ACTIN SEQUENCES AND ASSOCIATED ELEMENTS IN THE MOUSE GENOME

YING M. MAN

THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOCHEMISTRY,
UNIVERSITY OF GLASGOW

APRIL, 1987

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

ACKNOWLEDGEMENTS

I am particularly grateful to my supervisor, Dr. David Leader, for allowing me the opportunity to carry out this work. His invaluable guidance, constructive criticism and helpful assistance have been very useful throughout the course of this project, including the careful reading of this manuscript.

I also wish to thank Professor R.M.S. Smellie and the Department of Biochemistry for making available to me the facilities for this research. My thanks to Dr. Roger Adams, who as my auditor, has taken an interest in my progress, and to Dr. Bob Eason for providing assistance with the computing facilities.

My gratitude is also extended to all my friends in laboratories C36 and C35, both past and present, for their encouragement and many varied aspects of assistance. Especially to Carolyn Begg for being a 'pal' and for providing many useful suggestions, her continuous moral support was a great source of encouragement to me. Grateful acknowledgement is due to Irene Gall, whose skilful technical expertise has contributed in many ways to the completion of this project.

Finally, sincere thanks are expressed to my parents for their encouragement and support throughout these three years.

ABBREVIATIONS

The abbreviations recommended by the *Biochemical Journal* in its Instructions to Authors [*Biochemical Journal* (1985) **225**, 1-26] have been used throughout this thesis with the following additions.

BSA	bovine serum albumin
cDNA	complementary DNA
DNase	deoxyribonuclease
dNTP	deoxynucleoside-5'-triphosphate
PEG	polyethylene glycol
pfu	plaque forming unit
p.s.i.	pound-force per square inch
RNase	ribonuclease
rpm	revolutions per minute
SDS	sodium dodecyl sulphate

CONTENTS

FIGURES AND TABLES

SUMMARY

This work describes the structural analysis of four mouse genomic clones which had previously been shown by electron microscopic heteroduplex analysis to contain actin-like genes, each with a single interruption. The objective of this work was to determine the nature of these interruptions.

The first part of this work involved the characterisation of the actin-like DNA of three of these clones (λmA82, λmA118, and λmA119), in order to determine whether they were functional genes or pseudogenes. Restriction analysis of these clones was carried out and provided evidence to suggest that two of these clones, λmA82 and λmA119, were overlapping sequences from the same genomic region. λmA119 was chosen for further analysis because it contains more genomic DNA. The actin-like sequences in λmA118 and λmA119 were determined by the chemical method, and were found to resemble genes specifying the cytoplasmic γ-actin isoform. However, both contained mutations that would prevent them encoding a functional γ-actin. It was concluded therefore that λmA118 and λmA119 are pseudogenes, and the absence of multiple introns indicated that these were of the processed type.

The sequence of the actin pseudogene in λmA118 extended only from amino-acid 5, suggesting generation from an incomplete reverse transcript, as had been found previously for another γ-actin processed pseudogene. This suggests that there may be extensive secondary structure at the 5' end of the mRNA. The 3' untranslated region of λmA119 only extended for 108 of an expected 700 nucleotides, suggesting that it had suffered a deletion after integration. Making the assumption that these processed pseudogenes have accumulated neutral changes at a constant rate, free from any selective pressure, then the times at which λmA118 and λmA119 arose were estimated to be 6.8 and 4.4 million years ago, respectively. Comparison of three pseudogene sequences with the γ-actin cDNA sequence allowed certain of the differences found in λmA118 to be ascribed to silent mutations in the functional gene that

have occurred since λmA118 arose. The ratio of mutations in functional gene and pseudogene was consistent with a similar rate of mutation in the silent positions of the functional gene and in the pseudogene as a whole, in contrast to the results for some other pseudogenes. Examination of the proportion of mutations in λmA118 in positions corresponding to the silent and replacement positions of the functional gene showed an unexpected deviation from the ratio expected for totally neutral evolution of a pseudogene. This could be accounted for by a distorting effect of frequent transitions from CG doublets, thought to be due to deamination of 5-methyl cytosine.

The second part of this work involved the structural analysis of the DNA regions corresponding to the loops interrupting the genomic clones λmA118 and λmA119, and that interrupting λmA36, which had been shown by another worker also to contain a γ-actin processed pseudogene. The nucleotide sequences of the inserted elements (IEs) of λmA36, λmA118, and λmA119 were determined. Each was found to interrupt the actin-like DNA at a different position, none of which corresponded to that of an intron in the gene of any actin isoform yet sequenced. Furthermore, in no case was there a perfect match to the consensus sequence of intron/exon splice sites. The inserted elements were, however, flanked by short (4 to 6 base pair) direct repeats of actin-like sequence. Thus the inserted elements did not appear to represent residual introns, but rather transposon-like sequences that had inserted into the pseudogenes at staggered breaks.

It was found that IE 36 was 500 base pairs in length and was related to the long terminal repeat (LTR) of the retroviral-like intracisternal A-particle. This is the first such intracisternal A-particle solo LTR to be reported. However, IE 36 differs from a normal intracisternal A-particle LTR in containing a 46 nucleotide region which has undergone 5 successive duplications together with a subsequent deletion. This had occurred in the R region of the LTR, which appears particularly prone to rearrangement in intracisternal A-particle genes. It was estimated that there are 1,900 copies related to IE 36 per mouse haploid genome, consistent with the values for intracisternal A-particle genes estimated by others.

IE 119 was found to be 501 base pairs in length and was also related to

LTRs of the recently-described MS57 (632 base pairs in length) and GLN-3 (430 base pairs in length) retroviral-like elements. In the case of this family of retroviral-like genes, expansion appears to occur in the U3 region of the LTR. There are approximately 2,300 copies of IE 119 per mouse haploid genome.

IE 118 was found to be 865 base pairs in length and repeated 1,000 to 2,000 times in the mouse genome. Computer searches of the GenBank and EMBL nucleotide sequence databanks did not reveal any sequence of significant similarity to IE 118. However, several of the functionally important sequence motifs found in retroviral LTRs could be recognised in IE 118, albeit in an imperfect form. Therefore, the most likely possibility is that IE 118 is also a solo LTR of a hitherto unrecognised family of mouse retroviruses or retroviral-like elements. IE 118 also possesses a stretch of 27 out of 28 nucleotides identical to a region of the flanking actin pseudogene but in the opposite orientation. This may have arisen by a gene conversion event after the integration of IE 118 into its target pseudogene.

CHAPTER 1

INTRODUCTION

The concerns of this thesis are mouse actin genes and pseudogenes, and possible mobile elements associated with them. This is the basis for the choice of topics that are dealt with in this Introduction.

1.1 Actins and their Genes

1.1.1 Actins

Actin is an abundant, highly conserved protein that is found in all eukaryotic cells. In animals, actin is primarily involved in muscle contraction in differentiated striated and smooth muscle tissues. In non-muscle animal cells, actin is involved in a variety of processes, including maintenance of cytoskeletal structure, cellular motility, cell-surface mobility, intracellular transport, cytoplasmic streaming, cytokinesis, exocytosis, clot retraction, microvillar movement and, possibly, chromosomal condensation and mitosis (Schliwa, 1981; Lloyd, 1983; Ponte *et al.*, 1983; Stossel, 1984). Isoelectric focusing allowed resolution of isoforms of actin with different isoelectric points. Three different positions of migration (α, ß and γ) were observed, with striated muscle actins migrating as α, smooth muscle actins as α and γ, and non-muscle actins as ß and γ. Amino-acid sequencing studies of actins from mammalian sources have further shown the presence of at least three distinct α-actins (α-skeletal, α-cardiac and α-smooth) and two distinct γ-actins (γ-smooth and γ-cytoplasmic), bringing the number of known functional mammalian actin genes to six (Vandekerckhove and Weber, 1978a and 1978b). In birds and amphibians a third cytoplasmic isoform has been identified (Vandekerckhove *et al.*, 1981a; Bergsma *et al.*, 1985).

Each of the four muscle actins tends to predominate in a particular muscle tissue. Thus the tissue distribution of α-skeletal and α-cardiac actins reflects their names, and the different smooth muscle tissues tend to possess

predominantly either α or γ-smooth muscle actins. The striated muscle isoforms may, however, be coexpressed in a tissue under some circumstances. For example, the mouse α-cardiac actin is expressed not only in the adult cardiac muscle but also (along with the more abundant α-skeletal form) in foetal skeletal muscle (Minty *et al.*, 1982); and the human α-skeletal and α-cardiac actin genes are coexpressed in skeletal and cardiac muscles (Gunning, *et al.*, 1983). The smooth muscle actins appear to be similarly coexpressed, although in the genital and gastrointestinal tracts, γ-smooth muscle actin predominates, whereas in vascular tissue, such as aorta, α-smooth muscle actin is the primary isoform (Vandekerckhove and Weber, 1979a and 1984; Vandekerckhove *et al.*, 1981a; Gabbiani *et al.*, 1981). Cytoplasmic actins show an even more pronounced pattern of coexpression, with no non-muscle cell-type known to express predominantly either ß or γ-isoform (Vandekerckhove *et al.*, 1981b).

The vertebrate actin isoforms contain slight differences in their amino-acid sequences, and these are primarily located in the amino-terminal end of the proteins (Vandekerckhove and Weber, 1979a). The positions in the amino-acid sequence at which differences exist between the six actin isoforms are shown in Table 1.1. There are from 4 to 8 amino-acid differences between the four different muscle isoforms; 4 differences between the two cytoplasmic isoforms; and up to 25 differences between the cytoplasmic and muscle isoforms (Vandekerckhove and Weber, 1979a). Thus the muscle isoforms and cytoplasmic isoforms are more closely related to themselves than to one another. The amino-acid sequences for a single isoform of actin from diverse organisms are extremely similar. For example, chicken, bovine, and rabbit skeletal muscle actins have identical amino-acid sequences (Vandekerckhove and Weber, 1979a and 1979b). It is clear that among vertebrates the amino-acid sequences of actins are isoform specific, rather than species specific.

All eukaryotes synthesize one or more cytoplasmic actin isoform (Vandekerckhove *et al.*, 1981b). The vertebrate cytoplasmic ß and γ-actin are considered functionally and evolutionarily more closely related to the actins found in the lower eukaryotes. For example, yeast actin differs from the mammalian cytoplasmic γ-isoform in 41 positions, but from the α-skeletal muscle isoform in 49 positions, out of a total of 375 (Gallwitz and Sures, 1980; Ng and Abelson, 1980). In *Drosophila melanogaster*, actins with amino-acid sequences similar to the vertebrate cytoplasmic actins are utilised to form the actin filaments of sarcomeric muscle (Fyrberg *et al.*, 1981). It has been proposed that during early chordate evolution a novel actin isoform arose

Table 1.1 Differences in the amino-acid sequences of actin isoforms

Residue number	Actin isoforms					
	Skeletal muscle	Cardiac muscle	Smooth muscle (stomach)	Smooth muscle (aorta)	Non-muscle	
					ß-type	γ-type
1	Asp	Asp	-	Glu	Met	-
2	Glu	Asp	Glu	Glu	Asp	Glu
3	Asp	Glu	Glu	Glu	Asp	Glu
4	Glu	Glu	Glu	Asp	Asp	Glu
5	Thr	Thr	Thr	Ser	Ile	
6	Thr	Thr	Thr	Thr	Ala	
10	Cys	Cys	Cys	Cys	Val	Ile
16	Leu	Leu	Leu	Leu	Met	
17	Val	Val	Cys	Cys	Cys	
76	Ile	Ile	Ile	Ile	Val	
89	Thr	Thr	Ser	Ser	Thr	
103	Thr	Thr	Thr	Thr	Val	
129	Val	Val	Val	Val	Thr	
153	Leu	Leu	Leu	Leu	Met	
162	Asn	Asn	Asn	Asn	Thr	
176	Met	Met	Met	Met	Leu	
201	Val	Val	Val	Val	Thr	
225	Asn	Asn	Asn	Asn	Gln	
259	Thr	Thr	Thr	Thr	Ala	
266	Ile	Ile	Ile	Ile	Leu	
271	Ala	Ala	Ala	Ala	Cys	
278	Tyr	Tyr	Tyr	Tyr	Phe	
286	Ile	Ile	Ile	Ile	Val	
296	Asn	Asn	Asn	Asn	Thr	
298	Met	Leu	Leu	Leu	Leu	
357	Thr	Ser	Ser	Ser	Ser	
364	Ala	Ala	Ala	Ala	Ser	

The table indicates the positions in the amino-acid sequence at which exchanges have been detected between the different actin isoforms. Numbering of positions of the amino-acids in the actin sequence is made by analogy to rabbit skeletal muscle actin (Collins and Elzinga, 1975; Lu and Elzinga, 1977; Vandekerckhove and Weber, 1978c). Amino-acid residues in which the four muscle actins differ among themselves are underlined.

which now functions in the sarcomeres of muscle cells (Vandekerckhove *et al.*, 1983). In the time before the divergence of mammals and birds, this gene apparently underwent two successive duplications to produce the four muscle-actin isoforms found in mammals and birds today (Vandekerckhove *et al.*, 1983). Thus the muscle-actin isoforms must have been under strong selection pressure to maintain their amino-acid sequence since they arose.

1.1.2 Actin Genes

Actin cDNA clones that are specific for particular isoforms have been isolated (Ponte *et al.*, 1983; Gunning *et al.*, 1983), and have been used as hybridisation probes to assay the number and organisation of sequences related to actins in the genomes of different organisms, and to isolate individual genomic sequences.

(i) The Number of Actin Genes

Southern analysis and hybridisation of the genomic DNA with appropriate actin probes under low stringency washing conditions has allowed the estimation of the number of recognisable actin genes in different genomes. It was found that the number of actin genes in higher eukaryotes varies considerably. It was estimated that in chicken there are 4 to 7 actin genes (Cleveland *et al.*, 1980); in man, 20 to 30 actin genes (Moos and Gallwitz, 1983; Engel *et al.*, 1982); in rat, 12 or more actin genes (Nudel *et al.*, 1982); and in mouse, more than 20 actin genes (Minty *et al.*, 1983). In mammals these numerous actin sequences are scattered on different chromosomes throughout the genome (Soriano *et al.*, 1982). The number of actin genes in lower eukaryotes also varies between organisms. *Drosophila melanogaster* contains 6 actin genes (Fyrberg *et al.*, 1981), yeast contains 1 actin gene (Gallwitz and Sures, 1980; Ng and Abelson, 1980), *Dictyostelium discoideum* contains 17 actin genes (McKeown and Firtel, 1981), and sea urchin contains 11 actin genes (Scheller *et al.*, 1981).

When genomic DNA from higher eukaryotes was analysed by Southern blotting under high stringency conditions, so that only sequences highly homologous to the cDNA probe remain hybridised, single bands were obtained for a number of different isoforms (Minty *et al.*, 1983; Weydert *et al.*, 1983; Robert *et al.*, 1984). It is assumed that these represent the corresponding

functional genes, which, like those for most other structural proteins, appear to be present in a single copy per haploid genome (Minty *et al.*, 1983; Ponte *et al.*, 1983; Robert *et al.*, 1984).

In this kind of analysis, many of the multiple actin-related sequences detected under low stringency in the mammalian genome have been found to hybridise preferentially to probes for the cytoplasmic isoforms of actins. Thus it is more difficult to determine the number of functional genes for the cytoplasmic actin isoforms. However, it seems likely that these too are present as single copies, the related sequences being processed pseudogenes (see section 1.2.3) which are thought to be derived from ß or γ-actin mRNAs by reverse transcription and reintegration of the complementary DNA into the genome (Minty *et al.*, 1983; Moos and Gallwitz, 1982; Carmon *et al.*, 1982). The extent to which these sequences have diverged from the actin coding sequence and, hence, the time which has elapsed since their integration, varies. A family of highly diverged sequences of this type has been isolated from the mouse genome. These closely related sequences are probably the result of a recent amplification of a 17 kb region of mouse DNA containing a diverged actin pseudogene (Minty *et al.*, 1983). The high number of sequences related to actin is apparently restricted to the mammalian genome; in birds (Cleveland *et al.*, 1980) or in *Drosophila* (Fyrberg *et al.*, 1980), for example, the number of genomic actin sequences corresponds to the number of known proteins.

(ii) Structural Features of Actin Genes

As already mentioned, the coding sequences of the actin genes are highly conserved. On the other hand, the non-coding parts of the actin genes are quite divergent when compared along a wide evolutionary range. For example, the sizes of introns and their locations vary considerably. In protostomes such variability in intron positions is most apparent (Fyrberg *et al.*, 1981), although this is less so in deuterostomes (Fornwald *et al.*, 1982; Zakut *et al.*, 1982; see Table 1.2).

The heterogeneity in the location and number of introns in actin genes has led to the question of whether introns have been deleted or inserted during evolution. The available data are not sufficient to answer this question conclusively, although they have been interpreted in favour of intron deletion. A comparison of the intron positions in the actin genes of deuterostomes with those found in the recently sequenced α-smooth muscle

Table 1.2 Position of introns in the actin genes of various organisms

Actin gene	Organism	Intron position							
		5'UTR	41/42	84/85	121/122	150	204	267	327/328
α-smooth	chicken[1]	•	•	•	•	•	•	•	•
α-smooth	human[2]	?	•	•	•	•	•	•	•
α-skeletal	mouse[3], chicken[4], rat[5]	•	•			•	•	•	•
α-cardiac	chicken[6]	•	•			•	•	•	•
α-cardiac	human[7]	?	•			•	•	•	•
ß-cytoplasmic	chicken[8], rat[9], human[10]	•	•		•			•	•
SpG28	sea urchin[11]			•		•		•	•
SpG17	sea urchin[11]					•		•	
SfA	sea urchin[12]					•		•	

The intron positions in actin genes from various organisms are shown, the numbering being of the codons interrupted (relative to the vertebrate sequence). The key to references is :

1) Carroll *et al.*, (1986)
2) Ueyama *et al.*, (1984)
3) Hu *et al.*, (1986)
4) Fornwald *et al.*, (1982)
5) Zakut *et al.*, (1982)
6) Chang *et al.*, (1985)
7) Hamada *et al.*, (1982)
8) Kost *et al.*, (1983)
9) Nudel *et al.*, (1983)
10) Ng *et al.*, (1985)
11) Cooper and Crain, (1982)
12) Foran *et al.*, (1985)

actin gene (Carroll *et al.*, 1986), is consistent with this suggestion. It was demonstrated that the structural sequence of the chicken α-smooth muscle actin gene is interrupted by eight introns. Examination of the intron positions in vertebrate α-skeletal (Fornwald *et al.*, 1982; Zakut *et al.*, 1982; Hu *et al.*, 1986), α-cardiac (Hamada *et al.*, 1982; Chang *et al.*, 1985) and ß-cytoplasmic (Nudel *et al.*, 1983; Kost *et al.*, 1983; Ng *et al.*, 1985) actin genes, as well as those found in sea urchin genes (Cooper and Crain, 1982; Foran *et al.*, 1985), revealed that the intron positions in these latter genes represent subsets of the intron positions found in the chicken α-smooth muscle actin gene (Carroll *et al.*, 1986; Table 1.2). This observation suggests a common ancestral gene with multiple intron positions which have been partially lost during evolution. It was therefore concluded, at least for the case of the deuterostome actin genes, that intron deletion has been the dominant process influencing the placement of introns in modern actin genes (Zakut *et al.*, 1982; Blake, 1983; Nudel *et al.*, 1984; Carroll *et al.*, 1986). From this standpoint, the fact that lower organisms with high reproductive rates have lost more introns than higher organisms (Table 1.2) is rationalised in terms of a need to minimise the size of their genomes.

When the nucleotide sequences of different actin isoforms are compared in a single species, only the coding regions show a high degree of homology. No significant homology was detected in the 5' and 3' untranslated regions of genes for different actin isoforms. Moreover, the lengths of these parts of the genes often vary considerably, which suggests that at least part of the sequence heterogeneity is due to deletion and/or insertion of DNA. On the other hand, comparison of the genes for a single actin isoform in different mammals shows a considerable degree of homology between their untranslated regions. For example, the 3' untranslated regions of rat (Mayer *et al.*, 1984) and human (Hamada *et al.*, 1982) cardiac actin genes show a high degree of homology : two-thirds of the 3' part of these regions exhibit 92.5% homology and the 5' part of this region shows 85% homology. Similarly, a large part of the 3' untranslated regions of human (Hanukoglu *et al.*, 1983) and rat (Nudel *et al.*, 1983) ß-actins shows more than 85% homology. In fact, it has been shown that the 3' untranslated regions of actin mRNAs in birds and mammals are unique for each actin isoform (Cleveland *et al.*, 1980; Minty *et al.*, 1981; Ponte *et al.*, 1983; Yaffe *et al.*, 1985). Furthermore, the 3' untranslated regions of human skeletal, cardiac, ß and γ-actin mRNAs all hybridised to the corresponding genes of rodents (Ponte *et al.*, 1984). However, in the case of

the chicken α-smooth muscle actin gene (Carroll *et al.*, 1986), probes containing the 3' untranslated region did not hybridise to any sequences in human DNA, suggesting greater species divergence for this isoform. The biological significance of the conservation of the 3' untranslated regions in the majority of the genes is unclear, although a structural or a regulatory role has been suggested. If they are important in such ways, however, it is difficult to account for the apparent non-similarities of the 3' untranslated region of the α-smooth muscle actin genes.

Similar patterns of isoformic specificity are also observed with the 5' untranslated regions. Comparison of the 5' untranslated region of the human (Ponte *et al.*, 1984; Ng *et al.*, 1985) and rat ß-actin genes (Nudel *et al.*, 1983) revealed 80% homology, indicating a considerable conservation of this region of the gene. However, comparison of the 5' untranslated region of chicken α-actin with chicken and rat skeletal α-actin and ß-actin genes did not reveal any regions with substantial homology. A small but significant homology exists in the promoter regions for the skeletal α-actins (19 out of 20 nucleotides) and the promoter regions for ß-actins (21 out of 25 nucleotides) in chicken and rat (Kost *et al.*, 1983; Ordahl and Cooper, 1983; Eldridge *et al.*, 1985).

