kazazian, H.H., Antonarakis, S.E., Goff, S.C., Boehm,

C.D., Sexton, J.P., Waber, P.G. and Giardina, P.J.V. (1982). Linkage of β-thalassaemia mutations and β-globin gene polymorphisms with DNA polymorphisms in human β-globin gene cluster. Nature 296, 627-631.

Padgett, R.A., Hardy, S.F. and Sharp, P.A. (1983). Splicing of adenovirus RNA in a cell free transcription system. Proc. Natl.Acad.Sci. USA 80, 5230-5234.

Padgett, R.A., Mount, S.M., Steitz, J.A. and Sharp, P.A. (1983). Splicing of messenger RNA precursors is inhibited by antisera to small nuclear ribonucleoprotein. Cell 35, 101-107.

Padgett, R.A., Konarska, M.M., Grabowski, P.J., Hardy, S.F. and Sharp, P.A. (1984). Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. Science 225, 898-903.

Parefentjev, J.A. (1932). Calcium and nitrogen content in urine of normal and cancer mice. Proc.Soc.Exp.Biol.Med. 29, 1285-1286.

Payvar, F., Wrange, O., Carlstedt-Duke, J., Okret, S., Gustafsson, J.A. and Yamamoto, K.R. (1981). Purified glucocorticoid receptors bind selectively *in vitro* to a cloned DNA fragment whose transcription is regulated by glucocorticoids *in vivo*. Proc.Natl.Acad.Sci. USA 78, 6628-6632.

Payvar, F., De Franco, D., Firestone, G.L., Edgar, B., Wrange, O., Okret, S., Gustafsson, J.A. and Yamamoto, K.R. (1983). Sequence-specific binding of glucocorticoid receptor to MMTV DNA at sites within and upstream of the transcribed region. Cell 35, 381- 392.

Pelham, H.R.B. (1982). A regulatory upstream promoter element in

the Drosophila Hsp70 heat-shock gene. Cell 30, 517-528.

Pelham, H.R.B. and Bienz, M. (1982). A synthetic heat-shock promoter element confers heat-inducibility on the Herpes simplex virus thymidine kinase gene. EMBO J. 1, 1473-1477.

Pellicer, A., Robins, D., Wold, B., Sweet, R., Jackson, J., Lowy, I., Roberts, J.M., Sim, G.K. Silverstein, S. and Axel, R. (1980). Altering genotype and phenotype by DNA-mediated gene transfer. Science 209, 1414-1422.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980). The evolution of genes: the chicken preproinsulin gene. Cell 20, 555-566.

Petras, M.L. and MacLaren, I.A. (1976). Structural differences between albumin A and albumin C of the house mouse, Mus musculus. Biochemical Genetics 14, 67-73.

Pfahl, M. (1982). Specific binding of the glucocorticoid-receptor complex to the mouse mammary tumor proviral promoter region. Cell 31, 475-482.

Pfahl, M., McGinnis, D., Hendricks, M., Groner, B. and Hynes, N.E. (1983). Correlation of glucocorticoid receptor binding sites on MMTV proviral DNA with hormone inducible transcription. Science 222, 1341-1343.

Picard, D. and Schaffner, W. (1984). A lymphocyte-specific enhancer in the mouse immunoglobulin k gene. Nature 307, 80-82.

Polisky, B., Greene, P., Garfin, D.E., McCarthy, B.J., Goodman, H.M. and Boyer, H.W. (1975). Specificity of substrate recognition by the EcoRI restriction endonuclease. Proc.Natl.Acad.Sci. USA 72, 3310- 3314.

Potter, M., Finlayson, J.S., Bailey, D.W., Mushinski, E.B., Reamer, B.L. and Walters, J.L. (1973). Major urinary protein and immunoglobulin allotypes of recombinant inbred mouse strains. Genet.Res., Camb. 22, 325-328.

Proudfoot, N.J. and Brownlee, G.G. (1976). 3' non-coding region sequences in eukaryotic messenger RNA. Nature 263, 211-214.

Proudfoot, N.J. and Maniatis, T. (1980). The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication. Cell 21, 537-545.

Queen, C. and Baltimore, D. (1983). Immunoglobulin gene transcription is activated by downstream sequence elements. Cell 33, 741-748.

Radding, C.M. (1978). Genetic recombination: strand transfer and mismatch repair. Ann.Rev.Biochem. 47, 847- 880.

Ragg, H. and Weissmann, C. (1983). Not more than 117bp of 5' flanking sequence are required for inducible expression of a human IFN-α gene. Nature 303, 439-442.