1.2 Pseudogenes

1.2.1 Introduction

Pseudogenes are defined as DNA sequences with significant homology to functional genes, but possessing mutations that would prevent them expressing a functional product. Such mutations can, for example, cause premature termination of translation; the formation of polypeptides with little homology to the functional gene product; interference with transcriptional initiation; and interference with the processing of RNA transcripts. There are two types of pseudogenes, namely, duplicative and processed pseudogenes.

Pseudogenes of the first type, the duplicative pseudogenes, are thought to arise from duplication and divergence of functional genes, with silencing of one of the two copies. Therefore they are closely linked to their functional counterparts and, in the case of genes encoding proteins, retain the intervening sequences of the functional gene. The second type, the processed pseudogenes, apparently arose through incorporation of mRNA reverse

transcripts into the genome, probably at staggered breaks in the chromosome. Therefore they lack intervening sequences and have oligo(A) tracts at their 3' ends. Although processed pseudogenes may have intact coding regions, they can still be classed as pseudogenes by virtue of their transcriptional silence, a consequence of their mRNA origins.

In addition to processed pseudogenes corresponding to mRNAs, there are non-duplicative pseudogenes corresponding to RNA transcripts which serve non-coding functions. Although these are not necessarily 'processed', they are best considered with processed pseudogenes because of their similar origin. These pseudogenes include ones corresponding to small nuclear RNAs (snRNAs) and 7SL RNA (Denison and Weiner, 1982; Ullu and Tschudi, 1984).

1.2.2 Duplicative Pseudogenes

These pseudogenes are thought to arise by tandem duplication of functional genes, to which they are usually linked. Thus they show the major structural features of expressed genes, such as recognisable promoters, exons, introns, and RNA processing sites.

(i) The *Xenopus* 5S rRNA Pseudogene

The first gene-like sequence to be named a 'pseudogene' was that corresponding to the *Xenopus laevis* 5S rRNA gene (Jacq *et al.*, 1977). This pseudogene sequence is located downstream from the functional gene for 5S rRNA, and is part of the 700 nucleotide repeat unit that is expressed during oogenesis. The pseudogene lacks the last 20 base pairs from the 3' end of the functional gene (totalling 121 base pairs) and otherwise differs by 9 base changes (Miller *et al.*, 1978). Although RNA corresponding to this pseudogene is not found *in vivo*, a high rate of transcription (85% of that of the functional gene) can be achieved when the pseudogene is micro-injected into the *Xenopus* oocyte. However, most of the transcripts produced do not terminate correctly at the 3' end of the gene, but are of varying greater lengths and are unstable. Thus the apparent absence of transcripts *in vivo* may reflect a defect in the termination, rather than the initiation, of transcription (Miller and Melton, 1981).

(ii) Evolutionary Behaviour of Duplicative Pseudogenes

The α and ß-globin gene families of a variety of mammals provide typical examples of duplicative pseudogenes at different stages of their evolutionary decay and of the variety of processes by which the different gene clusters have evolved. All the globin pseudogenes, with the exception of two mouse α-globin pseudogenes (Leder *et al.*, 1981), are found linked to their functional counterparts. This is consistent with the origin of these pseudogenes from duplicated genes formed within the gene clusters, which have since diverged and become inactive.

Once a duplicated gene becomes inactive, it should be free from all selective constraint and then rapidly accumulate mutations at a rate characteristic of non-coding sequences. It is conceivable that certain pseudogenes still retain some regulatory functions within their parental gene cluster, although so far there are no data to support this suggestion. Another possibility is that, a silent gene may by chance undergo a reversion mutation and be reactivated (Ohno, 1970). This process does occur with 'cryptic' genes in bacteria, which under strong selective pressure can revert to a functional state (Hall, 1983). Such events are, however, likely to be rare, since defects accumulated in pseudogenes often include deletions and insertions which are most unlikely to undergo reversion. Reactivation of a pseudogene can occur by another mechanism termed gene conversion, which could lead to the 'correction' of a pseudogene by replacing a defective gene segment with functional sequences from a neighbouring gene (Jeffreys *et al.*, 1983). Gene conversion involves the non-reciprocal copying of information from one gene to another homologous gene within a cluster, as the result of inter- or intrachromosomal exchange (Lauer *et al.*, 1980; Slightom *et al.*, 1980). A number of instances of gene conversion have been detected among the α and ß-globin genes (Slightom *et al.*, 1980; Shen *et al.*, 1980; Leibhaber *et al.*, 1981; Schon *et al.*, 1982).

The evolutionary time spent by each pseudogene, first under selection as a functional gene, and then without selection as a pseudogene, has been estimated from the percentage of silent and replacement base changes in the coding sequence compared to the functional gene (Perler *et al.*, 1980). These estimations have assumed that pseudogenes accumulate mutations at the same rate as silent changes in functional genes. However, it seems that there may be some selective pressure against changes, even between synonymous codons in

functional genes. Thus, it has been reported that the rate of nucleotide substitution in globin pseudogenes is approximately twice the rate of substitutions in the third codon position in functional genes (Miyata and Yasunaga, 1981; Miyata and Hayashida, 1981; Li *et al.*, 1981). A further factor that has to be taken into account is the gene conversion events that may mask the true evolutionary age of genes or pseudogenes. For example, comparison of the coding regions of the two human adult globins, δ and ß, suggests that they arose from a duplication event not more than 40 million years ago (Spritz *et al.*, 1980; Efstratiadis *et al.*, 1980). However, various non-coding regions, the second intervening sequence, the mRNA 3' untranslated region, and the 5' sequences upstream of the CAAT box , all appear to have diverged over a much longer period of time (Martin *et al.*, 1983; Hardies *et al.*, 1984). Thus the δ-globin coding region appears to have undergone a recent conversion by the ß-gene, which has covered the traces of its more ancient origin. Reliable estimates of evolutionary divergence times can, therefore, only be derived from those regions of the gene that have not been subject to gene conversion.

(iii)Human α-Globin Pseudogenes

Besides the two active embryonic (ζ) and adult (α1, α2) genes, the human α-globin gene clusters also contains two pseudogenes, ψζ and ψα. Pseudogene ψζ is more than 99.5% homologous in its coding region to the functional ζ-globin gene and has a single deleterious mutation, a termination codon in its first exon; suggesting that the formation of this pseudogene is only recent (Proudfoot *et al.*, 1982). On the other hand, pseudogene ψα is only 75 to 80% homologous to the functional α-globin genes and has a considerable number of mutations. These include base substitutions that cause missense codons that affect translation and mRNA processing; deletions that cause frame shifts in the coding sequence, and that alter the spacing between CAAT and TATA boxes in the transcription promoter region (Proudfoot and Maniatis, 1980).

Comparative studies of the sequences surrounding the ψα pseudogene and the two functional genes α1 and α2 suggest that they arose by gene duplication and subsequent unequal crossing over (Lauer *et al.*, 1980; Proudfoot and Maniatis, 1980). Such events still appear to be operating in contemporary human populations, since chromosomes carrying either a single functional α-globin gene, or an α-globin gene triplication have been reported (Orkin *et al.*, 1979; Higgs *et al.*, 1980; Goosens *et al.*, 1980). Since the formation of the ψα, α1,

α2 cluster, the two functional genes have been maintained closely homologous by gene conversion events, while ψα has accumulated changes to become a pseudogene. Sequences in the intergenic regions upstream of α1 and α2 show strong homology and have been implicated in gene conversion, whereas their absence upstream of ψα may explain why it has apparently not been a subject to conversion (Proudfoot and Maniatis, 1980).

(iv) ß-Globin Pseudogenes

Detailed DNA sequence analysis of the ß-like pseudogenes from man and a number of the other primates has shown that a ψß gene is found in all primates and that this gene has probably been a pseudogene for the whole of primate evolution, suggesting that the ancestral primate ß-globin genes cluster comprised a five gene set, ε–γ–ψß–δ-ß (Chang and Slightom, 1984; Harris *et al.*, 1984). Comparison of the ψß sequences with those of the other ß-globin genes showed that the primate ψß gene is most closely related to the ε gene of goats (Goodman *et al.*, 1984). Both the primate ψß pseudogene and the embryonically expressed goat ε gene appear to be derived from a common ancestral gene, named η, that is distinct from the ε, γ, δ, and ß ancestral genes. The mouse and rabbit ß-globin gene clusters lack η-like genes and are thus derived from an ε–γ–δ-ß four gene set. The goat ß-genes, however, lack descendants of the γ-type gene, and are derived from a triplicated ε–η–δ–ß set of genes. Only in primates have descendants of all five types of ancestral genes been retained. It is curious that the descendants of the ancestral δ-type gene, mouse $ßh_2$ and $ßh_3$ pseudogenes, rabbit ψß2, goat $ψß^X$ and $ψß^Z$, and the minor adult δ-globin gene of primates have all shown a tendency to become silent.

The evolutionary history of the primate δ-globin gene is of particular interest. Comparative studies with other adult ß-genes show that although its 5' end has been subject to a relatively recent gene conversion by the ß-gene, its 3' end still bears significant homology to the pseudogenes of mouse ($ßh_2$) and rabbit ($ψß_2$). It appears that the δ-globin gene was originally a pseudogene and it became reactivated in the early primate lineage by a gene conversion with the adult ß-gene (Martin *et al.*, 1983). Although the δ-globin gene is still active in man, it has become silent in the Old World monkeys and will presumably evolve into a pseudogene. Thus, the δ-globin gene illustrates the possibility of both loss of activity and reactivation for the products of a gene duplication.

1.2.3 Processed Pseudogenes

Processed pseudogenes are quite distinct from the duplicative pseudogenes in bearing evidence of generation from RNA (Sharp, 1983). They are widely dispersed in the genome, and, where this has been examined, are generally on different chromosomes from their parents. Processed pseudogenes corresponding to sequences encoding proteins resemble DNA copies of mature mRNA in some or all of the following characteristics : extending only from the 5' CAP site to the site of polyadenylation; possessing a polyA tail in the expected position following a polyadenylation/processing signal; lacking all introns present in the functional gene; being flanked by direct repeat sequences (typically 11 to 15 base pairs in length) immediately preceding the transcriptional start and immediately following the polyA tail (Jeffreys and Harris, 1984). These characteristics clearly indicate that these pseudogenes originated from spliced polyadenylated mRNAs, DNA copies of which were inserted at sites in the chromosome created by a staggered endonucleolytic break. Most of such processed pseudogenes have been inserted at sites at which the lack of a promoter renders transcription impossible, and most have accumulated mutations that would, in any case, render any transcripts functionless (Vanin, 1984).

(i) Structural Characteristics

Processed pseudogenes corresponding to sequences encoding proteins may be further divided into two types. The first type are those which are more or less colinear with normal mRNAs, starting at the 5' mRNA CAP site and ending in an A-rich or oligoA stretch of 7 to 36 nucleotides, and are flanked by direct-repeat sequences of 9 to 25 bases. Processed pseudogenes of the second type, although clearly derived from RNA molecules, since they lack intervening sequences found in parent genes and end in oligoA or A-rich tracts; differ significantly from the normal cellular mRNAs of their parent genes.

A common feature of the members of the first class of processed pseudogene is that they correspond to mRNAs expressed in undifferentiated tissues. The first example of this type was a human ß-tubulin pseudogene (Wilde *et al.*, 1982b). Other examples include pseudogenes corresponding to the genes for the mouse cytochrome c (Limbach and Wu, 1985), cellular tumour

antigen p53 (Benchimol *et al.*, 1984; Zakut-Houri *et al.*, 1983), and ribosomal proteins L7 (Klein and Meynhas, 1984), L18 (Peled-Yalif *et al.*, 1984), L30 (Wiedemann and Perry, 1984), and L32 (Dudov and Perry, 1984); rat α-tubulin (Lemischka and Sharp, 1982) and cytochrome c (Scarpulla and Wu, 1983) ; and human metallothionein (Karin and Richards, 1982), nonmuscle tropomyosin (MacLeod and Talbot, 1983), γ-actin (Leube and Gallwitz, 1986), ß-actin (Moos and Gallwitz, 1982, and 1983), dihydrofolate reductase (Chen *et al.*, 1982; Masters *et al.*, 1983), argino-succinate synthetase (Freytag *et al.*, 1984), glyceraldehyde-3-phosphate dehydrogenase (Benham *et al.*, 1984; Hanauer and Mandel, 1984), and *c-ras* (McGrath *et al.*, 1983; Miyoshi *et al.*, 1984). There are, in fact, different rat cytochrome c (Scarpulla and Wu, 1983) and human ß-tubulin (Lee *et al.*, 1983) processed pseudogenes corresponding to mRNAs with 3' untranslated regions of different lengths.

The examples of the second type of processed pseudogene are less numerous. They include: (1) a human immunoglobulin λ light chain pseudogene (Hollis *et al.*, 1982), containing spliced J and C regions, but no V region (which is the present in the functional gene); (2) a human immunoglobulin ε heavy chain pseudogene (Ueda *et al.*, 1982; Battey *et al.*, 1982), comprising only the four spliced exons of the ε constant region, but no variable region coding elements (V, D, or J regions); (3) a mouse myosin light chain pseudogene (Robert *et al.*, 1984), consisting of five terminal exons common to both myosin alkali light chains LC1 and LC3, and lacking either of the two combinations of N-terminal exons normally present in the corresponding cellular mRNAs; (4) a mouse α-globin, α-ψ3, which extends at least 350 nucleotides 5' to the transcriptional start site (Vanin *et al.*, 1980; Nishioka *et al.*, 1980). The common feature of the processed pseudogenes of this class is that they correspond to mRNAs expressed only in specific differentiated cells.

The immunoglobulin J-C_λ and C_ε pseudogenes end in A-rich tracts of $(CA_X)_Y$ or $(GA)_X$, whereas the myosin light chain pseudogene has a short oligoA tract preceding an A-rich sequence. The pseudogenes are flanked by direct repeat sequences with the exception of the mouse α-globin α-ψ3 pseudogene. All these pseudogenes are truncated at their 5' ends relative to their parent genes, with the exception of the mouse α-ψ3 pseudogene. Thus they appear to have arisen from transcripts that initiated anomalously in the intervening sequence immediately upstream of those exons found in the pseudogene. The mouse α-ψ3 pseudogene also appears to be derived from an aberrant transcript,

initiated at a promoter upstream of the usual transcriptional start position.

(ii) Small Nuclear RNA Pseudogenes

Small nuclear RNAs (snRNAs) are a family of abundant discrete RNAs found associated with proteins in ribonucleoproteins in the nuclei of eukaryotes. A number of snRNA species have been identified (U1 to U6), and each hybridises to 10^2 to 10^3 sequences in the mammalian genome. The majority of these appear to be pseudogenes by the criterion of multiple base substitutions (Denison *et al.*, 1981; Hayashi, 1981; Lund and Dahlberg, 1984).

Different classes have been identified in snRNA pseudogenes according to their structural characteristics. Members of the first class of snRNA pseudogenes show significant homology to functional snRNA genes in their flanking regions, thus suggesting that they were generated by divergence of duplicated snRNA genes (Denison and Weiner, 1982). Members of the second class of more common snRNA pseudogenes were generated (see below) by the incorporation of reverse transcripts into the genome at either blunt or staggered chromosomal breaks (Van Arsdell *et al.*, 1981). These pseudogenes are characterised by containing sequences corresponding to snRNA molecules themselves. Their homology with snRNA genes begins precisely at the 5' end and extends either to the 3' end of the snRNA, or else they show some degree of 3' truncation. A number of these pseudogenes are flanked by short direct-repeats and contain short 3' terminal A-rich segments (Hayashi, 1981; Ohshima *et al.*, 1981; Monstein *et al.*, 1983; Nojima and Kornberg, 1983; Denison and Weiner, 1982). Since functional snRNA genes have a conserved 3' flanking sequence that is not A-rich, such pseudogenes must have been derived from aberrantly polyadenylated molecules (Ohshima *et al.*, 1981; Manser and Gesteland, 1982; Watanbe-Nagasu *et al.*, 1983).

(iii)Origins of Processed Pseudogenes

Processed pseudogenes are widespread in most individuals of a species in which they occur, and are transmittable as inheritable components in the genome. Hence they must have originally arisen in cells of the germ line. It follows from this that processed pseudogenes would be expected to be formed only from those genes that are expressed in germ line cells. Indeed, those processed pseudogenes that are essentially colinear with cellular mRNAs do

seem to be derived either from 'housekeeping' genes common to all cell types (*eg.* tubulins, cytoplasmic actins) or from genes that might be preferentially expressed in the germ line (*eg.*, tumour antigen p53, *c-ras*).

In contrast, those processed pseudogenes that appear to be derived from aberrant transcripts originate from genes that are not normally expressed in the germ line, since they encode products of highly differentiated somatic cells (*ie.*, lymphocyte immunoglobulin chains, erythrocyte α-globin, and muscle myosin light chain). Presumably, the aberrant nature of the transcripts from which they appear to be derived is a reflection of their abnormal transcription in the germ line.

The processed pseudogenes corresponding to human actin genes further emphasise the point that processed pseudogenes are usually only found corresponding to mRNAs that are expressed in the germ line. Processed pseudogenes seem to account for a large part of the genomic sequences hybridising to cytoplasmic ß and γ-actin cDNA probes (see section 1.1.2). In contrast, the α-cardiac and α-skeletal muscle actins, products of differentiated somatic tissues, are encoded by single copy genes with no related processed pseudogenes (Ponte *et al.*, 1983).

The overwhelming majority of processed pseudogenes have been found in mammals, examples from other vertebrates and invertebrates being very few indeed. A single calmodulin processed gene has been reported in chickens (Stein *et al.*, 1983), and some at least of the histone orphons of sea urchins appear to be derived from reverse-transcribed mRNAs (Liebermann *et al.*, 1983). In addition, the F elements of *D. melanogaster* appear to be dispersed by the integration of polyadenylated RNA transcripts, and hence the mechanism of their origin formally resembles that of processed pseudogenes (DiNocera *et al.*, 1983). Thus, although the mechanisms responsible for the generation of processed pseudogenes may not be exclusive to mammals, some feature of the metabolism of the mammalian gametes must make them peculiarly susceptible to the formation of processed pseudogenes.

(iv) Evolutionary Divergence and Possible Expression of Processed Pseudogenes

Unlike duplicative pseudogenes, which may show as little as 75% homology to their parent gene, most processed pseudogenes analysed seem to show very high (90 to 99%) homology to the genes from which they derive.

This suggests that they have arisen relatively recently in evolutionary history. For example, the myosin light chain pseudogene, which shows 99% homology to the functional gene, is found in *Mus musculus* but not the related species *Mus spretus*, which diverged less than 7 million years ago (Robert *et al.*, 1984). Similarly, the human dihydrofolate reductase pseudogene, hDHFR-ψ1, which has perfect homology to the functional gene, is only present in certain individuals of the species and shows an imbalance in its frequency in different racial groups (Anagnou *et al.*, 1984). However, as the processed pseudogenes studied to date have been detected and isolated using DNA hybridisation probes, there may have been bias towards those that are little diverged from their parent genes, especially when the probes were used under high stringency conditions. When genomic blots have been performed under low stringency conditions, genomic sequences with weaker homology to the probe have been detected (Lee *et al.*, 1983; Minty *et al.*, 1983; Wilde *et al.*, 1982a). Hence mammalian genomes may in fact contain whole series of processed pseudogenes varying quite widely in their divergence from their parental genes.

Processed pseudogenes have generally been assumed to be transcriptionally inactive since the time of their formation. Consistent with their inertness, pseudogenes may show a higher degree of DNA methylation than their functional counterparts (Lund and Dahlberg, 1984; Dudov and Perry, 1984). With the exception of the mouse α-ψ3 globin pseudogene, which retains upstream RNA polymerase II promoter sequences (Vanin, 1984), other processed pseudogenes analysed have generally lacked their original transcriptional promoters. Nevertheless it is possible to envisage integration occurring correctly downstream of a sequence that happens to correspond to that of a RNA polymerase II promoter. Apparent examples of this are a chicken processed calmodulin pseudogene (Stein *et al.*, 1983), and the rat preproinsulin I gene (Lomedico *et al.*, 1979; Soares *et al.*, 1985).

(v) Models for the Generation of Processed Pseudogenes

The basic mechanism for the formation of a processed pseudogene has been taken as the insertion of an mRNA or its cDNA copy into a staggered break in chromosomal DNA and subsequent repair of single stranded regions. Although this mechanism is widely accepted, it is difficult to define in greater detail the precise series of molecular events that gave rise to these

pseudogenes, since the only information concerning their mechanism of origin derives from the organisation of sequences flanking them.

Any model for the formation of processed pseudogenes must answer the following questions : What is the polymerase responsible for reverse transcription? How is the reaction primed? Where and how do the insertions occur in the genome? Is the inserted molecule an RNA or a cDNA (or an RNA-cDNA heteroduplex)?

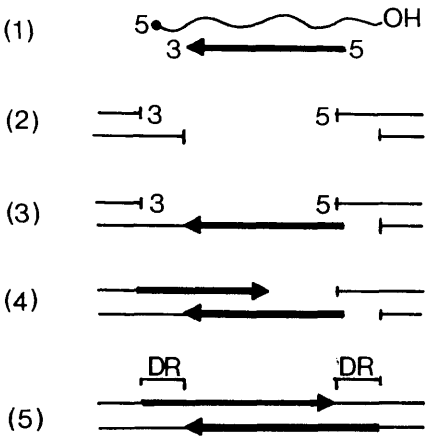
The source of the reverse transcriptase is difficult to decide as no such enzyme is detectable in normal cells. One possibility is that it is some secondary activity of a normal cellular DNA polymerase, since human DNA polymerase ß can copy synthetic RNA templates *in vitro* (Weissbach, 1977). A second possibility is that it derives from an endogenous retrovirus (Bernstein *et al.*, 1983), a supporting argument being that invertebrates, which are not thought to be subject to retroviral infection, generally lack processed pseudogenes. The weakness of this latter argument can be seen when one considers the paucity of chicken processed pseudogenes. A more attractive possibility is that the reverse transcriptase derives from the enzyme encoded by endogenous transposable elements such as the L1 family (see section 1.3.4ii).

The sites into which processed pseudogenes and other retroposons have integrated are frequently relatively AT-rich. Since such sequences are more prone to local melting of DNA strands and perhaps, therefore, to strand breakage, they might be expected to be a common source of sites for pseudogene insertion. It has also been suggested that DNA topoisomerase plays an important role in generating the transient breaks in DNA between which the insertion may occur (Van Arsdell and Weiner, 1984).

Questions concerning the primer for reverse transcription and the nature of the inserted molecule will be discussed together in comparing different models that have been proposed to account for the formation of processed pseudogenes (Figure 1.1). The first model was proposed for snRNA pseudogenes (Van Arsdell *et al.*, 1981) and suggested the following sequence of events: (1) synthesis of a cDNA copy of the snRNA; (2) covalent linkage of the cDNA 3' end to a 5' overhang of a staggered chromosome break; (3) second strand cDNA synthesis primed from the recessed 3' OH of the break; and (4) ligation and repair of the ends of the break, creating flanking direct repeats (Figure 1.1a). The use of the cDNA transcript in this model eliminated the need to propose mechanisms for decapping the snRNA and for the ligation of RNA to DNA, but it does not explain how the synthesis of the first cDNA strand is

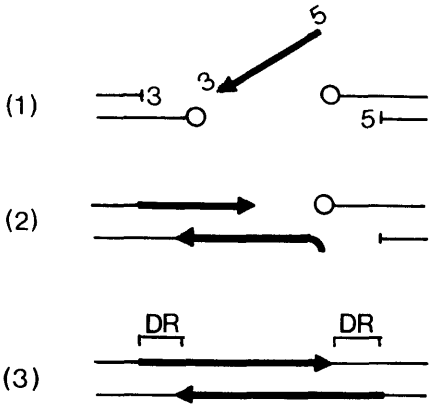
Figure 1.1 Models proposed for the formation of processed pseudogenes

(a)



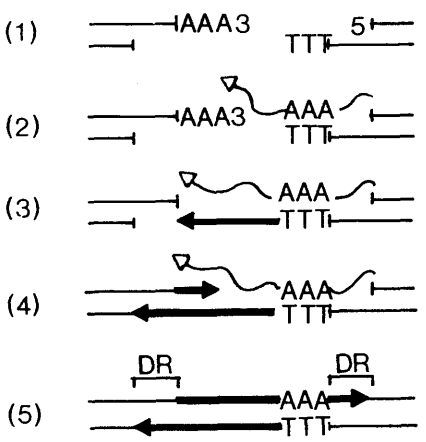
(1) Reverse transcription of RNA
(2) Staggered break in DNA
(3) Ligation of cDNA
(4) Second strand synthesis
(5) Ligation and repair

(b)



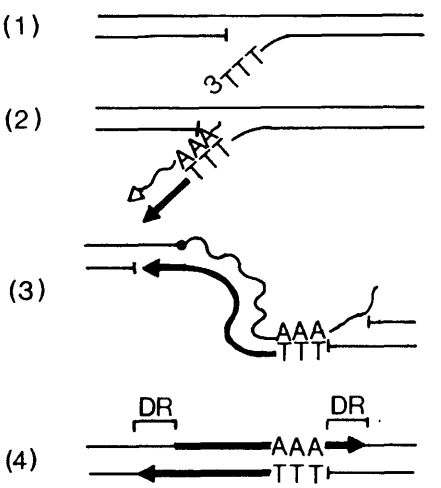
(1) Ligation of cDNA to topoisomerase activated break
(2) Second strand synthesis
(3) Downstream homology to cDNA blocks strand extension, ligation and repair

(c)



(1) Break in DNA at AT-rich region
(2) Hybridisation of T-rich end to mRNA polyA tail
(3) cDNA synthesis
(4) Second strand synthesis
(5) Ligation and repair

(d)



(1) Nick in DNA tailed with T-rich sequence
(2) Hybridisation to mRNA, cDNA synthesis
(3) Ligation to second DNA nick
(4) Replacement of RNA, ligation and repair

The four models proposed to account for the formation of RNA-derived pseudogenes are shown. Thin wavy lines represent RNA and thick lines represent new DNA (cDNA and second strand or repair DNA synthesis). Flanking direct repeats resulting from the insertion are indicated by parenthesis, and topoisomerase molecules by open circles. (a) and (b) are 'cDNA insertion' models for the generation of snRNA pseudogenes (Van Arsdell *et al.*, 1981; Van Arsdell and Weiner, 1984), (c) is a 'primed insertion' model for mRNA derived pseudogenes (Vanin, 1984), and (d) is a general retroposon insertion model (Rogers, 1985). For clarity, the second nick and its ligation to the RNA are shown as occurring after the first nick and cDNA synthesis, but they could occur concurrently.

primed. For some severely truncated snRNA pseudogenes, this presents no problem since the snRNAs from which they derive can act as a self-priming templates for reverse transcriptase. However, it is unsatisfactory to extend this model to processed pseudogenes that are full-length copies of the mRNAs, as it is necessary to invoke some exogenous T-rich primer molecule for synthesis of the first cDNA strand.

The minimal 'cDNA insertion' model has been elaborated to involve topoisomerase in the formation of staggered or blunt chromosomal breaks (Van Arsdell *et al.*, 1981; Van Arsdell and Weiner, 1984; Figure 1.1b). In addition, it was suggested that homology between the downstream direct repeat sequence and the incoming cDNA molecule might be instrumental in anchoring the cDNA relative to the staggered break (Moos and Gallwitz, 1983). This would account for the observation that flanking direct repeat sequences frequently overlap the 3' end of truncated U2 snRNA pseudogenes or the 3' oligoA or A-rich tails of full-length snRNA and processed pseudogenes.

This latter observation also points to a more attractive alternative model, which to a large extent overcomes the difficulty of 'cDNA insertion' (Figure 1.1c). The overlap between the 3' ends of pseudogenes and their flanking direct repeats suggests that 3' overhangs at staggered chromosomal breaks might themselves act as the primers for the initial cDNA synthesis by virtue of their partial homology to an RNA. Thus this model (Figure 1.1c) combines the two steps of cDNA synthesis and cDNA insertion. Since the cDNA molecule is primed by a single-stranded region of the genomic DNA itself, it is necessarily already linked into the chromosome. Subsequent steps would involve replacement of the RNA to generate a double-stranded cDNA and repair and ligation of the ends (Vanin, 1984).

A variation of this 'primed insertion' theme has been suggested by Rogers (1985). In this model (Figure 1.1d), a nick in chromosomal DNA becomes tailed with T-rich sequences, which then act as a primer for cDNA synthesis. To ensure complete copying of the mRNA, the 5' end of the inserted RNA is ligated to a second nick in the target DNA and repair synthesis completes the process to generate a DNA copy flanked by direct repeats.

Thus, several models have been proposed to account for the formation of processed pseudogenes. It appears that no one mechanism is likely to be universal, and the variety of pseudogenes structures and flanking 'tail' and repeat sequences probably reflects a variety of ways in which sequences contained in RNA may be reintroduced into the genome.

1.3 Transposition in Eukaryotes

1.3.1 Transposable Elements

The observation of an unstable phenotype in maize (McClintock, 1951) led to the first proposal of the existence of mobile genetic elements. Such mobile genetic elements are now more generally termed 'transposable elements', or transposons. Transposable elements were later detected in bacteria by the mutations they produce on moving to positions within a functional gene, and it was bacterial transposons that were first characterised in molecular detail. All bacterial transposable elements contain a gene coding for a transposase, which is required for the integration of the transposon into the target DNA. Certain classes of bacterial transposons (*eg.* Tn3), also contain a second gene which codes for a site-specific recombinase, the resolvase (Heffron, 1983; Grindley, 1985). More recently there has been marked progress in the characterisation of eukaryotic transposons. In eukaryotes, as well as transposons (*eg.* maize elements) that undergo DNA-mediated transposition, apparently analogous to that in bacteria, there are several examples of transposons in which transposition is RNA-mediated. Some examples of these eukaryotic transposable elements will be discussed in this section.

Transposable elements are defined as stretches of DNA that are flanked by specific terminal DNA sequences and that have the ability to move to new DNA sites with little or no specificity for the latter. They are usually identified by the effects they have on the structure or function of the genome. Thus they can cause interruption, deletion or inversion of regions of DNA in regulatory or coding regions of genes, and may have the ability to acquire and transpose other genomic DNA.

1.3.2 Eukaryotic Transposable Elements: DNA-mediated

The first transposable genetic elements to be discussed are the controlling elements of *Zea mays*. These were detected because they can cause unstable mutations affecting pigmentation of the aleurone layer of the maize kernel. McClintock (1951) coined the term 'controlling element' for elements of this type as they appeared to control gene expression. This occurs when a controlling element inserts within, or adjacent to, a gene and inhibits its expression. This inhibition is relieved when the controlling element is excised.

This may occur in either somatic or germline cells and may be correlated with insertion of the controlling element elsewhere in the genome. Some controlling elements can be detected even though they do not affect gene expression since they are sites of frequent chromosome breaks (McClintock, 1951).

Some controlling elements are termed autonomous or 'regulator elements' because they affect the expression of adjacent genes and mediate their own excision and/or transposition. Others are termed non-autonomous or 'receptor elements' because they can only transpose if a related regulator element is present elsewhere in the genome (Fincham and Sastry, 1974; Federoff, 1983). As regulators can promote their own excision and transposition as well as that of appropriate receptor elements, the receptors are believed to be derived from regulators by deletion.

The *Activator (Ac)* and *Dissociation (Ds)* elements constitute one of the best known family of the receptor-regulator systems of maize transposable elements (Doring and Starlinger, 1984). *Ds* is a transposable element in which the ability to move and cause chromosome breakage is dependent on the simultaneous presence of an *Ac* element elsewhere in the genome. In contrast, the *Ac* elements are able to promote their own transposition (McClintock, 1951; Federoff, 1983). Several *Ac* and *Ds* elements have been isolated (Doring *et al.*, 1984; Federoff *et al.*, 1983; Sutton *et al.*, 1984). The *Ac* elements have been shown to be structurally similar or identical to each other, whereas the *Ds* elements are quite variable in length and sequence. Most of the *Ds* elements are closely related to the *Ac* elements in structure, suggesting that they are derivatives of the *Ac* element defective in transposition (Federoff *et al.*, 1983). The *Ac* elements exhibit two properties that are common to transposable elements of many organisms: they are flanked by short direct repeats of target-site DNA (8 base pairs in this case), and possess short terminal inverted repeats (11 base pairs here).

Recent nucleotide sequence analysis indicates that *Ac* is 4.6 kb in length and contains two open reading frames encoding polypeptides of 839 and 210 amino-acids, respectively (Pohlman *et al.*, 1984). These genes in the *Ac* element are about the size of the transposase and the resolvase of the bacterial transposon Tn3. The *Ds* element is a complementable mutant of *Ac*, and it differs from the *Ac* element by a deletion of 194 nucleotides located entirely in the larger open reading frame (Pohlman *et al.*, 1984). Thus the protein which is inactivated by the deletion in the *Ds* element is probably a protein involved

in transposition, consistent with the analogy to bacterial transposons (Federoff, 1983). It is therefore reasonable to imagine that their mechanism of transposition is mediated by DNA in a similar way to that of bacterial transposons (Grindley and Reed, 1985).

1.3.3 Eukaryotic Transposable Elements : RNA-mediated

(i) Endogenous Retroviruses and Retroviral-like Elements

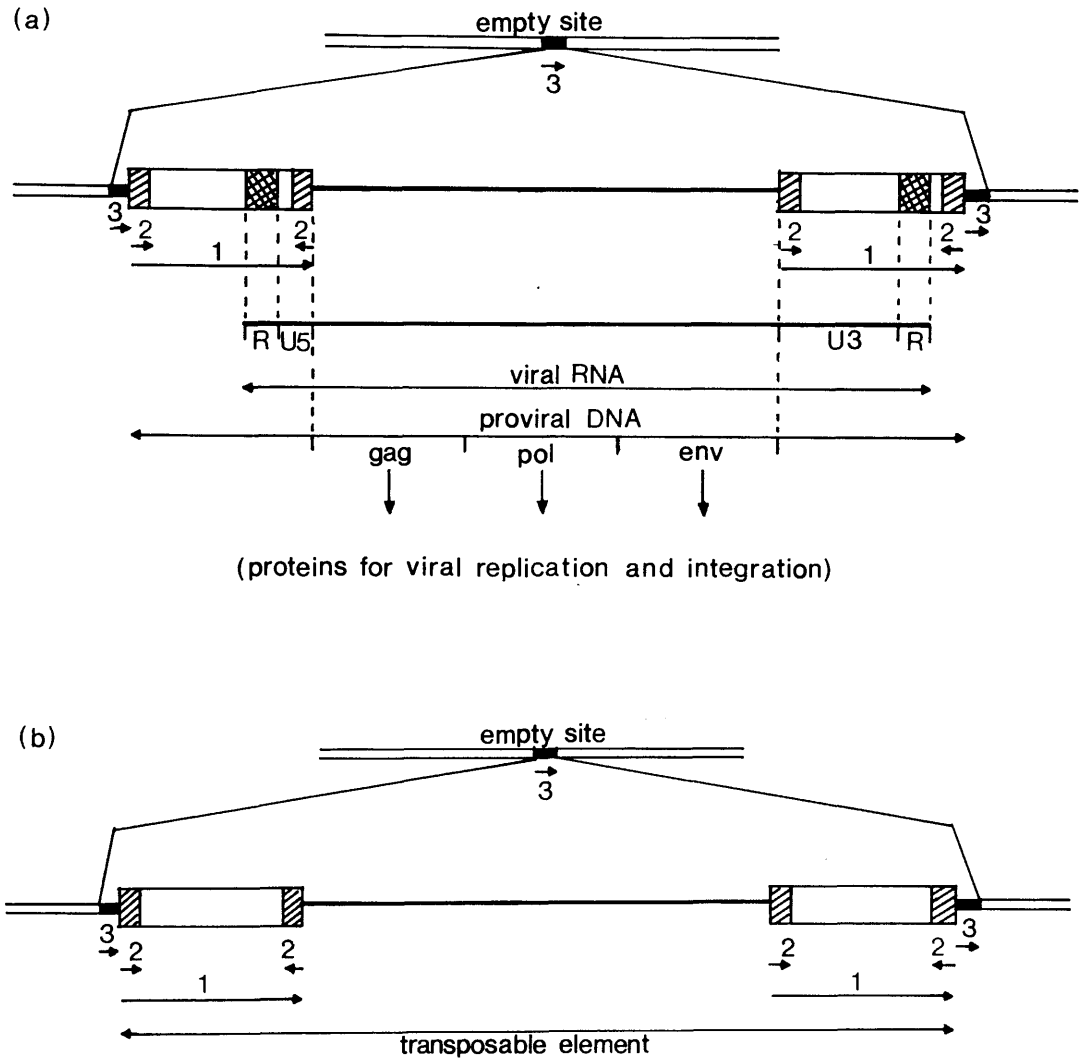
Retroviruses are RNA-containing animal viruses that replicate through a DNA intermediate incorporated into the host chromosome. Thus they have two alternative types of genomic nucleic acid, viral RNA and proviral DNA, both containing all of the genetic information of the retrovirus (Temin, 1981).

The proviral DNAs of the vertebrate retroviruses are approximately 7 kb in length, and are bounded by direct repeat sequences known as 'long terminal repeats' (LTRs) a few hundred nucleotides long (1 in Figure 1.2a). Each LTR includes a short inverted repeat sequence (5 to 13 base pairs in length) at its ends (2 in Figure 1.2a). All proviruses that have been analysed so far start with the sequence TG and end with CA (Weiss *et al.*, 1982). The proviral DNA is flanked on each side by a small number of bases (usually 5) which occur once at the site of insertion (3 in Figure 1.2a) in the absence of the element (target-site direct-repeats).

The LTRs of proviral DNA have three components, only two of which are found at the non-identical ends of the viral RNA (Figure 1.2a). At the left-hand end of an LTR is a unique sequence from the 3' end of the viral RNA. Adjacent to this is a sequence which is repeated at both ends of the viral RNA and a unique sequence from its 5' end. These are designated U3, R, and U5, respectively, in Figure 1.2a. It must be stressed, however, that the proteins required for transposition are encoded in the DNA between the LTRs of retroviruses and not in the LTRs themselves. There are three well-defined viral genes, *gag*, *pol*, and *env* (Figure 1.2a), which, respectively, encode a virion core protein, the reverse transcriptase and an envelope glycoprotein. However, certain of these genes actually encode more than one protein. Thus the *pol* gene also encodes a ribonuclease H, which is located at its 5' end, and an endonuclease which is located at its 3' end (Johnson *et al.*, 1986).

The transcription of the viral RNA from the proviral DNA is presumed to start at the beginning of R in the left-hand LTR of the provirus and terminate

Figure 1.2 Structural features of transposable elements



(a) Diagram showing the relationships between sequences in retroviral RNA and proviral DNA and between an integrated provirus and a site into which it has inserted. Putative U5, R, U3 regions are indicated. The coding domains of *gag*, *pol*, *env* are indicated on the proviral strand of DNA. The repeats are, 1, LTR; 2, short inverted repeat; 3, direct repeat of a host sequence present only once at the empty site. (b) Diagram of a *copia*-like transposable element, indicating the positions and orientations of the repeat sequences. The relationship between a full and an empty site is shown. Repeat sequences are described as in (a) above. (Modified from Figure 1 in Finnegan, 1981).

near the end of the right-hand LTR (Varmus, 1983). Putative promoter and polyadenylation signals have been found at appropriate positions in all LTRs that have been sequenced. Synthesis of the first DNA strand is primed by a host tRNA, hydrogen bonded to the viral RNA by its 3' terminal 18 nucleotides. The sequence to which this tRNA hybridises, the primer binding site, can be found two base pairs to the right of the left-hand LTR. A conserved purine-rich sequence adjacent to the right-hand LTR is believed to be involved in priming synthesis of the second DNA strand.

After infection, viral RNA molecules are first reverse-transcribed in the cytoplasm into linear double-stranded DNAs flanked by an LTR at each end. Following migration of the DNA into the nucleus, two prominent circular species appear with either one LTR, or two LTRs in tandem. The former species may be generated by homologous recombination between the two LTRs of the linear molecule, while the latter are apparently formed by blunt-end ligation of linear molecules (Panganiban, 1985) The formation of the circular molecules with two tandem LTRs result in the concurrent formation of a retroviral *att* site, which acts as a recognition site for integration into the host DNA (Panganiban and Temin, 1984). Therefore it is likely that only the two-LTR circular form serves as precursor to the provirus. Recently, reports have indicated that the 3' end of the *pol* gene function as an *'int'* locus, a region encoding a retroviral endonuclease 'integrase' (Schwartzberg *et al.*, 1984; Panganiban and Temin, 1984), which specifically recognises the *att* site and thus mediates integration (Panganiban, 1985). Integration and resolution of the *att* site results in the loss of the middle 4 base pairs of the *att* site. This precise deletion, as well as the duplication of target DNA at the site of insertion, is likely to reflect staggered cleavage of both DNAs, followed by exonucleolytic removal of all or part of each resulting viral DNA overhang, and filling-in of each cellular DNA overhang by DNA synthesis (Panganiban, 1985).

Endogenous proviruses occur in the genomes of many vertebrates. The chromosomal locations of endogenous proviruses can differ between individuals in populations of mice and chicken (Cohen and Varmus, 1979; Hughes *et al.*, 1979) and integration of both exogenous and endogenous proviruses can cause mutations. Proviral sequences have been found associated with mutations which inhibit the expression of genes affecting the coat colour (Jenkins *et al.*, 1981) or embryonic development (Jaenisch *et al.*, 1983) of mice, and which activate a cellular oncogene, *c-myc*, in chickens (Hayward *et al.*, 1981; Payne *et al.*, 1981).

Because the endogenous retroviral sequences are carried in the germline they formally resemble transposable elements as well as being proviruses. However, there are also defective elements related to retroviral genes but lacking the possibility of an extracellular phase, and so can only be classified as transposons. These are the genes that encode the intracisternal A-particles (IAP), which are retrovirus-like structures observed in mouse oocytes and preimplantation embryos (Calarco and Szollosi, 1973; Yotsuganagi and Szollosi, 1981) and in a variety of mouse tumour cells (Kuff *et al.*, 1972; Lueders *et al.*, 1977; Lueders and Kuff, 1977). Since these particles bud from the endoplasmic reticulum and remain within the cisternae, they are not infectious. Mouse IAPs contain a major protein of 73,000 daltons (Paterson *et al.*, 1978), a reverse transcriptase (Wilson and Kuff, 1972), and polyadenylated RNA molecules (IAP RNAs) (Kuff *et al.*, 1981; Ono *et al.*, 1980; Paterson *et al.*, 1978). Morphologically and biochemically, IAPs have retrovirus-like features, but IAP RNAs have no apparent sequence homology with either type B or type C murine retroviral RNAs (Lueders and Kuff, 1980).

DNA sequences homologous to IAP RNAs (IAP genes) are present in approximately 1,000 copies per haploid genome of *M. musculus*, and these genes appear to be interspersed throughout the chromosomes (Lueders and Kuff, 1977 and 1980; Kuff *et al.*, 1983a). Sequences homologous to the IAP genes of *M. musculus* are widely distributed in most rodent species and in some other mammals (Lueders and Kuff, 1981). The majority of DNA sequences related to the IAP genes can be categorised as Type I or Type II elements, the former being about 7 times more numerous than the latter. Type I elements are about 7.3 kb long whereas Type II elements are about 4.8 kb long. However these latter appear not to be simply deletion derivatives of Type I elements (Shen-Ong and Cole, 1982).

Several cases of transposition of mouse IAP elements have been reported and these have generally involved deleted IAP genomes. In two cases, the immunoglobulin κ light chain gene have been inactivated by the insertion of an IAP element within an intron of the gene (Kohler and Shulman, 1980; Hawley *et al.*, 1982; Kuff *et al.*, 1983a). Presumably these IAP elements prevent expression of the κ chain gene. However, in a mouse myeloma, the insertion of an IAP element within the cellular oncogene, *c-mos*, has apparently caused its activation (Rechavi *et al.*, 1982; Kuff *et al.*, 1983b). In addition, an IAP genome has been found associated with one of the two renin genes of DBA/2 mice, with circumstantial evidence for gene activation in this

case (Burt *et al.*, 1984). Recently, an IAP genome was found to be inserted upstream of the putative TATA box of the interleukin-3 gene promoter in a leukaemia cell line (Ymer *et al.*, 1985). Expression studies confirmed that the IAP genome was responsible for the constitutive expression of IL-3 in this leukaemia.