Rave, N., Crkvenjakov, R. and Boedtker, H. (1979). Identification of procollagen mRNAs transferred to diazobenzyloxy-methyl paper from formaldehyde agarose gel. Nucleic Acids Res. 6, 3559-3567.

Renkawitz, R., Schütz, G., von der Ahe, D. and Beato, M. (1984). Sequences in the promoter region of the chicken lysozyme gene required for steroid regulation and receptor binding. Cell 37, 503-510.

Reynolds, G.A., Basu, S.K., Osborne, T.F., Chin, D.J., Gil, G., Brown, M.S., Goldstein, J.L. and Luskey, K.L. (1984). HMG CoA reductase: a negatively regulated gene with unusual

promoter and 5' untranslated regions. Cell 38, 275-285.

Rogers, J. and Wall, R. (1980). A mechanism for RNA splicing. Proc. Natl.Acad.Sci. USA 77, 1877-1879.

Rosenfeld, M.G., Lin, C.R., Amara, S.G., Stolarsky, L., Roos, B.A., Ong, E.S. and Evans, R.M. (1982). Calcitonin mRNA polymorphism: Peptide switching associated with alternative RNA splicing events. Proc.Natl.Acad.Sci. USA 79, 1717-1721.

Roy, A.K., Demyan, W.F., Majumdar, D., Murty, C.V.R. and Chatterjee, B. (1983). Age-dependent changes in the androgen sensitivity of rat liver. In Steroid Hormone Receptors: Structure and Function, ed. Eriksson, H. and Gustafsson, J.A., 439- 459. Elsevier Science Publishers B.V.

Royal, A., Garapin, A., Cami, B., Perrin, F., Mandel, J.L., Le Meur, M., Bregegegre, F., Gannon, F., Le Pennec, J.D., Chambon, P. and Kourlisky, P. (1979). The ovalbumin gene region: common features in the organisation of three genes expressed in chicken oviduct under hormonal control. Nature 279, 125-132.

Rümke, P.H. and Thung, P.J. (1964). Immunological studies on the sex-dependent prealbumin in mouse urine and on its occurrence in the serum. Acta Endocrinologia 47, 156-164.

Ruskin, B., Krainer, A.R., Maniatis, T. and Green, M.R. (1984). Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. Cell 38, 317-331.

Sakano, H., Hüppi, K., Heinrich, G. and Tonegawa, S. (1979). Sequences at the somatic recombination sites of immunoglobulin light-chain genes. Nature 280, 288-294.

Sakano, H., Rogers, J.H., Hüppi, K., Brack, C., Trannecker, A.,

Maki, R., Wall, R. and Tonegawa, S. (1979). Domains and the hinge region of an immunoglobulin heavy chain are encoded in separate DNA segments. Nature 277, 627-633.

Salditt-Georgieff, M. and Darnell, J.E. (1983). A precise termination site in the mouse $\beta^{major}$-globin transcription unit. Proc. Natl.Acad.Sci. USA 80, 4694-4698.

Salser, W. (1977). Globin mRNA sequences: Analysis of base pairing and evolutionary implications. Cold Spring Harbor Symposia on Quantitative Biology 42, 985-1002.

Sampsell, B. and Held, W. (1984). Expression of major urinary proteins in wild-derived mice. Mouse News Lett. 71, 45-46.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc.Natl.Acad.Sci. USA 74, 5463-5467.

Scheidereit, C., Geisse, S., Westphal, H.M. and Beato, M. (1983). The glucocorticoid receptor binds to defined nucleotide sequences near the promoter of mouse mammary tumor virus. Nature 304, 749-752.

Searle, P.F., Davison, B.L., Stuart, G.W., Wilkie, T.W., Norstedt, G. and Palmiter, R.D. (1984). Regulation, linkage and sequence of mouse metallothionein I and II genes. Mol.Cell Biol. 4, 1221-1230.

Seed, B. (1983). Purification of genomic sequences from bacteriophage libraries by recombination and selection in vivo. Nucleic Acids Res. 11, 2427-2445.

Setzer, D.R., McGrogan, M., Nunberg, J.H. and Schimke, R.J. (1980). Size heterogeneity in the 3' end of dihydrofolate reductase

messenger RNAs in mouse cells. Cell 22, 361-370.

Setzer, D.R., McGrogan, M. and Schimke, R.T. (1982). Nucleotide sequence surrounding multiple polyadenylation sites in the mouse dihydrofolate reductase gene. J.Biol.Chem. 257, 5143-5147.

Shaw, P.H., Held, W.A. and Hastie, N.D. (1983). The gene family for major urinary proteins: expression in several secretory tissues of the mouse. Cell 32, 755-761.