No biological role has yet been attributed to IAP genes, other than the rather specious one of causing mutations. However, there is evidence that the gene for a IgE-binding factor is also a member of the *Mus musculus* IAP sequence family (Moore *et al.*, 1986). IgE-binding factors are produced by T lymphocytes and believed to regulate the production of IgE by B lymphocytes (Hirashima *et al.*, 1980; Ishizaka, 1984; Suemura *et al.*, 1980). Analysis of the coding region of the IgE-binding factor indicated that it is entirely derived from segments of the putative IAP *gag* and *pol* genes. The significance of this relationship is difficult to assess but the interesting possibility exists that some members of the mouse IAP sequence may have evolved to encode proteins with biological functions unrelated to retroviral replication.

(ii) *Copia*-like Elements in *Drosophila melanogaster*

The genome of *D. melanogaster* contains about 30 to 50 families of transposable elements, together making up approximately one-half of the moderately repetitive DNA (5 to 10% of the total genome) in this species (Finnegan, 1981). The best studied families (known as the *copia, 412, 297, mdg1, mdg3, and B104* families, after the first member of each to be studied) have several properties in common, although there is no detectable homology between them (Finnegan *et al.*, 1978; Strobel *et al.*, 1979; Ilyin *et al.*, 1980a and 1980b; Scherer *et al.*, 1982; Ikenaga and Saigo, 1982). They are usually referred to collectively as '*copia*-like elements'. The members of each family are well conserved and are located at 20 to 100 sites distributed throughout the genome. In general, there is only one element at each site but the number and location of these sites vary between strains of *D. melanogaster* and between embryonic and tissue culture cell DNA (Ilyin *et al.*, 1978; Potter *et al.*, 1979).

Copia is a 5 kb element (Finnegan *et al.*, 1978) and is flanked by two LTRs of 276 base pairs (Levis *et al.*, 1980). Evidence for its transposition is provided by mutations it causes, especially well-characterised being those involving the *white* locus. The *white* locus of *D. melanogaster* is responsible for the deposition of pigment in the eye. Insertion of a *copia* transposable

element into the small intron of the *white* gene results in a mutant (*white-apricot*) acquiring an eye colour which is considerably lighter than the red colour of the wild-type (Bingham and Judd, 1981). Furthermore, the mutations which are caused by the insertion of *copia* can be reverted by excision of the element, although, in the case of the *white-apricot* mutation, with a rather low frequency (Rubin *et al.*,1982). *In situ* hybridisation and Southern blotting experiments performed on different revertants do indicate that *copia* have been excised in these (Gehring and Paro, 1980; Goldberg *et al.*, 1982). Recent analysis from an isolated revertant of *white-apricot* has shown that *copia* was excised in such a way that one LTR has remained at the site of insertion (Carbonare and Gehring, 1985). The LTR has the same orientation as the *white-apricot* mutant and also retains the 5 base pair duplication in the flanking DNA. This structural arrangement is consistent with the hypothesis that *copia* has excised by a mechanism of intrachromosomal homologous recombination between the LTRs (Carbonare and Gehring, 1985). This mechanism of excision may also account for the occurrence of closed circular *copia* DNA molecules with one LTR only (Flavell and Ish-Horowicz, 1981), which would represent the reciprocal crossing-over product.

Copia-like elements possess many structural similarities in sequence organisation to that of proviruses of vertebrate retroviruses. They are of similar lengths to proviruses and flanked by LTR sequences a few hundred nucleotides long (1 in Figure 1.2b). The lengths and sequences of these LTRs are specific for each family of elements (Levis *et al.*, 1980; Bayev *et al.*, 1980). At the extreme ends of each element are short (about 10 base pairs) inverted repeats (2 in Figure 1.2b) which, except in the *mdg3* family, occur at both ends of the long direct-repeats. Immediately before and after each element is a short direct-repeat (3 in Figure 1.2b), the length, but not the sequence, of which is constant for all members of a particular family. Comparison between particular sites in the genome from different sources, one containing and one lacking a transposable element (the latter being referred to as an 'empty' site), indicates that the bases of the short direct repeat occur only once at the site into which an elements inserts (Dunsmuir *et al.*, 1980).

Recent sequence analysis of the *copia* element present at the *white-apricot* allele of the *white* locus in *D. melanogaster* has detected a single open reading frame of 4,227 nucleotides (Mount and Rubin, 1985). It has the potential to encode a polyprotein with several regions of homology to retroviral proteins, including good homology to a region of the *pol* gene

encoding integrase, recently shown to be distinct from reverse transcriptase and required for the integration of circular viral DNA to form proviruses (see above). However, the organisation of these coding regions within *copia* is different from their organisation in retroviruses and in the other *copia*-like element *17.6*, but more closely resembles the Ty transposable elements of yeast (see below).

(iii)Ty Elements in Yeast

Ty elements are a family of approximately 30 transposable elements, which are dispersed throughout the genome of *Saccharomyces cerevisiae* (Cameron *et al.*, 1979; Roeder and Fink, 1980; Farabaugh and Fink, 1980; Gafner and Philippsen, 1980). They consist of a central region approximately 5.3 kb in length, flanked by terminal direct repeats, known as δ sequences, about 330 base pairs long. Upon transcription, the Ty elements generate duplications of 5 base pairs in the target DNA as a consequence of integration (Farabaugh and Fink, 1980). The 5 base-pair duplications among different Ty elements have been shown to be unrelated apart from being A/T rich, indicating that there is little sequence specificity associated with transposition (Roeder and Fink, 1983).

Recent sequence analysis has detected two open reading frames in the Ty element, occupying the same relative positions as the *gag* and *pol* coding sequences in retroviruses (Hauber *et al.*, 1985). The overlap between the two open reading frames is 38 base pairs, and recent experiments (Mellor *et al.*, 1985) suggest that they are expressed as a fusion protein, possibly by a mechanism common to retroviruses. As in the case of *Rous sarcoma* virus, the production of this fusion protein would require translational frameshifting within the region of overlap between the genes (Clare and Farabaugh, 1985; Mellor *et al.*, 1985). Although putative transposition intermediates have not yet been found for Ty, it was possible to decide which transposition mechanism was used by Ty elements by following the fate of a foreign intron inserted into a galactose-promoted Ty element (Boeke *et al.*, 1985). The intron (plus its flanking exons) was from a yeast ribosomal protein gene, and was inserted under the artificial control of a yeast GAL1 promoter, so that transcription could be induced by galactose . It was found that this induced transposition and caused the intron to be correctly spliced out of the transposed copy. It is thus beyond doubt that RNA is an intermediate in the transposition of Ty. The clear implication of an RNA intermediate in Ty transposition is that yeast cells

induced for Ty transposition contain the enzyme reverse transcriptase; and, consistent with this, it has been found that there is significant amino-acid sequence homology between the Ty element and retroviral and other reverse transcriptases (Clare and Farabaugh, 1985; Hauber *et al.*, 1985; Warmington *et al.*, 1985; Mount and Rubin, 1985).

1.3.4 Retroposons

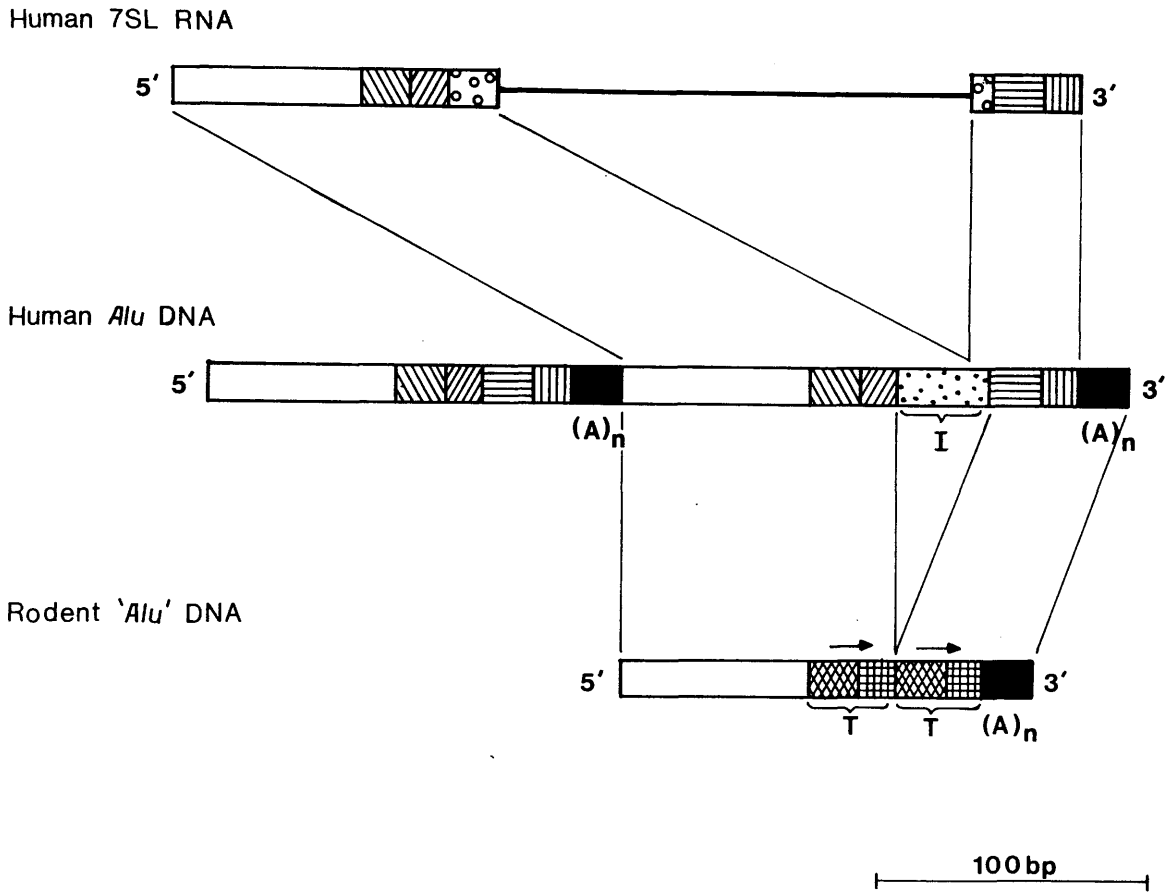
The name 'retroposon' was coined to describe certain dispersed sequences in the mammalian genome that shown common features which suggest that they are generated from RNA intermediates by a common mechanism involving reverse transcription (Rogers, 1983). These differ from retroviruses and retrotransposable elements in that they lack LTRs and, in most cases, appear to lack the ability to code the machinery of their own transposition. Processed pseudogenes fall into this category (see section 1.2.3), as do the interspersed nuclear elements described below. The interspersed nuclear elements are subdivided into the short interspersed nuclear elements (SINES), several hundred base pairs in length, and the long interspersed nuclear elements (LINES), some thousands of base pairs in length.

(i) Short Interspersed Nuclear Elements

The SINES are typified by the *Alu* family found in the primate genome, so called because most of its members contain AluI restriction sites (Houck *et al.*, 1979). Each *Alu*-repeat is about 300 base pairs in length and there are 3 to 5 X 10^5 copies, constituting 3 to 6% of the human genome (Schmid and Jelinek, 1982). The structure of *Alu* contains a head to tail tandem arrangement of two related sequences about 130 base pairs long, each terminated by an A-rich tail (Figure 1.3). The 3' tandem repeat contains an additional, internal, segment of 31 base pairs which is usually absent from the 5' repeat (Duncan *et al.*, 1981). In primates and rodents, all the genomic *Alu*-like sequences carry variable A-rich 3' tails and are flanked by direct repeats (Bell *et al.*, 1980; Krayev *et al.*, 1980), except where there is clear evidence of deletion.

Recent studies have revealed highly-significant (about 80%) sequence homology between the longer unit of the *Alu* consensus sequence and the 5' and 3' portions of the 7SL RNA. The 7SL RNA is an abundant cytoplasmic RNA, 300 base pairs in length, and forms part of the signal recognition particle

Figure 1.3 The structural features of human 7SL RNA and the consensus sequence of human and rodent *Alu* DNA



The structural relationship of human 7SL RNA to the consensus structures of human and rodent *Alu* DNA is shown. Homologous regions are indicated by identical shading. $[(A)_n]$ is the A-rich segment at the 3' end of two related tandem repeats of a head to tail dimer in human *Alu* DNA. (I) is the insert from the right monomer which is absent in the left monomer. A segment of 155 base pairs from the centre of the 7SL RNA sequence is absent from the human *Alu*-repeat unit. Arrows above the mouse *Alu* DNA indicate the position of the 32 base pair tandem duplication.

(Walter and Blobel, 1980). As shown in Figure 1.3, the central 155 base pairs of the 7SL RNA primary sequence is not represented in the *Alu*-repeat (Ullu and Tschudi, 1984). Although it has been suggested that the *Alu* sequences may have originated from something equivalent to processed pseudogenes of 7SL RNA (Gundelfinger *et al.*, 1983), it is important to emphasise that the processed pseudogenes of 7SL RNA that have been described contain the central portion lacking from *Alu* and are of the 3' truncated variety, assumed to have been generated by self-priming facilitated by a suitable 3' secondary structure (Ullu and Weiner, 1984).

Most *Alu*-elements are capable of being transcribed by RNA polymerase III and thus are likely to terminate at oligo(dT) stretches in flanking 3' DNA (Jagadeeswaran *et al.*, 1981). Transcripts would thus have the potential to hybridise to the corresponding A-rich region, which could act as a priming site for reverse transcription, leading to a cDNA copy that could be reintegrated into the genome (see section 1.2.3v). Transposed *Alu* sequences still retain their RNA polymerase III internal promoters and thus have the potential for further transposition. However, the accumulation of mutations may inactivate the promoters of some *Alu* sequences and render them incapable of further transposition.

'*Alu*-equivalent' families have been identified in the genomes of many rodents, for example in mouse (Krayev *et al.*, 1980), rat (Lemischka and Sharp, 1982) and chinese hamster (Haynes *et al.*, 1981). Comparison of the *Alu*-equivalent sequences in human and rodent shows that the *Alu*-equivalent sequences in rodents is derived from just one of the 130 base pair units found in *Alu* (Figure 1.3), but contains a tandem repeat formed by duplication of an internal 32 base pair sequence (Kalb *et al.*, 1983). This internal 32 base pair insert is, however, different from the 31 base pair insert of the human *Alu*-repeat. Disregarding these insert regions, the human and rodent *Alu*-equivalent DNA sequences show a homology of approximately 80%.

(ii) Long Interspersed Nuclear Elements

The LINEs are typified by the L1 family, which is believed to be the only major family of this type in the primate and rodent genome. Human L1 sequences (designated L1Hs) have been shown to be homologous to the L1 family of mouse (designated L1Md), and a wide variety of other mammals (Singer *et al.*, 1983; Martin *et al.*, 1984), suggesting that L1 is ancient and has

been conserved throughout the mammalian genome (Katzer *et al.*, 1985; Witney and Furano, 1984). There are 10^4 to 10^5 copies of these elements, and they constitute 2 to 3 % of the mammalian genome (Singer, 1982). Most members of the LINE families are less than the full length of 6 to 7 kb. Different segments of L1Hs and L1Md elements were cloned independently as separate sequences by digestion with particular restriction enzymes, and were originally thought to represent different repeated DNA families. These were referred to as the HindIII (Manuelidis, 1982) and KpnI (Grimaldi *et al.*, 1984) families of primates, and the BamHI (Soriano *et al.*, 1983), MIF-I (Brown and Piechaczyk, 1981), Bam-5 (Fanning, 1982), BstNI (Cheng and Schildkraut, 1980) and R-families (Gebhard *et al.*, 1982) of rodents. These separate repeat sequences were later shown to be colinear (Fanning, 1983; Bennett and Hastie, 1984).

Transposed copies of the L1 family are heterogeneous in length. Smaller versions generally have the peculiarity that they lack varying amounts of the 5' end of the full-length version, but contain the same 3' sequences terminated by an A-rich segment of variable length (Lerman *et al*, 1983; Grimaldi *et al*, 1984; DiGiovanni *et al*, 1983). This, coupled with the observation that individual L1 elements are surrounded by small (less than 15 base pair) direct-repeats, suggests that individual L1 elements are generated from RNA intermediates of different lengths and inserted at staggered breaks dispersed throughout the genome (Voliva *et al*, 1984; Grimaldi *et al*, 1984; Wilson and Storb, 1983).

Recently, several full-length L1Md elements have been isolated, and have been found to contain multiple copies of a 208 base-pair tandemly repeating region at the 5' end (Loeb *et al.*, 1986). The two examples described each contain 4 2/3 and 1 2/3 copies of this tandem repeat, the truncated 2/3 copy being the most 5' member in both cases. The full-length L1Md member has two large overlapping open reading frames of lengths 1,137 and 3,900 base pairs (Loeb *et al.*, 1986). The ratio of amino-acid replacement to amino-acid silent (R/S) site differences between the 3,900 base-pair open reading frame and a composite consensus primate L1 sequence was determined and indicated that this portion of L1 is evolving under pressure to conserve protein function. It is not yet known whether this gene encodes a protein required for transposition of L1, but the predicted product from the 3,900 base-pair open reading frame bears some similarity to reverse transcriptase (Loeb *et al*, 1986). Such an encoded reverse transcriptase activity would be consistent with the previous

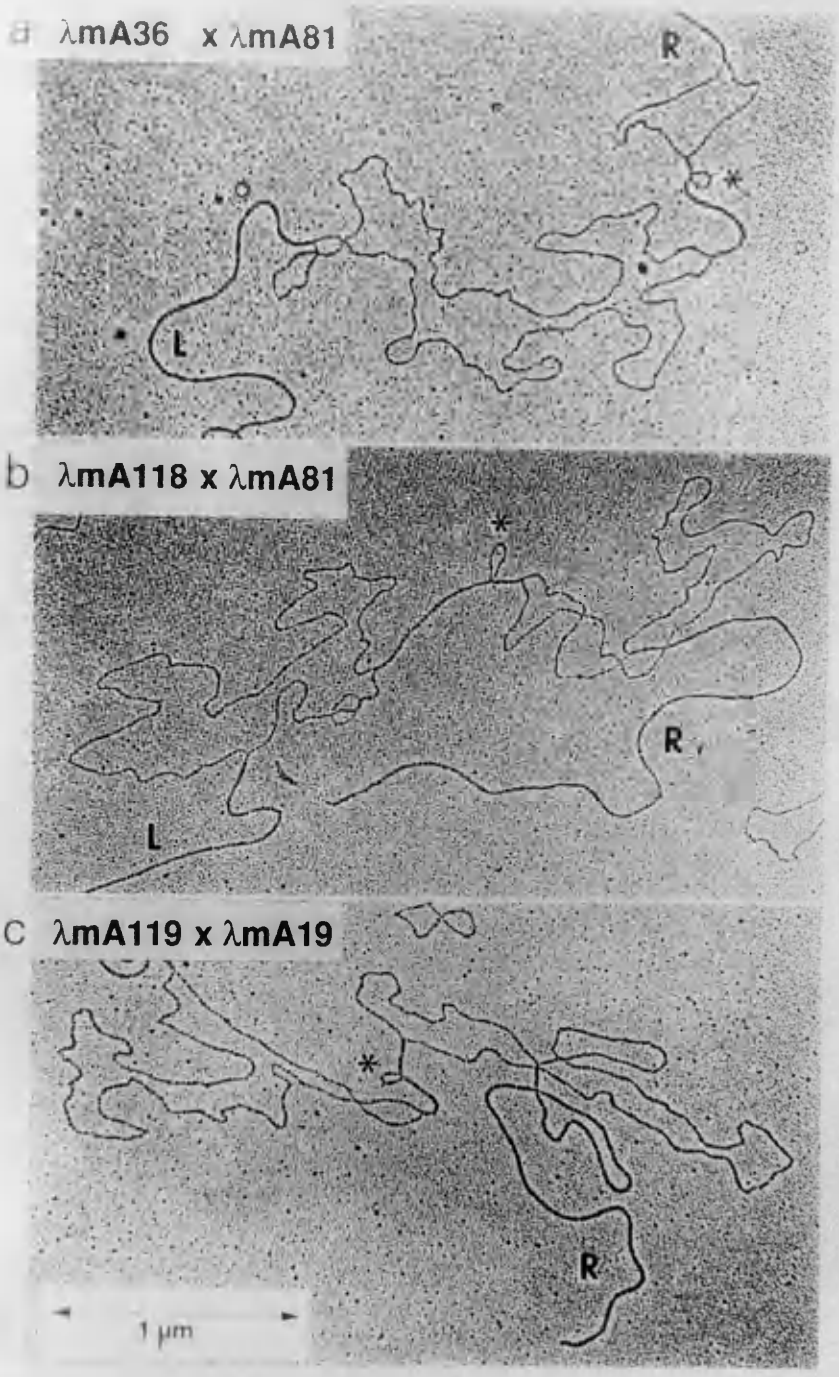
proposal that L1 is a mobile genetic element which moves via an mRNA intermediate, and would provide at least a partial explanation of the abundance of L1 elements.

1.4 Objectives of the Project

The work described in this thesis is concerned with the structure of four actin-like genes previously isolated from a bacteriophage lambda library of mouse genomic DNA. These clones were chosen for study because electron microscopic heteroduplex analysis (performed by Dr. H. Delius, EMBL, Heidelberg) had shown them to be interrupted - each at a different position - by single regions of DNA not contained in the reference cDNA-like clone. (This was actually a γ-actin processed pseudogene, λmA19, the nucleotide sequence of which has been determined by Leader *et al,* (1985), or a similar clone, λmA81, with a γ-actin pseudogene in the opposite orientation). The interpretation of these heteroduplexes in terms of the structures of the actin-like genes in λmA36, λmA82, λmA118 and λmA119 is shown in Figure 1.4.

The objective of this project was to determine the nature of these interruptions. Had they been found to be introns in functional actin genes, then attention would have focused on characterising these genes. In fact, they were found not to be introns, suggesting the interesting possibility that they might be mobile elements inserted into pseudogenes ; and the main objective became that of determining the nature of these inserted sequences.