Shermoen, A.W. and Beckendorf, S.K. (1982). A complex of interacting DNAase I-hypersensitive sites near the Drosophila glue protein gene, Sgs4. Cell 29, 601-607.

Sirover, M.A. and Loebb, L.A. (1977). On the fidelity of DNA replication. Effect of metal activators during synthesis with avian meyeloblastosis virus DNA polymerase. J.Biol.Chem. 252, 3605-3610.

Slighton, J.L., Blechl, A.G. and Smithies, O. (1980). Human fetal $^{G}\gamma$- and $^{A}\gamma$-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21, 627-638.

Smith, G.P. (1976). Evolution of repeated DNA sequences by unequal crossover. Science 191, 528-535.

Smith, G.R., Schultz, D.W. and Creasemann, J.M. (1980). Generalized recombination: nucleotide sequence homology between Chi recombinational hotspots. Cell 19, 785-793.

Smith, G.R., Kunes, S.M., Schultz, D.W., Taylor, A. and Triman, K.L. (1981). Structure of Chi hotspots of generalized recombination. Cell 24, 429-436.

Smith, G.R. (1983). Chi hotspots of generalized recombination. Cell

34, 709-710.

Spandidos, D.A. and Paul, J. (1982). Transfer of human globin genes to erythroleukemic mouse cells. EMBO J. 1, 15-20.

Stahl, F.W. (1979). Special sites in generalized recombination. Ann. Rev.Genet. 13, 7-24.

Stein, J.P., Catterall, J.F., Kristo, P., Means, A.R. and O'Malley, B.W. (1980). Ovomucoid intervening sequences specify functional domains and generate protein polymorphism. Cell 21, 681-687.

Stein, R., Gruenbaum, Y., Pollack, Y., Razin, A. and Cedar, H. (1982). Clonal inheritance of the pattern of DNA methylation in mouse cells. Proc.Natl.Acad.Sci. USA 79, 61-65.

Steinmeitz, M. and Hood, L. (1983). Genes of the major histocompatibility complex in mouse and man. Science 222, 727- 733.

Steinmeitz, M., Malissen, M., Hood, L., Orn, A., Maki, R.A., Dastoornikoo, G.R., Stephan, D., Gibb, E. and Romaniuk, R. (1984). Tracts of high or low sequence divergence in the mouse major histocompatibility complex. EMBO J. 3, 2995-3003.

Südhof, T.C., Goldstein, J.L., Brown, M.S. and Russell, D.W. (1985a). The LDL receptor gene: a mosaic of exons shared with different proteins. Science 228, 815-822.

Südhof, T.C., Russell, D.W., Goldstein, J.L. and Brown, M.S. (1985b). Cassette of eight exons shared by genes for LDL receptor and EGF precursor. Science 228, 893-895.

Supowit, S.C., Potter, E., Evan, R.M. and Rosenfeld, M.G. (1984).

Polypeptide hormone regulation of gene transcription: specific 5' genomic sequences are required for epidermal growth factor and phorbol ester regulation of prolactin gene expression. Proc.Natl.Acad.Sci. USA 81, 2975-2979.

Sweet, R.W., Chao, M.V. and Axel, R. (1982). The structure of the thymidine kinase gene promoter: Nuclease hypersensitivity correlates with expression. Cell 31, 347- 353.

Swift, G.H., Hammer, R.E., MacDonald, R.J. and Brinster, R.L. (1984). Tissue-specific expression of the rat pancreatic elastase 1 gene in transgenic mice. Cell 38, 639-646.

Szoka, P. and Paigen, K. (1978). Regulation of mouse major urinary protein production by the Mup-a gene. Genetics 90, 597-612.

Szoka, P. and Paigen, K. (1979). Genetic regulation of MUP production in recombinant inbred mice. Genetics 93, 173-181.

Taniguchi, T., Mantei, N., Schwarzstein, M., Nagata, S., Murumatsu, M. and Weissmann, C. (1980). Human leukocyte and fibroblast interferons are structurally related. Nature 285, 547-549.

Thomas, P.S. (1980). Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. Proc.Natl.Acad.Sci. USA 77, 5201-5205.

Treisman, R., Proudfoot, N.J., Shander, M. and Maniatis, T. (1982). A single-base change at a splice site in a $\beta^{0}$ - thalassemic gene causes abnormal RNA splicing. Cell 29, 903-911.

Treisman, R., Orkin, S.H. and Maniatis, T. (1983). Specific transcription and RNA splicing defects in five cloned β-thalassaemia genes. Nature 302, 591-596.