Figure 1.4 Electron micrographs of heteroduplexes containing mouse actin-like sequences



Electron micrographs of heteroduplexes between recombinant lambda DNAs containing mouse actin-like sequences are shown. (a) Heteroduplex between λmA81 and λmA36 DNA. (b) Heteroduplex between λmA81 and λmA118 DNA. (c) Heteroduplex between λmA19 and λmA119 DNA. The double-stranded left and right arms of the lambda are labelled 'L' and 'R', respectively. The homologous actin sequences form duplexes in the insert regions, only interrupted by one deletion/insertion loop, which is marked by a star in the electron micrographs.

CHAPTER 2

MATERIALS AND METHODS

Common chemicals were AnalaR grade supplied by BDH Chemicals, Poole, Dorset, or Fisons Scientific, Loughborough, Leics. Exceptions to this are noted in the text.

Suppliers of speciality reagents :

Amersham - Amersham International, Amersham, Buckinghamshire
Boehringer - BCL, Lewes, East Sussex
BRL - Bethesda Research Labs., Paisley, Scotland
Difco - Difco Labs., Detroit, USA
Sigma - Sigma Chemical Company, Poole, Dorset

2.1 Media and Antibiotics

2.1.1 Liquid Media

All media and solutions used in the handling of nucleic acids were sterilized by autoclaving for 15 min at 15 p.s.i..

L-Broth (per litre) :
10g Bactotryptone (Difco 0123-01)
5g Yeast Extract (Difco 0127-01)
5g NaCl
Adjusted to pH 7.2 with NaOH

M9 Medium (per litre) :
6g Na_2HPO_4
3g KH_2PO_4
0.5g NaCl
1g NH_4Cl

2YT Medium (per litre)	:	16g	Bactotryptone (Difco 0123-01)
		10g	Yeast Extract (Difco 0127-01)
		5g	NaCl
10X Hogness Medium	:	6.3g	K_2HPO_4
(per litre)		0.45g	Na citrate
		0.09g	$MgSO_4.7H_2O$
		0.9g	$(NH_4)_2SO_4$
		1.8g	KH_2PO_4
		44.0g	glycerol
TE Buffer	:	10mM	Tris-HCl,pH8.
		1mM	EDTA
Lambda Diluent	:	10mM	Tris-HCl,pH7.5
		10mM	$MgSO_4$
		1mM	EDTA
Phage Buffer (per litre)	:	3g	KH_2PO_4
		8.77g	$Na_2HPO_4.2H_2O$
		5g	NaCl
		1ml	1M $MgSO_4$
		0.1ml	1M $CaCl_2$
		1ml	1% gelatin

2.1.2 Media Containing Agar

L-Agar (per litre)	:	1litre	L-Broth (pH7.2)
		15g	Agar (Difco 0140-01)
Minimal Agar (per litre)	:	100ml	10X M9 Medium
		11g	Agar (Difco 0140-01)
		1ml	1M $MgSO_4$
		1ml	0.1M $CaCl_2$
		*1ml	1M Thiamine-HCl(VitB1, Sigma)
		*5ml	40% glucose

BBL-Agar (per litre)	:	10g	BBL Tripticase (BBL 11921)
		5g	NaCl
		10g	Agar (Difco 0140-01)
		Phenol Red added to 0.02g/l	
BBL-Top Layer Agar (per litre)	:	10g	BBL Tripticase (BBL 11921)
		5g	NaCl
		6.5g	Agarose (BBL)
		*10ml	1M $MgSO_4$
		Phenol Red added to 0.02g/l	
H-Agar (per litre)	:	10g	Bactotryptone (Difco 0123-01)
		8g	NaCl
		12g	Agar (Difco 0140-01)
H-Top Agar (per litre)	:	10g	Bactotryptone (Difco 0123-01)
		8g	NaCl
		8g	Agar (Difco 0140-01)
Hammersmith Stabs (per litre)	:	9g	Nutrient Broth (Difco 0003-02)
		7.5g	Agar (Difco 0140-01)
		5g	NaCl
		*10ml	10mg/ml Thymine
		Adjusted to pH 7.2 with NaOH	

*Sterilized separately as a concentrated solution.

2.1.3 Antibiotics

These were obtained from Sigma and used at the following concentrations.

Ampicillin : A stock solution was made up of 10mg/ml of the sodium salt of ampicillin in water. It was sterilized by filtration and stored in aliquots at -20 ^{o}C . The working concentration was 30μg/ml .

Tetracycline : A stock solution was made up of 5mg/ml of tetracycline hydrochloride in ethanol/ water (50% v/v). It was sterilized by filtration and stored in aliquots at -20 ^{o} C. The working concentration was 20μg/ml.

2.2 Maintenance of Bacteria and Plasmids

2.2.1 Bacterial Strains

The bacterial strains used in this work were all derivatives of *E.coli* K12. They are listed in Table 2.1 along with the markers carried on the strains, and the source of the strain.

2.2.2 Storage of Bacteria

Most bacteria could be kept for at least one month in the coldroom on tightly wrapped L- agar plates. Similarly overnight cultures could be kept for a few weeks in the coldroom.

Long term storage of bacteria of one year was maintained in Hammersmith stabs (see section 2.1.2). A single colony was inoculated into the stab and stored at room temperature.

Any bacterial strains of significance, such as strains carrying foreign DNA as plasmids, were in addition kept as frozen stock culture in Hogness modified freezing medium (see section 2.1.1). To 2.5ml of an exponentially growing culture, 0.1 volume 10X Hogness medium was added, mixed, shock frozen in liquid nitrogen and stored at -70 ^{o}C. These bacteria remain viable for several years.

2.2.3 Storage of Plasmid and Phage DNA

Native plasmid and phage DNA were stored in TE buffer (see section 2.1.1) in a tight fitting capped Eppendorf tube. Plasmid DNA was stored at -20 ^{o}C and lambda DNA stored at 4 ^{o}C. DNA stored in this way remains stable for several years, and could be used to re-transform bacterial host cells should the need arise.

2.3 Preparation of Plasmid DNA

Two major differences between *E.coli* DNA and plasmid DNA are exploited in the method to isolate pure plasmid DNA. The *E.coli* chromosome is much larger than the DNA of commonly used plasmids. The bulk of *E.coli* DNA extracted from cells is obtained as broken, linear molecules, whereas plasmid

Table 2.1 *E.coli* strains described in this study

E.coli Strain	Genotype	Reference
DH1	F^-, rec A1, end A1, gyr A96, thi, hsdR17, Sup E44, rel A1, λ^-	Low, 1968; Meselson and Yuan, 1968.
HB101	F^-, hsd S20 (r_B^-, m_B^-), rec A13, ara-14, pro A2, lac Y1, gal K2, rps L20 (Sm^r),xyl-5, mtl-1, Sup E44, λ^-	Bolivar and Backman, 1979.
JM103	Δlac pro, thi, str A, Sup E, end A, sbc B15,hsd R4, F'tra D36, pro AB, lac Iq, ZΔM15	Messing *et al* ., 1981.
JM105	thi, rps L, end A, sbc B15, hsp R4, Δ(lac-pro AB), [F', tra D36, pro AB, lac IqZΔM15]	Yanisch-Perron *et al.*, 1985.
JM109	rec A1, end A1, gyr A96, thi, hsd R17 Sup E44, rel A1, λ^-,Δ(lac-pro AB), [F^-, tra D36, pro AB, lac IqZΔM15]	Yanisch-Perron *et al.*, 1985.
Q358	hsd R_k^-, hsd M_k^-, Sup F, Ø80^r, rec A^+	Karn *et al.*, 1980.
W8850	F^-, gal^-, str^R, Tl^R, λ^R	Allet *et al.*, 1973.
Y1090	Δlac U169, pro H^+,Δlon, ara D139, stra A, Sup F, trp C22:Tn10(pMC9)	Young and Davis, 1980.

DNA is generally extracted in a covalently closed-circular form. All methods devised here were based on three basic steps. Growth of bacteria and amplification of the plasmids, harvesting and lysis of the bacteria and lastly purification of plasmid DNA.

2.3.1 Large Scale Preparation of Plasmid DNA (Birnboim and Doly,1979)

Growth of Bacteria and Amplification of the Plasmid

An overnight culture was prepared from a single colony of transformed bacteria in L-broth supplemented with ampicillin (30μg/ml). The main culture was set up by inoculating 2 X 5ml of overnight culture into 2 X 800ml of L-broth. The bacteria were shaken at 37 °C until late log phase was reached (A_{600} = 0.8). Chloramphenicol (25mg/ml in 50% ethanol) was added to a final concentration of 165μg/ml and incubation continued a further 16 to 20 hr.

Harvesting and Lysis of Bacteria

Bacteria were harvested by centrifugation at 5000 rpm for 10min at 4 °C, and resuspended in 9.5ml of a solution containing 50mM glucose, 10mM EDTA, 25mM Tris-HCl (pH 8.0) and 0.5ml lysozyme (Sigma grade I : 40mg/ml in the same solution). After 30 min on ice, 20 ml of a solution of 0.2M NaOH, 1% sodium dodecyl sulphate was added , and the mixture left for a further 5 min on ice. Finally, 15ml of 3M sodium acetate (pH 4.8) was added and the mixture left for 1hr on ice. The high molecular weight DNA and bacterial debris were removed by centrifugation at 30,000 rpm for 30 min in a Beckman Ti60 rotor at 4 °C. Total nucleic acid was precipitated from the resultant supernatant by the addition of 0.6 volume of isopropanol and left standing for 10 min at room temperature. DNA was precipitated by centrifugation at 8000 rpm for 15 min at room temperature.

Purification of Plasmid DNA

Plasmid DNA behaves differently from *E.coli* DNA when the two are centrifuged to equilibrium in caesium chloride (CsCl) gradients containing the intercalating dye, ethidium bromide. Covalently closed circular plasmid DNA binds less ethidium bromide than linear *E.coli* DNA and therefore bands at a higher density in CsCl gradients.

The DNA was dissolved in 30ml TE (see section 2.1.1), 28.9g of CsCl and

1.8ml of ethidium bromide (10mg/ml) were added, and the solution clarified by centrifugation in a Beckman 'Table-top' centrifuge at 2000 rpm for 30 min. Having avoided the surface scum and the precipitated material, the solution was then carefully transferred to sealable tubes and centrifuged at 50,000 rpm for 16hr at 20 °C in the VTi50 rotor of a Beckman ultracentrifuge.

The DNA was visualised under UV (long wave) illumination, where two bands were usually seen. The upper band consisted of linear bacterial DNA and nicked plasmid DNA, while the lower band consisted of closed-circular DNA . The plasmid DNA was collected after piercing the tube with a 21g needle just below the band. A second CsCl centrifugation was usually performed using sealable tubes, and centrifugation at 65,000 rpm overnight in the VTi65 rotor. Plasmid DNA was isolated as described above.

Ethidium bromide was removed by repeated extraction with an equal volume of isoamyl alcohol (3 methyl,1-butanol : Koch Light). The DNA solution was diluted with four volumes of TE and precipitated with 2.5 X the total volume of ethanol. DNA was sedimented by centrifugation at 10,000 rpm for 10min at 4°C, re-precipitated, washed with 70% ethanol, and briefly dried under vacuum. The DNA was redissolved in 100µl TE.

The concentration of DNA was determined by measuring the A_{260} using the assumption that a 50µg/ml solution of DNA has an A_{260} of 1.0 when measured in a spectrophotometer cell with a 1.0cm light-path. It was then adjusted to 1mg/ml by addition of TE. A sample of 0.5µg was subjected to electrophoresis on a 1% agarose mini-gel to check the quality of the preparation.

2.3.2 Small Scale Preparation Plasmid DNA

(i) Mini-preparation of Plasmid DNA

For small scale plasmid preparations, a modification of the procedure described by Holmes and Quigley (1981) was followed. A bacterial colony containing plasmid was streaked thoroughly onto an antibiotic plate, and allowed to grow overnight.

Bacterial cells were carefully scraped off the plate and resuspended in 1ml of a solution of : 8% glucose; 50mM Tris-HCl,pH 8.0; 50mM EDTA; 5% Triton X-100 in a 1.5ml plastic Eppendorf tube, and 10µl lysozyme (20mg/ml in water) was added. The tube was transferred to a 95 °C block for 7 min, and centrifuged

for 15 min. 0.6ml portion of the supernatant was transferred to a fresh tube, 2µl of boiled RNase (1mg/ml) was added and incubated for 15 min at 37 °C, then 1µl of diethylpyrocarbonate was added and incubated for a further 10 min at 65 °C. The plasmid DNA was collected by addition of 0.24 ml 5M ammonium acetate and 0.54ml isopropanol, and precipitated by centrifugation. After washing the DNA with 0.3M ammonium acetate/70% isopropanol, it was briefly dried under vacuum and resuspended in 50µl TE. A sample of 5µl (1.0µg) was sufficient for a single restriction digest.

(ii) Midi-preparation of Plasmid DNA

This small scale method for plasmid preparation was modified from Birnboim and Doly (1979). Half antibiotic plates with transformed bacteria were prepared as above, and the scraped cells were inoculated into 100ml L-broth and grown to confluence. Bacteria were harvested by centrifugation in the 'Table-top' centrifuge for 5 min, then resuspended in 1ml of a solution containing 50mM glucose, 10mM EDTA, 25mM Tris-HCl (pH 8.0) in a 15ml Falcon tube containing 10µl of lysozyme (50mg/ml). After standing for 15 min on ice, 3ml of a solution of 0.2M NaOH, 1% sodium dodecyl sulphate was added. The mixture was left standing on ice for 15 min, followed by addition of 2.3ml of 3M sodium acetate (pH 4.8) and left to stand on ice for a further 10 min. The cellular DNA and debris were precipitated by centrifugation (10,000 rpm for 15 min at 4 °C), the plasmid DNA was precipitated from the supernatant with 0.6 volume of isopropanol. After standing for 15 min on ice, the DNA was recovered by centrifugation in the 'Table-top' centrifuge for 10 min, washed with 70% ethanol, and briefly dried under vacuum. the DNA was redissolved in 0.9ml 2.5M sodium acetate, centrifuged to remove debris, and the supernatant re-precipitated with 540µl isopropanol. After centrifugation and drying under vacuum, the DNA pellet was redissolved in 100µl TE.

The plasmid DNA was usually contaminated with tRNA, which could either be removed by passing the sample through a column of Biogel P-60 (see section 2.9.2), or by CsCl gradient centrifugation at 65,000 rpm (see section 2.3.1).

2.4 Preparation of Bacteriophage Lambda and its DNA

2.4.1 Preparation of Bacteriophage from Lytic Infection

A single colony of the bacterial host *E.coli* Q358 (Table 2.1), susceptible to infection with λ1059, was inoculated into an overnight culture of L-broth containing 10mM $MgSO_4$. A suitable dilution of λ1059 giving 10 to 100 plaques, was absorbed onto 200µl Mg treated cells containing 10mM $MgSO_4$, mixed with 3ml of BBL-top agar, poured onto a BBL-plate and inoculated overnight.

A single plaque was removed and added to 200µl of a freshly saturated overnight culture of bacterial host and left standing at room temperature for 20 min. The culture was then transferred to a fresh flask with 20ml L-broth containing 5mM $MgSO_4$. The flasks were shaken at 37 °C until lysis occurred (between 4 to 6 hr), when 1ml of chloroform was then added. After 5 min of shaking, the upper layer was sedimented by centrifugation in a 'Table-top' centrifuge for 20 min and the supernatant was transferred to a fresh tube which could be stored at 0 °C. The supernatant was titred and found to be approximately 10^{10} pfu/ml.

Two 500ml batches of L-broth containing 5mM $MgSO_4$ were inoculated with 2 X 3.5ml of saturated overnight bacterial culture. The culture was grown until the A_{630} value reached 0.3 and then a total of 5 X 10^{10} pfu of phage λ1059 added. The infected culture was shaken until lysis occurred (usually about 3.5 hr) and then 5.0ml of chloroform was added to each flask and shaking continued a further 5 min.

The lysed cultures were decanted, sedimented by centrifugation at 4,000 rpm for 20 min, and pancreatic DNase and RNase (Sigma) were added to a final concentration of 10µg/ml. After 30 min incubation, solid NaCl was added to 2%, solid polyethylene glycol (Serva) was added to 8%, and the culture was left standing overnight at 4 °C to precipitate the phage particles.

The phage was sedimented by centrifugation at 6,000 rpm for 30 min and carefully resuspended in 20ml of lambda diluent. Solid CsCl (0.71g/ml) was added, giving a final density of 1.5g/ml. After a clarifying centrifugation in the 'Table-top' centrifuge for 30 min, the solution was transferred into sealable tubes and centrifugation was performed at 50,000 rpm and 25 °C overnight using a VTi50 rotor of a Beckman ultracentrifuge.

The white phage band was removed (see section 2.3.1) and a second CsCl centrifugation was performed at 65,000 rpm and 25 °C overnight. The phage

suspension was extracted and then dialysed against four changes of 500ml buffer (10mM Tris-HCl, pH 7.5; 1mM EDTA; 10mM $MgSO_4$) where the CsCl was removed. DNA was extracted with phenol/chloroform twice (see section 2.5.1), and precipitated with ethanol (see section 2.5.2). DNA was sedimented by centrifugation at 10,000 rpm for 10 min, briefly dried under vacuum, and resuspended in 400µl TE. Boiled pancreatic RNase A (Boehringer, grade I) was added to 10µg/ml and left standing at room temperature for 30 min. The phage DNA was stored at 0 °C.

2.4.2 Preparation of Bacteriophage from Lysogenic Strain

An overnight 30 °C culture of the *E.coli* strain W8850 (Table 2.1), lysogenic for $\lambda cI_{857}S_7$ was prepared. The overnight culture (4 X 10ml) was inoculated into 4 X 200ml of L-broth containing 10mM $MgSO_4$. The cultures were shaken at 30 °C until the A_{630} reached 0.7, and the lysogen was induced by incubation at 42 °C for 30 min. The induced culture was shaken a further 90 min at 37 °C. The cells were harvested by centrifugation at 6,000 rpm for 15 min and resuspended in 4ml of supernatant fluid. Chloroform (0.3ml) was added and then shaking was continued at room temperature until the solution became very viscous. DNase (Boehringer, grade II) was added to 5µg/ml and incubated at 37 °C for 5 min to reduce the viscosity.

The volume was adjusted to 20ml with lambda diluent and 14.2g of CsCl was added to give a density of 1.5g/ml. Subsequent procedures continued as described in the standard phage preparation (see section 2.4.1).

2.5 Extraction and Precipitation of DNA

DNA was routinely purified free of enzyme by extraction with phenol/chloroform or with phenol/ether, followed by precipitation with ethanol.

2.5.1 Phenol/Chloroform Extraction

Phenol was redistilled before use, saturated with TE and stored at -20 °C. Extraction with phenol/chloroform was performed as follows: an approximately equal volume of phenol was added to the DNA sample to be extracted, mixed; centrifuged for 1 min; the upper aqueous layer transferred to a fresh tube and

the extraction with phenol was repeated; 500µl chloroform was added and mixed; centrifuged for 10s; the aqueous layer transferred to a fresh tube and the extraction with chloroform repeated.

2.5.2 Phenol/Ether Extraction

Extraction with phenol/ether were performed as above with ether saturated with water replacing the chloroform. In this case, the ether was removed as the upper phase after extraction.

2.5.3 Ethanol Precipitation

The ethanol was of the absolute alcohol 100 grade (James Burrough) and stored at -20 °C. To the solution of DNA 0.1 volume of 3M sodium acetate (pH 5.2) and 2.5 volumes of cold ethanol was added and mixed; placed at -20 °C overnight or -70 °C for 15 min; centrifuged for 10 min at 0 °C; the precipitate washed with 70% ethanol and dried under vacuum for 5 min. The DNA was usually stored in TE at -20 °C.

2.6 Digestion with Specific Restriction Endonucleases

Restriction enzymes were obtained from the following:

Anglian Biotechnology Ltd. : Unit 8, Hawkins Rd, Colchester, Essex CO2 8JX

Boehringer : The Boehringer Corporation (London) Ltd.,Lewes, E. Sussex

BRL : Bethesda Research Labs., PO Box 35, Trident House, Renfrew Rd., Paisley PA3 4EF

NBL : NBL Enzymes Ltd., South Nelson International Estate, Cramlington, Northumberland, NE23 9HL

New England Biolabs : CP Labs. Ltd. (UK distr.), Bishop Stortford, Herts

Pharmacia Ltd : Pharmacia House, Midsummer Boulevard, Central Milton Keynes, Bucks MK9 3HP

2.6.1 Reaction Buffers

Restriction digests were generally set up using one of three convenient buffers and at the temperature specified by the manufacturer.

The composition of the restriction enzyme buffers are shown below:

Buffer	NaCl	Tris-HCl	$MgSO_4$	Dithiothreitol
low	0	10mM, pH7.4	10mM	1mM
medium	50mM	10mM, pH7.4	10mM	1mM
high	100mM	50mM, pH7.4	10mM	0

Buffers were usually stored as a 10 X concentrated stock at -20 °C.

2.6.2 Restriction Digestions

Restriction digestions were routinely carried out in 1.5ml Eppendorf tubes in the presence of the appropriate reaction buffer. One unit of enzyme activity is defined as the amount of enzyme required to digest 1μg of DNA to completion in 1 hr. However a several fold excess of enzyme per digest was usually added to ensure complete digestion.

A typical reaction mixture contained 1μg of DNA and 10 units of restriction enzyme in a final volume of 25μl of the appropriate restriction enzyme buffer. The mixture was incubated for 1 to 2 hr and the extent of digestion was monitored by electrophoresis of a small aliquot in a 1% agarose mini-gel (see section 2.7.2).

In multiple digestions where different buffers were required, the digestion which required the lowest ionic strength was carried out first, then the ionic conditions adjusted, the second enzyme added and the incubation continued.

2.6.3 Restriction Mapping

This was carried out mainly by logical interpretation of the results of a combination of single and double digestions. Sequential digests of fragments recovered from agarose gels (see section 2.7.3) were very useful.

2.7 Separation of DNA Fragments by Agarose Gel Electrophoresis

Electrophoresis through agarose gels was the standard procedure used to separate, identify, and purify DNA fragments.