Unterman, R.D., Lynch, K.R., Nakhasi, H.L., Dolan, K.P., Hamilton, J.W., Cohn, D.V. and Feigelson, P. (1981). Cloning and sequence of several $\alpha_{2u}$-globulin cDNAs. Proc.Natl.Acad. Sci. USA 78, 3478- 3482.

van der Ploeg, L.H.T. and Flavell, R.A. (1980). DNA methylation in the human γδβ-globin locus in erythroid and nonerythroid tissues. Cell 19, 947-958.

Vanin, E.F., Goldberg, G.I., Tucker, P.W. and Smithies, O. (1980). A mouse α-globin-related pseudogene lacking intervening sequences. Nature 286, 222-226.

Vardimon, L., Kressman, A., Cedar, H., Machler, M. and Doerfler, W. (1982). Expression of a cloned adenovirus gene is inhibited by in vitro methylation. Proc.Natl.Acad.Sci. USA 79, 1073-1077.

Waechter, D.E. and Baserga, R. (1982). Effect of methylation on expression of microinjected genes. Proc.Natl.Acad.Sci. USA 79, 1106-1110.

Wahl, G.M., Stern, M. and Stark, G.R. (1979). Efficient transfer of large DNA fragments from agarose gels and rapid hybridization by using dextran sulphate. Proc.Natl.Acad.Sci. USA 76, 3683-3687.

Wahli, W., Dawid, I.B., Wyler, T., Weber, R. and Ryffel, G.U. (1980). Comparative analysis of the structural organization of two closely related vitellogenin genes in Xenopus laevis. Cell 20, 107-117.

Walker, M.D., Edlund, T., Boulet, A.M. and Rutter, W.J. (1983). Cell-specific expression controlled by the 5'- flanking region of insulin and chymotrypsin genes. Nature 306, 557-561.

Wasylyk, B., Derbyshine, R., Guy, A., Molko, D., Roget, A., Teoule, R. and Chambon, P. (1980). Specific in vitro transcription of conalbumin gene is drastically decreased by single-point mutation in TATA box homology sequence. Proc.Natl.Acad.Sci. USA 77, 7024-7028.

Weaver, R.F. and Weissman, C. (1979). Mapping of RNA by a modification of the Berk-Sharp procedure: the 5' termini of 15S β-globin mRNA precursor and mature 10S β-globin mRNA have identical map coordinates. Nucleic Acids Res. 7, 1175-1193.

Weil, P.A., Luse, D.S., Segall, J. and Roeder, R.G. (1979). Selective and accurate initiation of transcription at the Ad2 major late promoter in a soluble system dependent on purified RNA poly-merase II and DNA. Cell 18, 469-484.

Weintraub, H. and Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. Science 193, 848-856.

Weintraub, H., Larsen, A. and Groudine, M. (1981). α-globin gene switching during the development of chicken embryos: expression and chromosome structure. Cell 24, 333-344.

Weintraub, H., Beug, H., Groudine, M. and Graf, T. (1982). Temperature-sensitive changes in the structure of globin chromatin in lines of red cell precursors transformed by ts-AEV. Cell 28, 931-940.

Weiss, E.H., Mellor, A., Golden, L., Fahrner, K., Simpson, E., Hurst, J. and Flavell, R.A. (1983). The structure of a mutant H-2 gene suggests that the generation of polymorphism in H-2 genes may occur by gene conversion-like events. Nature 301, 671-674.

Weissmann, C. (1984). Excision of introns in lariat form. Nature 311, 103-104.

Wellauer, P.K., Reeder, R.H., Dawid, I.B. and Brown, D.D. (1976). The arrangement of length heterogeneity in repeating units of ribosomal DNA from Xenopus laevis. J.Mol.Biol. 105, 487-505.

Wellauer, P.K. and Dawid, I.B. (1977). The structural organization of ribosomal DNA in Drosophila melanogaster. Cell 10, 193-212.

Wieringa, B., Meyer, F., Reiser, J. and Weissmann, C. (1983). Unusual sequence of cryptic splice sites utilized in the β-globin gene following inactivation of an authentic 5' splice site by site-directed mutagenesis. Nature 301, 38-43.

Wieringa, B., Hofer, E. and Weissmann, C. (1984). A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. Cell 37, 915-925.

Wilde, C.D., Crowther, C.E., Cripe, T.P., Lee, M.G.S. and Cowan, N.J. (1982). Evidence that a human β-tubulin pseudogene is derived from its corresponding mRNA. Nature 297, 83-84.

Woodbury, C.P.Jr., Hagenbuchle, O. and von Hippel, P. (1980). DNA site recognition and reduced specificity of the EcoRI endonuclease. J.Biol.Chem. 255, 11534-11546.

Wright, S., de Boer, E., Grosveld, F.G. and Flavell, R.A. (1983). Regulated expression of the human β-globin gene family in murine erythroleukemic cell hybrids. Nature 305, 333-336.

Wright, S., Rosenthal, A., Flavell, R. and Grosveld, F.G. (1984). DNA sequences required for regulated expression of β-globin genes in murine erythroleukemia cells. Cell 38, 265-273.

Wu, C., Bingham, P.M., Livak, K.J., Holmgren, R. and Elgen, S.C.R. (1979a). The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell 16, 797-806.

Wu, C., Wong, Y.C. and Elgin, S.C.R. (1979b). The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. Cell 16, 807- 814.

Wu, C. (1980). The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNAse I. Nature 286, 854-860.

Wu, C. (1984). Two protein-binding sites in chromatin implicated in the activation of heat-shock genes. Nature 309, 229-234.

Yamada, Y., Avvedimento, V.E., Mudryj, M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I. and de Crombrugghe, B. (1982). The collagen gene: evidence for its evolutionary assembly by amplification of a DNA segment containing an exon of 54bp. Cell 22, 887-892.

Yamada, Y., Liau, G., Mudryj, M., Obici, S. and de Crombrugghe, B. (1984). Conservation of the sizes for one but not another class of exons in two chicken collagen genes. Nature 310, 333-337.

Yang, V.W., Lerner, M.R., Steitz, J.A. and Flint, S.J. (1981). A small nuclear ribonucleoprotein is required for splicing of adenoviral early RNA sequences. Proc.Natl.Acad.Sci. USA 78, 1371-1375.

Zaret, K.S. and Yamamoto, K.R. (1984). Reversible and persistent changes in chromatin structure accompany activation of a glucocorticoid-dependent enhancer element. Cell 38, 29-38.

Zimmer, E.A., Martin, S.L., Beverley, S.M., Kan, Y.W. and Wilson, AC. (1980). Rapid duplication and loss of genes coding for the chains of haemoglobin. Proc.Natl.Acad.Sci. USA 77, 2158- 2162.

Zinn, K., Di Maio, D. and Maniatis, T. (1983). Identification of two distinct regulatory regions adjacent to the human β-interferon gene. Cell 34, 865-879.

## Abbreviations used in text

| | |
|---|---|
| AEV | avian erythroblastosis virus |
| ASV | avian sarcoma virus |
| ATP | adenosine 5' triphosphate |
| BSA | bovine serum albumin |
| BPV | bovine papiloma virus |
| cAMP | cyclic (3' - 5') adenosine monophosphate |
| cDNA | DNA copy of RNA |
| CaMV | cauliflower mosaic virus |
| cpm | counts per minute |
| dATP | deoxyadenosine triphosphate |
| dCTP | deoxycytidine triphosphate |
| DEAE | diethylaminoethyl |
| DNA | deoxyribonucleic acid |
| DNase | deoxyribonuclease |
| DTT | dithiothreotol |
| EDTA | diaminoethanetetra-acetic acid |
| GF/C | glass fibre filter |
| HMG CoA | 3-hydroxy-3-methylglutaryl coenzyme A reductase |
| HSV | herpes simplex virus |
| IEF | isoelectric focusing |
| kbp | kilobase pair |
| LB | L broth |
| LTR | long terminal repeat |
| MHC | major histocompatibility complex |
| MMTV | mouse mammary tumor virus |
| MOPS | morpholinopropane sulphuric acid |

| | |
|---|---|
| mRNA | messenger RNA |
| MUP | major urinary protein |
| OAc | acetate |
| $OD_x$ | optical density at a wavelength of x nanometers |
| PEG | polyethylene glycol |
| pfu | plaque forming unit |
| pH | -log [H+] |
| phage | bacteriophage |
| PPO | 2,5 - diphenyloxazole |
| POPOP | 1,4-bis-2-(4-methyl-5-phenyloxazolyl)-benzine |
| poly(A) | polyadenylic acid |
| poly(A) RNA | polyadenylated RNA |
| S1 | single strand specific nuclease |
| rDNA | DNA coding for ribosomal RNA |
| rev. $min^{-1}$ | revolutions per minute |
| RNA | ribonucleic acid |
| RNase | ribonuclease |
| rRNA | ribosomal RNA |
| SDS | sodium dodecyl sulphate |
| TCA | trichloroacetic acid |
| Tris | tris-[hydroxymethyl]-aminomethane |
| ts | temperature sensitive |
| V | volts |
| vol. | volume |
| W/V | weight per volume |