2.7.1 Preparation of Agarose Gels

(a) Electrophoresis Buffers

Loening's phosphate buffer (Loening, 1967) contains : 36mM Tris-HCl; 30mM NaH_2PO_4; 1mM EDTA; 0.5µg/ml ethidium bromide; should be pH 7.7 without further adjustment.

Acetate buffer contains : 40mM Tris-HCl; 5mM sodium acetate; 1mM EDTA; 0.5µg/ml ethidium bromide; adjusted to pH 7.4 with acetic acid.

TBE buffer contains : 100mM Tris-HCl; 100mM boric acid; 1mM EDTA; 0.5µg/ml ethidium bromide; should be pH 8.3 without further adjustment.

These buffers were prepared as a 20 X (10 X for TBE) concentrated stock solution and stored at room temperature.

(b) Loading Buffers

These were 1:1 ratios of glycerol and 0.025% bromophenol blue in the appropriate electrophoresis buffer.

(c) Agarose Solutions

The required concentration of agarose (Sigma, type II) was made up in electrophoresis buffer. The mixture was dissolved by heating and could be stored at room temperature.

(d) Electrophoresis Conditions

Size Range	Gel Concentration	Recommended Voltage
10-100kb	0.5%	40V
0.8-15kb	0.7%	40V
0.4-8kb	1.0%	60V
0.1-3kb	2.0%	80V

1% agarose gels were routinely used for plasmid DNA, while 0.5% was best for restriction digests of lambda DNA, and 0.7% was used to handle genomic DNA.

(e) Horizontal Gel Electrophoresis Apparatus

The dimensions of the mini-agarose gel were 12 X 12cm, and this was

submerged in a 18 X 12cm mini-gel tank. This was either used as an analytical gel or a preparative gel, depending on the percentage of agarose and the capacity of the comb size.

The dimensions of the large agarose gel accompanying the large horizontal gel tank (Pharmacia GNA-200) were 22 X 22cm, and this was generally used in the analysis of genomic DNA.

2.7.2 Electrophoresis in Agarose Gels

The DNA sample was mixed with 3µl (or 0.3 volume) of loading buffer and applied to the sample well. Electrophoresis was carried out until the bromophenol blue had travelled the desired distance, dependent upon the size of the specific sample.

On completion of electrophoresis, the gel was removed and examined on a UV transilluminator (UV Products, Bishops Stortford, Herts) and photographed (if desired) using a Polaroid camera (Cu-5 Hand camera) with type 665 positive/negative film.

The sizes of the restriction fragments were determined by comparing with DNA marker fragments of known size subjected to electrophoresis alongside the unknown fragments.

The distances between the well and the positions where the DNA fragments of known sizes had travelled were measured and plotted on semi-log graph paper as distance travelled (mm) against log size of DNA (kb). Similarly, the distance travelled by DNA fragments of unknown sizes were then measured, and their sizes were determined from the standard curve.

2.7.3 Recovery of DNA from Agarose Gels

(i) Electroelution

This method is after McDonnell *et al* (1977) and involves electroelution of the DNA band in a dialysis bag.

The DNA band of interest was excised from the agarose gel (prepared and electrophoresed in acetate buffer) with a small scalpel, and placed into a piece of dialysis tubing (9mm in width). A minimal volume of acetate buffer was added (200µl) and the bag was sealed with clips at both ends. The bag was immersed in a shallow layer of acetate buffer on the platform of the gel tank,

and electrophoresis was performed at 60V for 1 hr. The solution was removed (and retained) and a further 200µl of acetate buffer was added and electrophoresis was continued for 15 min. The polarity of the current was reversed for 2 min, and then the solution was collected. The pooled DNA solution was centrifuged to remove agarose, then precipitated with ethanol. The DNA isolated was suitable for the nick-translation reaction.

(ii) Recovery of DNA from Low Melting Agarose

Low melting agarose (BRL) gels were prepared exactly as agarose gels (see section 2.7.1), except that they were poured and electrophoresis conducted in the cold room.

The DNA band of interest was excised and transferred to an Eppendorf tube. Two volumes of TE were added, and the sample was placed in a 65 °C heating block to melt the agarose. The sample was mixed, extracted twice with phenol/chloroform and precipitated with ethanol.

2.8 Separation of DNA Fragments by Polyacrylamide Gel Electrophoresis

Electrophoresis through polyacrylamide gels was another method used to separate, identify and purify DNA fragments. The method was routinely used to isolate ^{32}P-labelled DNA fragments for sequencing by the method of Maxam and Gilbert, and to resolve DNA fragments of similar size where agarose gel electrophoresis was inadequate.

2.8.1 Preparation of Acrylamide Gels

(a) Electrophoresis Buffer

This was TBE buffer (see section 2.7.1).

(b) Acrylamide Loading Buffer

This was a 1:1 ratio of glycerol and 0.05% Xylene Cyanol, 0.05% bromophenol blue in TBE buffer.

(c) Acrylamide Stock (20%)

This was a mixture of 19% v/v acrylamide (BRL, ultra pure grade) and 1%

N,N-methylene bisacrylamide (BRL) dissolved in distilled water, filtered, and stored in the dark at 4 °C.

(d) Electrophoresis Conditions

Size Range	Gel Concentrations
100-1000kb	4%
80-600kb	6%
60-400kb	8%
40-200kb	12%

Acrylamide gels (4%) were routinely used for the isolation of labelled fragments. A typical acrylamide gel mixture contained 4% acrylamide, 10% glycerol and 0.04% ammonium persulphate in a final volume 50 ml in TBE buffer.

(e) Vertical Gel Apparatus

The vertical gel apparatus (BRL model V161) was assembled with 1.5mm spacers. TEMED (NNN'N'-tetramethylethylenediamine; 40µl) was added to 5ml of the gel mixture, and poured down the gel plates to form the plug. TEMED (30µl) was then added to the remaining gel mixture and the main gel was poured.

2.8.2 Electrophoresis in Acrylamide Gels

The gel was subjected to pre-electrophoresis for 30 min at 150V. The DNA sample was mixed with the loading dye (0.3 volume) and applied to the gel. Electrophoresis was continued for 2 to 3 hr at 200V, until the required resolution was obtained. The gel was then stained in a solution of ethidium bromide for 15 min, and photographed on a UV illuminator (see section 2.7.2).

2.8.3 Recovery of DNA from Polyacrylamide Gels

Having stained and photographed a gel, the desired band was cut out and placed in a 1ml automatic pipette tip (Eppendorf), previously sealed at the narrow end and plugged with glass wool. The gel piece was ground with a glass rod, then 600µl of elution buffer (500mM ammonium acetate; 10mM

$Mg(acetate)_2$; 1mM EDTA; 0.1% SDS) was added. The top of the tip was sealed with Nescofilm, placed in a siliconised 15ml Corex tube, and incubated overnight at 37 °C.

The DNA of interest was eluted by cutting the end of the tip and allowing the elution buffer to drain into the Corex tube. The tip was further rinsed with 4 X 200µl portions of elution buffer. The pooled eluate was precipitated with ethanol, and the DNA sedimented by centrifugation at 10,000 rpm for 30 min at -10°C. The DNA was then suspended in 400µl of 0.3M sodium acetate, centrifuged to remove any acrylamide debris and the supernatant transferred to a fresh Eppendorf tube. The supernatant was then re-precipitated with ethanol, centrifuged to sediment the DNA, and briefly dried under vacuum.

2.9 Southern Blotting, Radiolabelling and Hybridisation of DNA

2.9.1 DNA Transfer to Nitrocellulose (Southern Blotting)

This method is based on that of Southern (1975). After electrophoresis the agarose gel was placed in 250ml of denaturing solution (1.5M NaCl; 0.5M NaOH) and left to soak for 30 min. The gel was then placed in 500ml of neutralising solution (1.5M NaCl; 0.5M Tris-HCl, pH 7.6) for a further 30 min.

The arrangement of the blotting components were as follows: 500ml of 20 X SSC (SSC = 0.15M NaCl; 15mM Na citrate, pH 7.3) was added to the buffer reservoirs; two strips of Whatman 3MM paper were connected over the solid support into both reservoirs to form the wicks; the gel was inverted and placed onto the bridge (ideally the same width); nitrocellulose filter (Schleicher and Schuell, BA85; cut to size) was carefully placed onto the gel after previously wetting in 2 X SSC; four sheets of 3MM paper were further placed on top; a stack of absorbent pads (cut to size and 6cm in thickness) was then placed on top; and finally the whole system was compressed using a glass plate and a 1.5kg weight. Transfer of DNA was allowed to proceed for up to 16 hr at room temperature.

After blotting, the nitrocellulose filter was removed, washed in 2 X SSC to remove any adhering agarose and allowed to dry at room temperature on 3MM paper. This was then baked in a vacuum oven for 2 hr at 80 °C.

2.9.2 Radiolabelling DNA Fragments

(i) Nick-translation of DNA

This method is according to Rigby *et al.* (1977). It was applied to insert DNA derived from agarose separation gels (see section 2.7.3).

The labelling reaction was assembled in the following order on ice: probe DNA (0.3 to 1.0μg); 50μCi α^{32}P-dNTP (1mCi/100μl, Amersham); 50μM of each of the remaining dNTPs; DNaseI (10^{-7}mg/ml); and 5 units of DNA polymerase I (Boehringer) in a total volume of 18μl in medium salt restriction enzyme buffer. The reaction was allowed to proceed for 4 hr at 15 °C, after which time 100μl of NE (50mM NaCl; 0.5mM EDTA, pH7.0) was added. The labelled DNA fragment was purified by applying the mixture to a column of Biogel P-60 (Biorad) in a blue 1ml Eppendorf tip (equilibrated with NE). After the solution had adsorbed the column was eluted with 9 X 100μl of NE, each fraction being collected separately.

The ^{32}P in each fraction was estimated from its Cherenkov radioactivity in the ^{3}H channel of a scintillation spectrometer. The values obtained are approximately 30% of those by scintillation counting in the ^{32}P channel. The first peak of radiation fractions were pooled.

The nick-translated DNA was stored at -20 °C. The specific activity of DNA labelled by this method was between 5 X 10^{7} and 10^{8} cpm per μg of DNA.

(ii) Oligo-nucleotide labelling of DNA

A modified method for labelling DNA fragments without purifying the DNA from agarose was developed by Feinberg and Vogelstein (1984) and used in the later phases of this work. This modified technique is even more efficient than nick-translation, eliminates the loss of DNA and time involved in recovering the DNA fragments from agarose gels and can be used with very small amounts of template DNA.

Plasmid DNA was cleaved with an appropriate restriction enzyme and the fragments were electrophoretically separated in a single lane of a 1% low melting point agarose gel in acetate electrophoresis buffer (see section 2.7.1). The desired band was excised cleanly and placed into an Eppendorf tube. Water (2 X volume) was added and the tube was placed in a boiling water bath for 10 min to dissolve the gel and denature the DNA. It was stored at -20 °C.

Preparatory to subsequent labellings, the gel was reboiled for 7 min and maintained in a 37 ^{o}C block for at least 30 min until initiating the labelling reaction.

The labelling reaction was carried out at room temperature by addition of the following reagents in the order stated: 20% of oligo-nucleotide labelling buffer (see below); 20µg of bovine serum albumin ; 10 to 30ng of insert DNA; 20µCi of α^{32}P-dNTP (Amersham 1mCi/100µl); 5 units of large fragment of *E.coli* DNA polymerase I (BRL) in a final volume of 50µl. The reaction was left standing overnight at room temperature, after which time 50µl of a solution of: 20mM NaCl; 20mM Tris-HCl, pH7.5; 2mM EDTA; 0.25% SDS; and 1µM dNTP was added to stop the reaction. Purification of labelled DNA was by chromatography through a small column of Biogel P-60 as described previously (see section 2.9.2i).

Oligo-nucleotide labelling buffer was made up from the following components: 0.018% 2-mercaptoethanol and 50mM each of the remaining cold dNTPs in a final volume of 1ml in a solution of 1.25M Tris-HCl (pH 8.0) and 0.125M $MgCl_2$; 2M Hepes (pH 6.6); and a solution of hexadeoxyribonucleotide (Pharmacia) in TE at 90 OD units/ml, mixed in a ratio of 2:5:3 and stored at -20 ^{o}C.

The specific activity of the DNA labelled by this method was between 10^8 and 5 X 10^9 cpm per µg of DNA.

2.9.3 Hybridisation of Blotted DNA

The following hybridisation and working conditions were used for the detection of blotted nucleic acid sequences with DNA probes. The relatively low stringency gave optimal hybridisation signals, but the non-specific background was nevertheless low.

The blotted nitrocellulose filter was pre-wetted in 5 X SSPE (SSPE= 0.18M NaCl; 10mM $Na_2H_2PO_4$; 10mM NaOH; 1mM EDTA; adjusted to pH 7.2) and placed into a polythene bag. The bag was then sealed and prehybridised for 2 hr at 42^{o}C with 10ml of hybridising solution containing : 5 X SSPE; 10 X Denhardt's solution (Denhardt's = 0.02% Ficoll; 0.02% Polyvinyl-pyrolidine; 0.02% BSA; filtered and stored at -20 ^{o}C); 0.1% SDS; and 50% deionised formamide.

The appropriate DNA probe was denatured by adding 0.1 volume of 1M NaOH for 10 min, then neutralised by adding 0.1 volume of 1M Tris-HCl (pH 7.6) and 0.1 volume of 1M HCl.

After pre-hybridisation the filter was then hybridised overnight at 42 ^{o}C

with the hybridising solution (approximately 1ml/cm^2 of filter) and denatured probe. The hybridised filter was washed in 2 X SSC, 0.1% SDS for 5 X 10 min at room temperature, followed by 0.1 X SSC, 0.1%SDS at 42^{o}C for 2 X 30 min. After washing the filter, the filter was air dried, wrapped with cling film, and exposed to Kodak-X-Omat H film using an intensifying screen (Cronex-lighting) and left overnight at -70 ^{o}C.

The DNA fragments that were complementary to the sequence of the DNA probe were identified by the autoradiograph.

2.10 Subcloning into pUC Plasmid Vectors

DNA fragments of interest and of suitable size (usually less than 4kb) from mouse genomic lambda clones were derived and used to construct subclones. The plasmid vector pUC18 (Yanisch-Perron, Vierra and Messing, 1985), Figure 2.1, was used in the construction of the subclones in this project, the DNA to be cloned being inserted into one of the unique restriction sites in the polylinker of this vector.

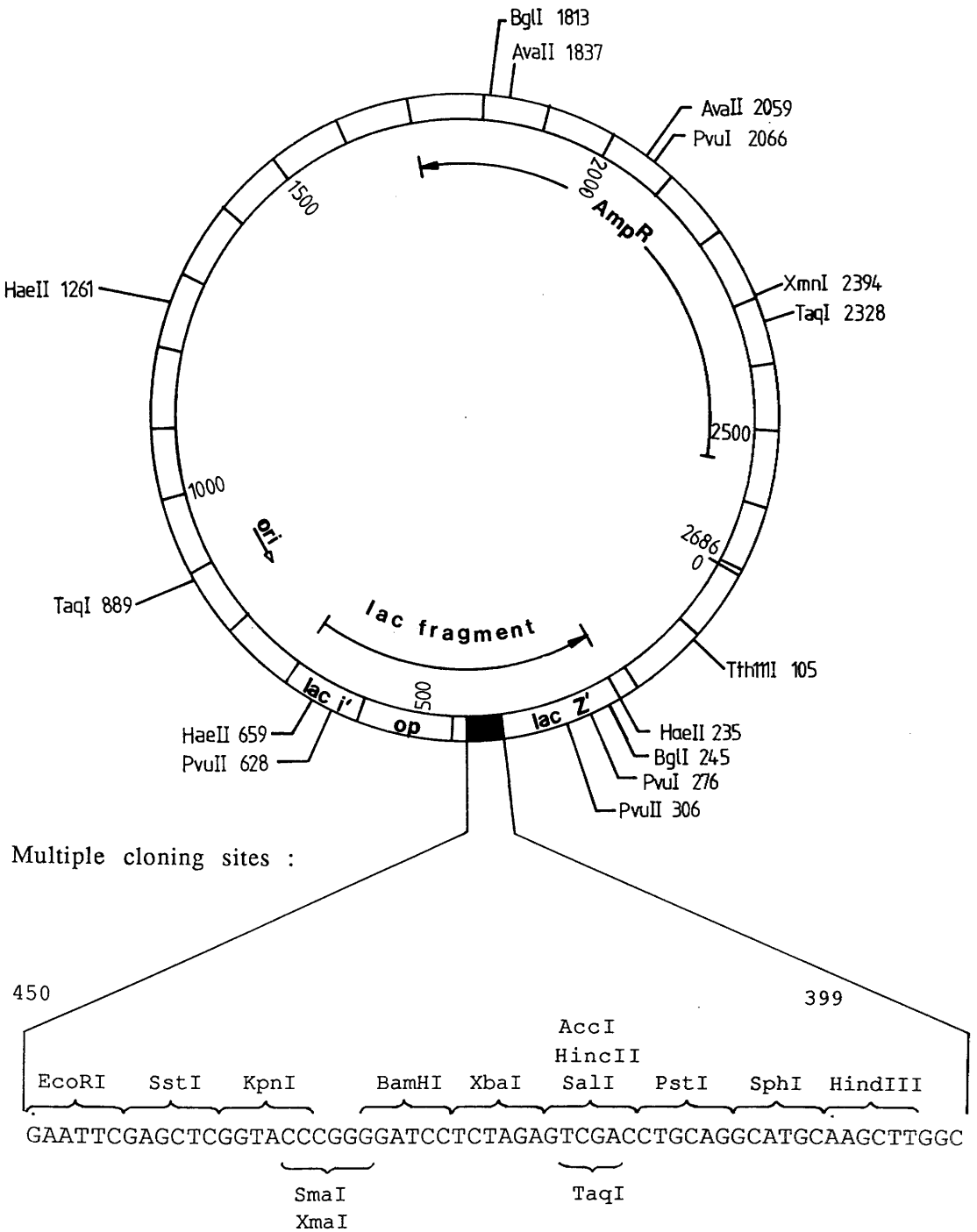
2.10.1 Preparation of Insert and Plasmid Vector DNA

Plasmid vector pUC18 DNA (5 to 10μg) and the cloned lambda DNA (2μg) were separately digested with the appropriate enzyme(s), and agarose gel electrophoresis was performed to check the completeness of the digest. Both vector and insert DNA were purified by extraction with phenol/ether, and then precipitated with ethanol. After lyophilisation, the insert DNA was redissolved in TE at 0.2μg/μl, while the vector DNA was kept as a lyophilised precipitate.

2.10.2 Alkaline Phosphatase Treatment of Vector DNA

Any undigested circular molecules of vector DNA would subsequently transform with a high efficiency to give a high background of blue colonies. This was reduced by complete digestion of vector DNA or purification of the linear form by agarose gel electrophoresis. If the vector DNA had been linearised by digestion with a single enzyme, then subsequent ligation and transformation would result in a high background of blue colonies due to religation of the vector. This could be reduced by removing the 5' phosphates with alkaline phosphatase.

Figure 2.1 Plasmid vector pUC18



The plasmid vector pUC18 (Yanisch-Perron *etal.*, 1985), used in the construction of the subclones in this project. This is a double-stranded circular DNA molecule, 2686 base pairs in length. It carries a 54 base pair multiple cloning site (polylinker) that contains sites for 13 different hexanucleotide-specific restriction enzymes. The overall map shows the restriction sites of those enzymes that were used in this project. The polylinker is shown below the map. The map also shows the positions of the ampicillin resistance gene and the lac gene fragment.

The restricted vector DNA precipitate was redissolved in 20µl of 50mM Tris-HCl (pH 9.5), 1mM spermidine, 0.1mM EDTA and 1 unit of alkaline phosphatase (Calf intestinal, Boehringer grade I) was added. After incubation at 37 °C for 30 min, the incubation volume was increased to 100µl with TE and then extracted with phenol/ether and precipitated with ethanol. The DNA was redissolved in TE at 0.3µg/µl.

2.10.3 Ligation of DNA Fragments

For high efficiency cloning, it was essential to have the correct molar ratios of clonable ends of vector and insert. Suitable molar ratios of total clonable ends of vector and insert were 1:1 and 3:1. Linearised pUC vector is approximately 2.5kb, so there would be 2 X clonable ends for 2.5µg. Lambda is approximately 50kb, thus if an enzyme cuts at n sites, generating n+1 fragments, n-1 of these would have clonable ends. Thus there are 2 X(n-1) clonable ends per 50µg lambda DNA, or 2 X(n-1)0.05 clonable ends per 2.5µg lambda DNA, and 2 clonable ends per 2.5µg pUC DNA. Hence 1:1 weight ratio = 1: 0.05(n-1) vector : insert molar end ratio, and 3:1 weight ratio = 3 : 0.05(n-1) vector : insert molar end ratio.

The ligation reaction was assembled in the following order on ice: insert DNA(a suitable amount); cut phosphatased vector DNA (0.3µg); 0.5mM ATP ;1 unit of T4 DNA ligase (BRL) in a final volume of 30µl ligase buffer which contains 40mM Tris-HCl (pH 7.6), 1mM $MgCl_2$, and 1mM dithiothreitol.

The ligation mixture was incubated overnight at 15 °C. Control ligations of digested vector, and vector treated with alkaline phosphatase were carried out.

2.10.4 Transformation of *E.coli* by Plasmid DNA

(i) Preparation of Cells Competent for Transformation

The bacterial strains used to make 'competent' cells were *E.coli* JM103 and JM109.

An overnight culture of the bacterial host cells (2.5ml) was inoculated into 500ml L-broth. It was shaken at 37 °C and allowed to grow until an A_{600} of 0.2 was reached. The cells were harvested by centrifugation at 5,000 rpm for 15 min at 4 °C, the supernatant removed and the cells resuspended in 0.5 volume

(250ml) of cold 100mM $CaCl_2$. After allowing to stand on ice for 20 min, the cell suspension was re-centrifuged at 6,000 rpm for 10 min at 4 °C, and the cells were gently resuspended in 0.01 volume (5ml) of cold 100mM $CaCl_2$. Sterile glycerol was added to a final concentration of 10% (v/v). The cells were aliquoted in 1ml portions, frozen in liquid nitrogen and stored at -70 °C.

(ii) Transformation of *E.coli* by Plasmid DNA

An aliquot of frozen competent JM109 cells were allowed to thaw slowly on ice for 30 min. Then 0.1 and 0.5 volumes of both 1:1 and 3:1 ligation mixes of vector and insert, plus the two controls, were used to transform 100µl aliquots of competent cells. The transformation reaction mixture was allowed to stand on ice for 30 min, maintained for 2 min at 37 °C, and plated directly onto ampicillin/X-gal/IPTG plates. The plates were prepared 5 min before use, 0.5% IPTG (isopropyl-thiogalactoside) in sterile water, 0.5% X-gal (5-bromo-4-chloro-3-indolyl-ß-D-galactoside) in dimethyl formamide was spreaded over the surface of the ampicillin L-plates.

The plates were left at room temperature until all the liquid had been adsorbed, then they were inverted and incubated overnight at 37 °C.

Self-religating and undigested vector give blue colonies, while recombinants containing the insert DNA normally give white colonies in 8 to 10 hr.

2.10.5 Selection of Recombinant Clones

The pUC plasmids have been constructed as cloning vectors using ß-galactosidase activity as the basis of selection. The vector has a fragment of the *E.coli* lac operon containing the regulatory region and the coding information for the first 146 amino-acids of the ß-galactosidase (Z) gene. The amino-terminal peptide is able to complement the product of a defective ß-galactosidase gene present on the F' episome in the host cell. A 'polylinker' DNA fragment containing several unique restriction sites for cloning have been inserted, in phase, into the amino-terminal portion of the ß-galactosidase gene. This insertion does not affect the complementation. However, insertion of additional DNA into the 'polylinker' region generally destroys the complementation.

The complementation produces active ß-galactosidase which gives rise to

a blue colour when the transformed cells are grown in the presence of the inducer IPTG and the chromogenic substrate X-gal. However, when DNA is cloned into the 'polylinker' region, the ß-galactosidase is inactive and the colonies appear white. False positive white colonies occur at low frequency, probably due to incorrect self-ligation of the vector.

2.10.6 Identification of Recombinant Subclones

Bacteria containing the recombinant plasmid were identified by picking white colonies, and isolating their DNA (see section 2.3.2 i). The recombinant DNA was screened by limited restriction analysis and hybridisation to the blotted DNA with a ^{32}P-labelled probe (see section 2.9). An example is described below.

An XbaI fragment of 2.5 kb from the genomic clone λmA119, containing the actin-like DNA, was subcloned into the XbaI site of plasmid vector pUC18 (Figure 2.1).

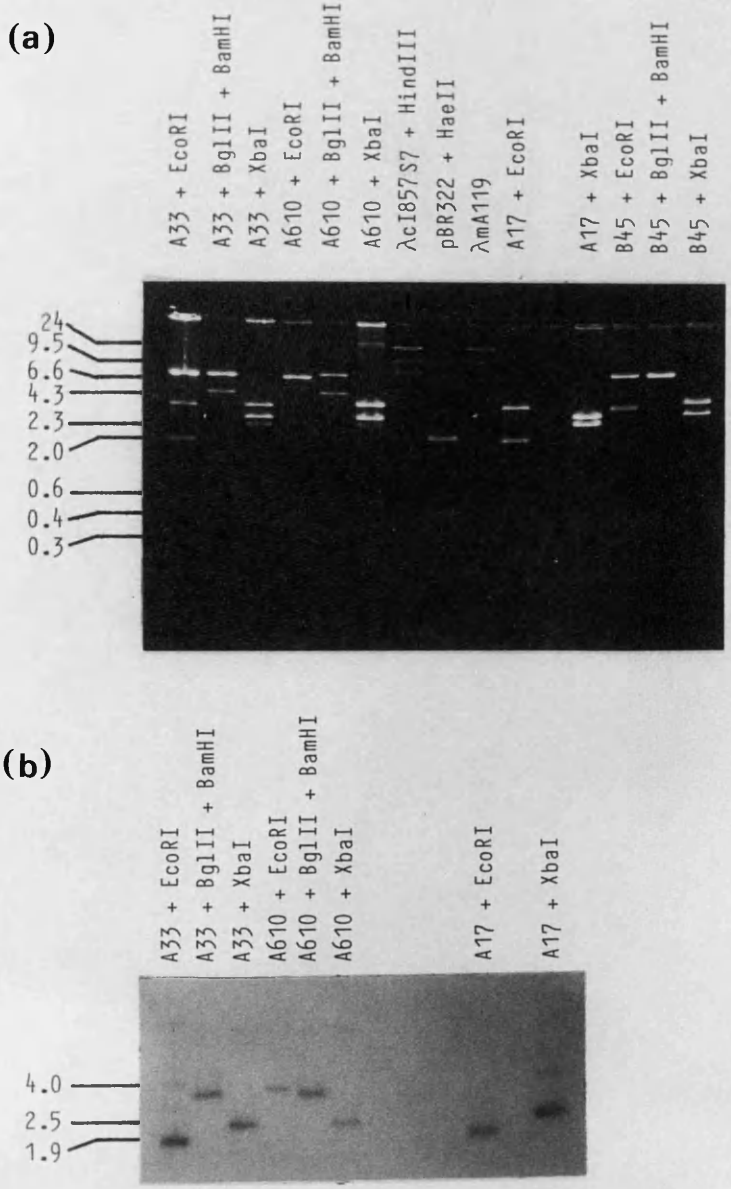
The DNA of λmA119 (2µg), and pUC18 (5µg) were digested with XbaI (see section 2.10.1). The restricted vector DNA was then treated with alkaline phosphatase (see section 2.10.2). The XbaI fragments from λmA119 were ligated into the restricted pUC18 (see section 2.10.3), and then used to transform 'competent' JM109 cells (see section 2.10.4),

The efficiency of transformation was approximately 3 X 10^4 transformants per µg of genomic DNA. A total of 144 white colonies were initially picked and plated onto master ampicillin plates. Small scale preparation of DNA were made for 12 of these colonies (see section 2.3.2).

The DNA was subjected to electrophoresis through a 1% agarose gel to check the quality of the preparation, and to select suitable subclones which might contain the recombinant plasmid. A number of the chosen subclone DNAs (1µg) were digested with suitable restriction enzymes, and then subjected to electrophoresis as shown in Figure 2.2a. The restricted DNA was then transferred to nitrocellulose filters by the Southern blotting method (see section 2.9.1).

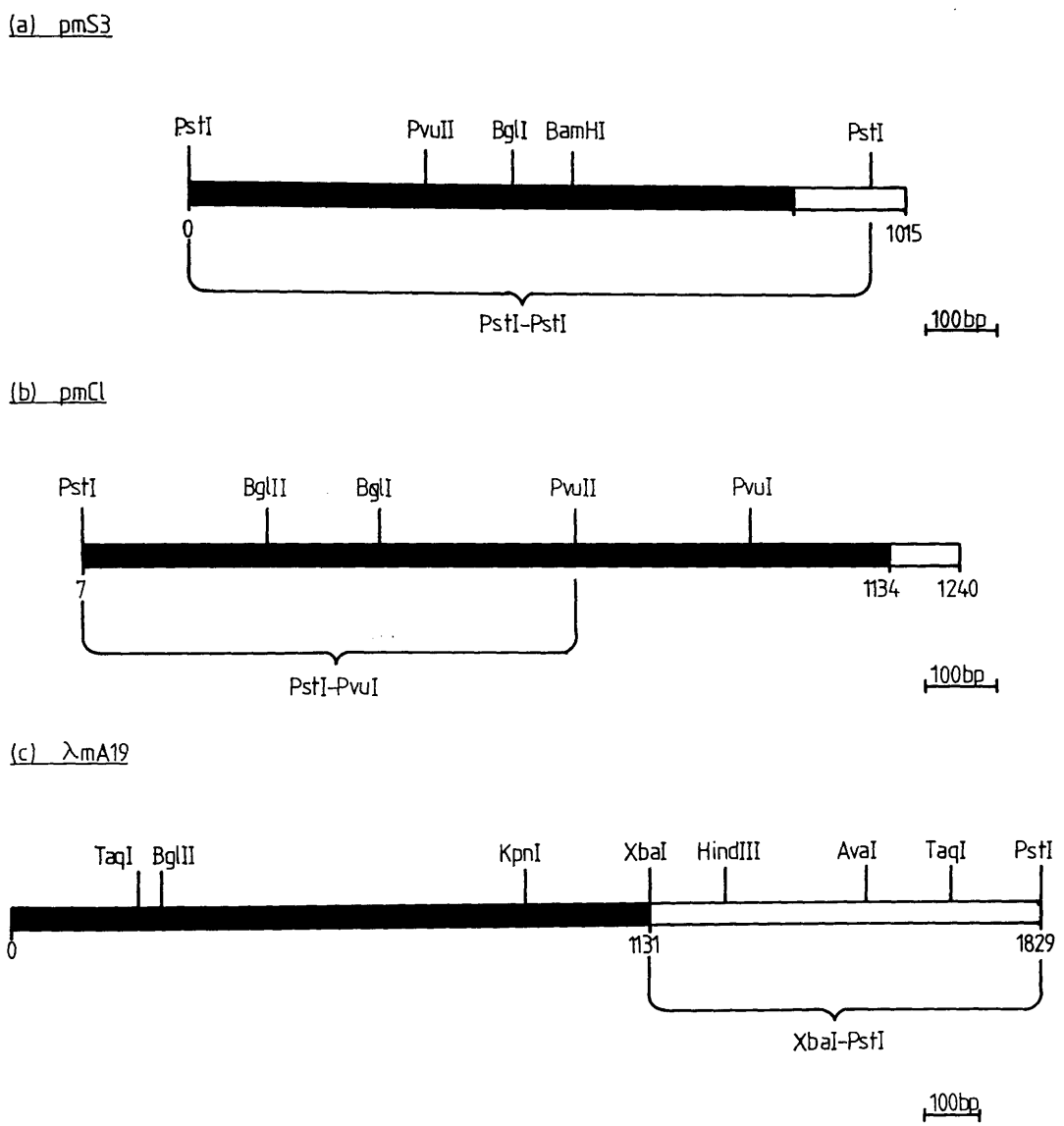
Subclones designated A33 and A610 can be seen to contain a similar 2.5kb XbaI fragment (Figure 2.2a) which could correspond to that in λmA119. This was confirmed by the fact that they hybridised to a ^{32}P-labelled PstI-PvuII fragment from clone pmC1 (see sections 2.9.2 and 2.9.3; Figure 2.3b), corresponding to cDNA specific for amino-acids 1 to 231 of cardiac muscle actin

Figure 2.2 Restriction digestions of subclones derived from clone λmA119 and hybridisation to radioactive-labelled actin probes



A number of subclones derived from subcloning the 2.5 kb XbaI fragment from the genomic clone λmA119 (Figure 3.6; section 3.2.2), were subjected to restriction digestion, and hybridised to a ^{32}P-labelled PstI-PvuII fragment of actin cDNA clone pmC1 (Figure 2.3). (a) shows the eletrophoretic pattern of the subclones given the preliminary designations A33, A610, B45, and A17 digested with EcoRI, BglII and BamHI, and XbaI. (b) shows the autoradiograph of the hybridisation.

Figure 2.3 **Restriction maps of actin clones used as probes in this work**



The partial restriction maps of the actin clones used in this project are listed. (a), pmS3, a mouse skeletal muscle actin cDNA clone (Leader *etal.*, 1986a). (b), pmC1, a mouse cardiac muscle actin cDNA clone (Leader *etal.*, 1986b). (c), λmA19, a mouse γ-actin pseudogene (Leader *etal.*, 1985). The vectors are excluded from the diagrams. The coding region of actin is indicated by solid blocks, and the 3' untranslated region is indicated by open blocks. The fragment of insert used as a probe is indicated by parenthesis.

(Figure 2.2b).

Further restriction analysis of the subclones A33 and A610 was carried out to characterise them in more detail, and to determine the orientation of the cloned fragment relative to the vector in the two subclones. The subclone A33 produced a 1.9kb fragment when digested with EcoRI (for which there is one site in the polylinker of the vector), while the subclone A610 produced a 0.6kb fragment (Figure 2.2a). This is consistent with the two subclones representing both orientations of a 2.5kb XbaI insert possessing a single asymmetrically positioned EcoRI site.

Purified DNA from the subclones A33 and A610 was prepared (see section 2.3.1) in order to allow for further analysis including sequence determination.

2.11 Preparation of Fragments for Sequencing by the Method of Maxam-Gilbert

2.11.1 Polynucleotide Kinase End-labelling of DNA

DNA for end-labelling must be free of low molecular weight RNA. This was removed where necessary using a Biogel A-15 (Biorad) column.

(i) Phosphatase Treatment

The terminal 5' phosphate of DNA was removed by treatment with alkaline phosphatase (Calf intestinal, Boehringer grade I).

The DNA sample (5 to 10μg) with 1 to 50 pmol of 5' protruding ends was made up to 100μl with TE. Alkaline phosphatase (1 unit) was added, mixed well, and incubated for 75 min at 37 ^{o}C. After incubation the mixture was extracted with phenol, and the phenol layer was re-extracted with 100μl TE. The aqueous phases were pooled and precipitated with 100μl of 0.3M Na acetate and 300μl of ethanol, washed with 70% ethanol, and briefly dried under vacuum.

(ii) Polynucleotide Kinase Labelling

This method is specific for labelling 5' protruding ends. The reaction was assembled by adding the following to the dried DNA (1 to 50 pmol ends): 5μM dithiothreitol; 60μCi γ^{32}P-ATP (Amersham, 1mCi/100μl); and 5 units of polynucleotide kinase (PL Biochemicals) in a final volume of 11μl kinase

buffer (50mM Tris-HCl, pH8.0; 10mM $MgCl_2$). The mixture was incubated at 37°C for 30 min. After incubation, 40µl of 2.5M ammonium acetate was added and precipitated with 160µl of ethanol. The DNA was then re-precipitated with 100µl of 0.3M sodium acetate and 300µl of ethanol, washed with 70% ethanol, and briefly dried under vacuum.

2.11.2 Klenow End-labelling of DNA

The removal of tRNA from the DNA preparation, or the elimination of the 5' phosphate from DNA were not required for this method, which involves the 3' to 5' 'filling-in' reaction of the Klenow large fragment of the restricted DNA.

This method could only be used for labelling 5' protruding ends. The radioactive α^{32}P-dNTP with high specific activity (Amersham, 1mCi/100µl) must be complementary to one of the nucleotides in the restricted 'sticky end'. It was routine to add all three of the remaining cold nucleotides to the reaction. However, it was sometimes possible to label a single end of a fragment with different 5' sticky ends specifically by judicious choice of hot and cold nucleotides to fill in.

The reaction was assembled by the addition of the following to the dried DNA (5 to 10µg) : 50µCi α^{32}P-dATP (Amersham, 1mCi/100µl); 4µM of each of the remaining three cold dNTPs; and 2 units Klenow fragment (Boehringer) in a final volume of 25µl in 1.5 X medium restriction enzyme buffer (see section 2.6.1). The mixture was incubated for 30 min at room temperature. After incubation, 90µl of 2.5M ammonium acetate was added and precipitated with 360µl of ethanol. The DNA was then re-precipitated with 100µl of 0.3M sodium acetate and 300µl of ethanol, washed with 70% ethanol, and briefly dried under vacuum.

2.11.3 Secondary Digestion and Separation of Labelled Ends

Chemical sequencing could only be performed on a fragment of DNA labelled at one end, therefore DNA fragments labelled at both ends were cleaved and the fragments separated.

A restriction enzyme was chosen that would cleave the DNA fragment asymmetrically, and preferably into two. The restricted fragments were separated by polyacrylamide gel electrophoresis, usually employing 4%

acrylamide (see sections 2.8.1 and 2.8.2). The gel was stained and the DNA fragments were eluted as described previously (see section 2.8.3). If the amount of DNA was too small to be seen by staining with ethidium bromide, then the gel was subjected to autoradiography for 15 min at room temperature.

The radioactivity of the dried DNA was estimated by measuring its Cherenkov radiation. The minimum activity required to proceed to the next stage was 2 X 10^4 cpm.

2.12 Sequencing DNA by the Chemical Method of Maxam and Gilbert

This method of sequencing involves base modification and strand scission by chemical means. A detailed description is given by Maxam and Gilbert (1977 and 1980).

2.12.1 Reagents and Solutions

Dimethlysulphate - DMS (Aldrich Chemical Co., Dorset)

Hydrazine - HZ (Kodak Ltd.) : stored at -70 °C.

Piperidine - (Koch Light Labs., Bucks.) : stored at 4 °C.

Pyridine formate : 4% v/v formic acid was adjusted to pH2.0 with pyridine (BDH,AnalaR).

DMS Buffer : 50mM Na cacodylate; 10mM $MgCl_2$; 0.1mM EDTA; adjusted to pH8.0 and stored at 4 °C.

DMS Stop : 1.5M Na acetate, pH7.0; 1M ß-mercaptoethanol (Koch Light); 100µg/ml yeast tRNA; stored at -20 °C.

HZ Stop : 0.3M Na acetate; 0.1mM EDTA; 50µg/ml yeast tRNA; stored at 4 °C.

2.12.2 Modification Reactions and Strand Scission

The four reactions used for full sequence determination were specific for guanine (G), guanine and adenine (G+A), cytosine and thymine (C+T), and cytosine (C). Chain cleavage was achieved using 1M piperidine. The precise procedure followed for each of the four reactions was as follows.

The dried labelled DNA (1µg) was dissolved in 11 µl of water and 4µg of calf thymus carrier DNA was added. The mixture was aliquoted equally into four siliconised Eppendorf tubes labelled G, A(+G), T(+C) and C. Each tube then

received different components : 98µl DMS buffer into tube G; 11µl water into tube A(+G); 6µl water into tube T(+C); 8µl water saturated with NaCl into tube C.

Pyridine formate (2.5µl) was added to tube A(+G), and the mixture was incubated for 70 min at 30 °C. The reaction was stopped by freezing the mixture at -70 °C for 5 min followed by drying under vacuum. The sample was washed with 10µl water, frozen and dried as before.

Dimethylsulphate (0.5µl) was added to tube G, and the mixture was incubated for 5 min at 20 °C. The reaction was stopped by the addition of 24µl DMS Stop, 400µl cold ethanol and then left at -70 °C for 15 min.

Hydrazine (15µl) was added to tubes T(+C) and C, mixed and incubated at 20°C for 8 and 10 min respectively. The reactions were stopped by the addition of 60µl HZ Stop, 250µl cold ethanol and then left at -70 °C for 15 min.

After precipitation, tubes G, T(+C) and C were centrifuged for 5 min and the supernatant was discarded. The DNA was re-precipitated in 60µl 0.3M sodium acetate and 200µl cold ethanol, and then briefly dried under vacuum.

To all four tubes G, A(+G), T(+C) and C, 100µl of 1M piperidine was added, and the mixtures were heated at 90 °C for 30 min. After a brief centrifugation (5s), the samples were frozen at -70 °C and dried under vacuum for 2 to 3 hr. The residual piperidine was removed by washing twice with 20µl water followed by drying under vacuum (2 hr). The Cherenkov radiation of each tube was measured.

2.12.3 DNA Sequencing Gels

DNA fragments differing in length by only one nucleotide were separated by electrophoresis on 6% polyacrylamide urea denaturing gels (40cm X 20cm X 0.4mm) according to Sanger and Coulson (1978).

A typical gel mixture contained 6% acrylamide, 7M urea and 0.01% ammonium persulphate in a final volume of 100ml in TBE buffer.

The gel plates were siliconised before use with Repelcote (BDH) and assembled with 0.4 mm spacers. TEMED (NNN'N'-tetramethylethylenediamine; 40µl), was added to the gel mixture before pouring, a comb (14 X 7mm) was inserted , and the gel was allowed to set at room temperature.

The gel was subjected to pre-electrophoresis at 25 to 30mA for 1 to 2 hr (LKB 2103 power supply, LKB Instruments Ltd., Surrey). During this time, the samples were dissolved in the appropriate volume of sequencing loading dye (99% deionised formamide; 0.05% xylene cyanol) to give 10,000 cpm

(Cherenkov) per µl.

When the gels were ready, the samples were boiled for 2 min, chilled on ice, and 1.5µl of each was loaded on the first set of wells. The length of the run was chosen so as to be appropriate for the total length of the DNA fragment. For DNA fragments between 200 base pairs and 1.2 kb in length, three consecutive loadings were usually made. The first dye front was allowed to travel down the gel for approximately 20cm before the second loading was made, and this allowed to travel 15cm before the third loading was made. Electrophoresis was stopped when the dye front from the third loading had travelled about 13cm. The samples had to be reboiled before each loading, and 1µl of each was used for the second and third loading. Xylene cyanol migrates with a mobility equivalent to a DNA fragment of 60 base pairs on a 6% acrylamide gel.

2.12.4 Autoradiography

After electrophoresis, one of the glass plate was removed and the exposed gel was covered with clingfilm. Autoradiography was performed at -70 °C using Kodak X-Omat H film and an intensifying screen (Cronex Lighting Plus, Dupont, Huntingdon). An overnight exposure was required for 10,000 cpm (Cherenkov) per loading, the exposure time was increased accordingly if less than 10,000 cpm were used.

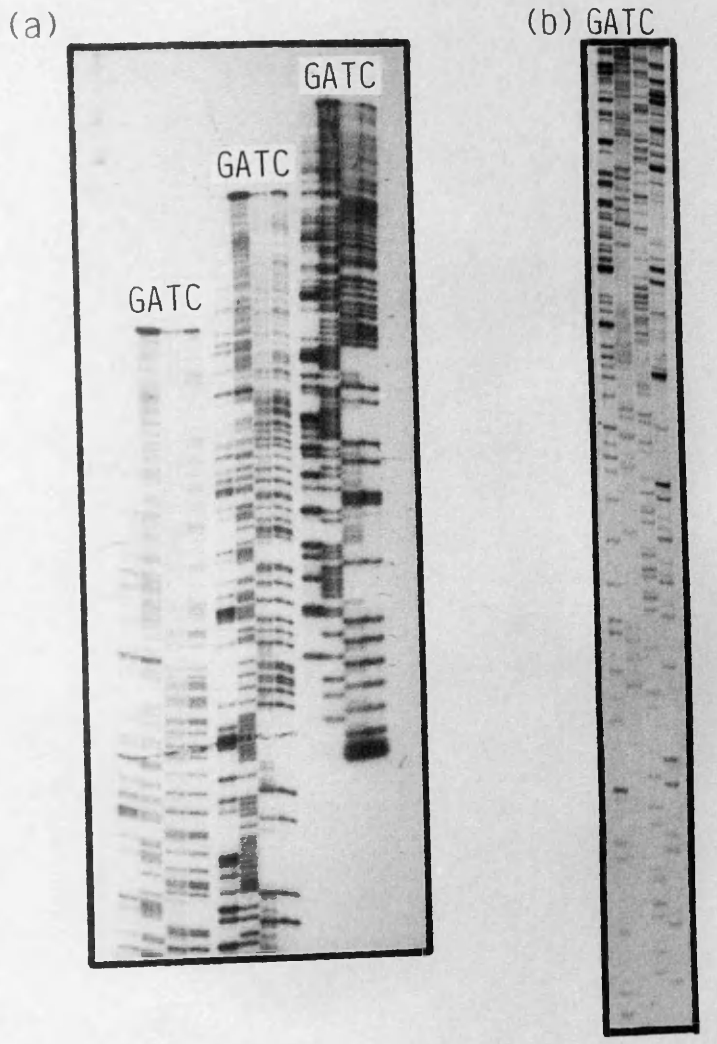
Following autoradiography, it was possible to read between 150 to 200 nucleotides from one gel. Figure 2.4a shows an example of an autoradiograph of a sequencing gel.

2.13 Cloning into M13 and Preparation of Single-stranded Template

The aim of cloning into bacteriophage M13 (mp18 and mp19; Figure 2.5) was to take fragments of double-stranded DNA, and using M13 RF (replicative form) DNA as a vector, produce from the resulting virus, pure single-stranded DNA template suitable for the Sanger "Dideoxy Sequencing" method (Sanger, 1981; Messing, 1983). All the protocols for cloning were supplied in the form of "M13 Cloning and Sequencing Handbook" (Amersham International) and these were strictly adhered to.

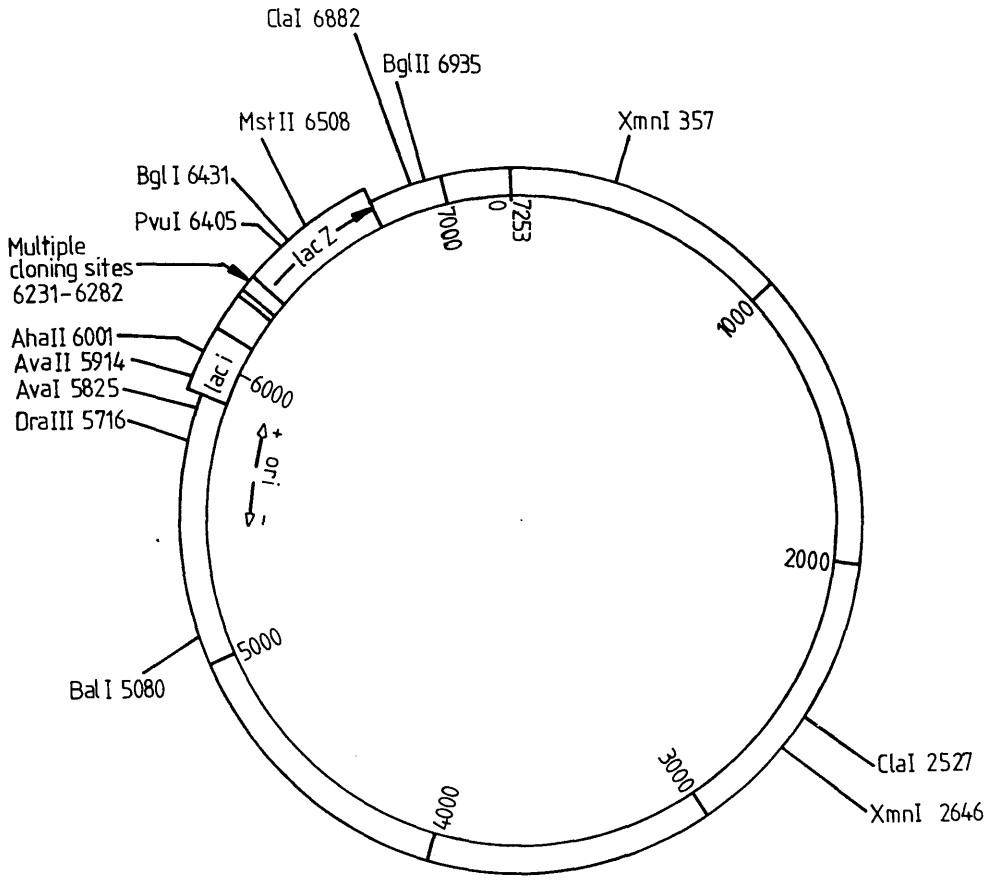
The paired vectors M13 mp18 and M13 mp19 (Yanisch-Perron *et al.*, 1985;

Figure 2.4 Example of polyacrylamide gel separation of radioactively labelled nested fragments of DNA generated for nucleotide sequence determination by the methods of Maxam and Gilbert and of Sanger



(a) The subclone 119XB from clone λmA119 (Figure 3.6) was restricted with StyI, 5' Kle-now end labelled, and secondary cleaved with BglII. Maxam and Gilbert sequencing was performed from the StyI site, allowing determination of sequence number 4 in Figure 3.9. The resulting autoradiograph is shown. Nucleotides are numbered according to the complete sequence in Figure 3.10. (b) An EcoRI fragment from the subclone 36KK in clone λmA36 (Figure 4.8) was cloned into bacteriophage vector M13 mp18, and single-stranded templates prepared. Sanger "Dideoxy" sequencing was performed, thus allowing determination of sequence number 10 in Figure 4.9. The resulting autoradiograph is shown. Nucleotides are numbered according to the complete sequence in Figure 4.10.

Figure 2.5 Bacteriophage vectors M13 mp18 and M13 mp19



Multiple cloning sites in M13 mp18:

```
6230                                                              6289
                                        AccI
                                       HincII
  EcoRI    SstI    KpnI      BamHI  XbaI   SalI    PstI    SphI  HindIII
GAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTGGC
          SmaI
          XmaI                          TaqI
```

The bacteriophage vectors M13 mp18 and M13 mp19 (Yanisch-Perron *et al.*, 1985), used for the nucleotide sequence determination by Sanger "Dideoxy-Sequencing" method. This are single-stranded circular molecules, 7253 bases in length, and differ only in the orientation of the 54 base polylinker that they carry. The polylinker includes 10 discrete hexanucleotide recognition sites for 13 different enzymes. The map shows a number of restriction sites of enzymes that cleave the molecule once or twice. The multiple cloning sites in the orientation present in M13 mp18 are shown below the map. In M13 mp19 the same sites are present in the opposite orientation. The map also shows the positions of the lac gene fragment and the origins of plus and minus strand replication.

Figure 2.5) were used in this project, and their double-stranded replicative form was a gift from Dr Ken Duncan.

2.13.1 Preparation of Insert and Vector DNA

Insert DNA (1μg) and M13 vector DNA (2μg) were digested with the appropriate enzyme(s) and the extent of digestion checked by gel electrophoresis. After digestion, the insert and vector DNA were purified by extraction with phenol/chloroform and precipitated with ethanol. The restricted insert DNA was resuspended in TE to 20ng/μl.

If the vector had been linearised by digestion with a single enzyme, 5' phosphates would have to be removed by alkaline phosphatase to reduce the high background of the plaques due to religation of the vector.

The restricted vector DNA was redissolved in a final volume of 40μl in a solution of 10mM Tris-HCl (pH9.2), 0.1mM EDTA and 1 unit of alkaline phosphatase (calf intestinal, Boehringer grade I) was added. After incubation at 45 °C for 30 min, 1 unit of alkaline phosphatase was further added and re-incubated at 45 °C for 30 min. The vector DNA was then purified by extraction with phenol/ether, precipitated with ethanol, and redissolved in TE to 10ng/μl.

2.13.2 Ligation of RF DNA to Insert DNA

Ligation mixes for ligation of RF DNA to insert DNA were assembled on ice as follows : Insert DNA (100ng); vector DNA (20ng);1mM ATP; 5mM dithiotheitol; and 1 unit of T4 DNA ligase (BRL) in a final volume of 10μl in ligase reaction buffer (50mM Tris-HCl,pH7.4; 10mM $MgCl_2$). The ligation mixture was incubated at 14 °C for 4 to 12hr.

2.13.3 Transformation of *E.coli* and Plating Out

Competent cells were prepared from *E.coli* JM109 strain as in section 2.10.4. Alternatively, frozen competent cells were used routinely. 10ml of 2 X TY was also inoculated with a drop of overnight culture to provide exponentially growing cells for plating.

To 300μl of competent cells, 5μl of DNA ligation mix was added, mixed, and left on ice for 40 min. The cells were then maintained at 42 °C for 3 min and returned to ice. During this time the following were added to 3ml of molten H-

top agar (kept at 42 °C) in sterile culture tubes : 0.03% IPTG in water; 0.03% X-gal in dimethyl formamide; 0.06% fresh JM109 cells. After mixing, the molten agar mix was added to the transformation mix, and directly spreaded onto a pre-warmed H-plate (see section 2.1.2). The plate was allowed to set, and incubated at 37 °C overnight.

Transformed cells showed up as a colourless plaque on a lawn of uninfected cells, while self-religating vectors showed up as blue plaques.

2.13.4 Preparation of Single-Stranded Template

A single white plaque was inoculated using a sterile Eppendorf tip into 1.5 ml 2 X TY containing 0.01 volume of an overnight *E.coli* JM109 culture.

This culture was shaken for 5 hr at 37 °C, and the cells sedimented by centrifugation, while the supernatant was transferred to a fresh tube and re-centrifuged. The second supernatant (1.0ml) was added to 200µl of a solution of 20% polyethylene glycol 6,000 and 2.5M NaCl, mixed, and left standing at room temperature for 15 min. The viral DNA was then sedimented by centrifugation, and the remaining supernatant was removed by a drawn out pasteur pipette in order to remove all traces of polyethylene glycol. The viral DNA was redissolved in 100µl TE, purified by extraction with phenol/chloroform, precipitated with ethanol, washed with 1ml of cold ethanol, dried at room temperature, redissolved in 30µl TE and stored at -20 °C.

A sample (3.0µl) of the viral DNA template was subjected to electrophoresis on a 1% agarose gel to check the condition of the template, and to determine whether the DNA insert was incorporated into the vector.

2.14 Sequencing by the Sanger Chain Termination Method

Sequencing was carried out using the protocols supplied by Amserham in the form of "M13 Cloning and Sequencing Handbook" as in the previous cloning section.

2.14.1 Working Solutions

All nucleotide stocks and working solutions were stored at -20 °C.

Deoxy NTP working solutions : 10mM stocks supplied were diluted to 0.5mM working solutions. 0.5mM dATP was not required when

sequencing with α^{35}S-dATP.

Deoxy NTP mixes (A^o, C^o, G^o, T^o) :

	A^o	C^o	G^o	T^o
0.5mM dCTP	20µl	1µl	20µl	20µl
0.5mM dGTP	20µl	20µl	1µl	20µl
0.5mM dTTP	20µl	20µl	20µl	1µl
TE, pH8.0	20µl	20µl	20µl	20µl

Dideoxy NTP working solutions : 10mM stocks supplied were diluted to 0.1mM ddATP, 0.1mM ddCTP, 0.3mM ddGTP, and 0.5mM ddTTP. These concentrations were altered for the sequencing reaction if the need required.

Deoxy NTP/Dideoxy NTP mixes : An equal volume of dNTP was added to the corresponding ddNTP working solution.

2.14.2 Annealing Primer to Template

The primer used was a 17mer universal primer with the sequence 5'd[GTAAAACGACGGCCAGT] 3'. The first stage of the sequencing reaction was to anneal the primer to the single-stranded template. The following reaction was assembled : Single-stranded template DNA (5µl of preparation); and 1µg of primer (Amersham, 1µg/µl) in a final volume of 10µl in Klenow reaction buffer (10mM Tris-HCl, pH8.5; 5mM $MgCl_2$). The mixture was incubated at 60 ^oC for 1 to 2 hr.

2.14.3 Sequencing Reactions

To the annealed primer/template mixture, 15µCi α^{35}S-dATP (Amersham 1mCi/100µl) and 1 unit of Klenow fragment (Boehringer) were added and mixed. The mixture (2.5µl) was placed into each of the four tubes marked A, C, G, and T in a microcentrifuge rotor. The relevant dNTP/ddNTP mix (2µl) was placed inside the rim of each tube and a brief spin mixed the contents. After 20 min, 2µl of chase mixture (0.5mM of all four dNTPs) was placed into each tube, mixed, and allowed to stand for a further 15min. The chase reaction was stopped by the addition of 4µl of formamide dye (0.03% xylene cyanol; 0.03% bromophenol blue; and 20mM EDTA in deionised formamide).

2.14.4 DNA Sequencing Gels

Polyacrylamide urea denaturing gels (6%) were prepared as in section 2.13.3, using a 32 X 2.5mm comb. The samples were boiled for 3 min, then loaded immediately onto the gel. Electrophoresis was performed at 25mA and 40W until the bromophenol blue reaches the bottom of the gel (approximately 2 hr).

Following autoradiography, it was possible to read between 180 to 220 nucleotides from one loading (Figure 2.4b). Two separate loadings were necessary to maximise the length of the sequence that could be read: the first loading was subjected to electrophoresis for 3 to 4 hr; and the second loading for 2 hr. A total of 280 to 330 nucleotides could be read from two loadings. Buffer gradient gels were sometimes used to give up to 280 nucleotides with increased resolution in the lower section of the gel.

2.14.5 Autoradiography

After electrophoresis, the gel was fixed after removing the notch plate by soaking in a 2 litre bath of 10% v/v acetic acid and 10% methanol for 30 min to remove the urea. The gel was drained for a few min, transferred onto a sheet of Whatman 3MM paper, and dried under vacuum on a gel drier (Biorad, model 1125; California) for 30 min at 80 ^{o}C.

After drying, the gel was exposed directly onto Kodak X-Omat H film overnight at room temperature. A longer exposure was sometimes subsequently required.

2.15 Isolation of High Molecular-weight DNA and Genomic Southern Transfer

High molecular-weight genomic DNA was extracted from mouse liver and subsequently purified according to Blattner *et al.* (1978). Southern blotting was used to identify sequences of interest within digests of the genomic DNA.

2.15.1 Isolation of High Molecular-weight DNA from Mouse Liver

Six mice were starved overnight to reduce the glycogen content of their

livers. The mice were killed and their livers quickly removed and frozen in liquid nitrogen. The frozen liver was ground to a fine powder and then added to 100ml of pre-prepared medium as follows. To 0.5M EDTA pH 8.0, 0.5% N-lauroyl sarcosine (Sigma), and proteinase K (100µg/ml) were added and the solution heated for 30 min at 55 °C.

The mixture was incubated for 2 hr at 55 °C in a rotory stirring water bath (200 rpm). After incubation, the DNA was extracted with phenol three times, then dialysed overnight against 4 X 500ml of : 0.05M Tris-HCl, pH 8.0; 0.01M EDTA; and 0.01M NaCl at 4 °C. The solution was removed from the dialysis bag and CsCl was added to 1.273 volume, giving a final density of 1.7g/ml. After mixing carefully, the solution was clarified by centrifugation in a 'Table-top' centrifuge for 15 min, transferred to a sealable tube and centrifuged at 50,000 rpm for 16 to 20 hr at 20 °C in a VTi50 rotor (Beckman).

The DNA was extracted by piercing with a large bore needle (21 gauge) near the bottom of the tube and collecting all fractions. The fractions containing DNA were detected by their high viscosity. These were pooled and dialysed overnight against 4 X 500ml of : 0.01M Tris-HCl, pH 7.0; 0.01M NaCl; 1mM EDTA at 4 °C. The dialysed DNA was stored at 4 °C.

2.15.2 Genomic Southern Transfer

Genomic DNA (10µg) was digested with the appropriate restriction enzyme(s), and loaded into a single lane of a 0.7% agarose gel in a large horizontal tank (Pharmacia GNA-200) with acetate electrophoresis buffer. Electrophoresis was performed at 30V overnight. After electrophoresis the agarose gel was denatured, transferred to a nitrocellulose filter and immobolised (see section 2.9.1). The appropriate ^{32}P-labelled DNA probes (see section 2.9.2) were then hybridised to the genomic DNA attached to the filter (see section 2.9.3). The specific radioactivity of the probes were usually at least 10^8 cpm/µg, and were added to the hybridisation buffer at approximately 10^6 cpm/ml. The filters were continued to hybridise for 48 hr at 42 °C.

After hybridisation, the filters were washed as in section 2.9.3, and autoradiography was performed to locate the position of any bands complementary to the radioactive probe.

2.16 Screening a Bacteriophage Genomic Lambda Library

The genomic lambda library was screened with various ^{32}P-labelled DNA probes to identify the number of plaques that contain the complementary sequence. The plaques were picked and purified for further experiments.

2.16.1 Preparation of Filter Replicas

The library of EMBL 3 was titred (see section 2.4.1), then a dilution of the phages were plated out to give approximately 1,000 plaques on each of six BBL-plates with BBL-top layer agarose containing 10mM $MgSO_4$. A fresh overnight culture (L-broth supplemented with 10mM $MgSO_4$ and 4% maltose; 200µl) of the bacterial host *E.coli* Y1090 (susceptible to the EMBL 3 library) was used for the phage infection.

After incubating the plates at 37 °C overnight, they were cooled at 4 °C for 1 hr. Nitrocellulose filters (Schleicher and Schuell, 9cm diameter) were placed grid-side down onto the agarose surface, and the plates were returned to 4 °C for 20 min. After carefully marking the plate and filter on various asymmetric positions, the filters were removed and transferred through a series of solutions : 0.2M NaOH, 1.5M NaCl for 20s to 5 min; 0.2M Tris-HCl (pH 7.6), 1.5M NaCl for 1 min; 2 X SET (SET = 0.15M NaCl; 30mM Tris-HCl, pH 8.0; 1mM EDTA) until ready for the next stage. The filters were allowed to dry on Whatman 3MM paper for 1 hr at room temperature.

The complete procedure was repeated to obtain the required number of filters per plate. The nitrocellulose filters were baked for 2 hr at 80 °C in a vacuum oven.

2.16.2 Hybridisation of Replica Filters

The replica filters were pre-wet in 4 X SET and then soaked for 1 hr at 65 °C in 10 ml per filter of : 4 SET; 10 Denhardt's solution (see section 2.9.3); and 0.1% SDS.

The filters were pre-hybridised by shaking at 100 rpm for 3 hr at 65 °C in a buffer containing : 4 X SET; 10 X Denhardt's solution; 0.1% SDS; 0.1% sodium pyrophosphate; 50µg/ml sonicated salmon sperm DNA; 50µg/ml each of poly rA, rI, rU, and rC. Each set of filters was stacked together in a 1 litre wide-neck plastic bottle with approximately 1 ml of buffer per filter.

The appropriate ^{32}P-labelled DNA probes were prepared (see section 2.9.2) and denatured (see section 2.9.3), then added to the filters. The probes prepared usually have a specific radioactivity of at least 10^9cpm/μg, and were added to the hybridisation buffer at approximately 1.5 X 10^6 cpm/ml. The filters were continued to hybridise overnight at 65 °C.

After hybridisation, the probe/hybridisation solution was poured off and subsequent washes were performed as follows : 4 X SET, 0.1% SDS, 0.1% sodium pyrophosphate for 2 X 20 min at 65 °C; 2 X SET, 0.1% SDS for 4 X 15 min at 45 °C; 0.2 X SET, 0.1% SDS for 15 min at 45 °C; 3mM Tris-HCl (unbuffered) for 1 hr at room temperature.

The filters were then allowed to dry on Whatman 3MM paper for 30 min. Autoradiography was performed to identify the clones in the genomic library that contained the complementary sequence of the DNA probes.

2.16.3 Plaque Purification

The plaques giving the positive signal on autoradiography were identified and their positions marked on the photograph film. After lining the film to its respective plate, the area of plaques that coincided with the positions were picked using the wide end of a pasteur pipette, and inoculated into 1ml of phage buffer (see section 2.1.1) in a glass tube. The glass tubes were left to stand at room temperature for 1 hr, then a drop of chloroform was added to kill off the bacteria. The phage was stored at 4 °C until further use.

Dilutions of the phage solution were plated to obtain 100 to 300 plaques on BBL-plates (see section 2.16.1). The plates were re-hybridised as before, and the positives identified. This procedure was repeated until a low density plate containing more than 50% positive plaques were obtained, then a well isolated plaque was picked and used.

2.17 Computer Programs for the Analysis of DNA Sequences

The following programs were used in the compilation, manipulation and analysis of DNA sequences. A number of programs devised by Staden (1978), were run on a Digital PDP 11-34 computer, with a multi-user facility in the Biochemistry Department, University of Glasgow. Other programs of the UWGCG (University of Wisconsin Genetics Computer Group) package (Devereux *et al.*, 1984) were run on the EMBL (European Molecular Biology Laboratory) VAX

11/785 and VAX 8600 computers. This package contains programs for the analysis and investigation of DNA sequences and comparison of sequences with those in the EMBL database (EMBL, Heidelberg, W. Germany).

2.17.1 Staden Programs

SEQEDT :this program was used to create and edit a file for DNA sequences.

SEQLST : lists the sequence file created by SEQEDT in the Staden format.

TRNTRP: translates nucleotide sequences into peptide sequences in any desired reading frame using the three-letter amino-acid code.

SEARCH: searches sequences for restriction sites and strings of sequences of no more than 20 bases.

SEQFIT : searches sequence for similarities with a string of sequence less than 200 bases, and can also be used for percentage complementation.

SQRVCM: generates a sequence complementary to the sequence in question.

CUTSIT : compares given sequence file with restriction enzyme file and lists all known restriction sites within the sequence.

2.17.2 Other Programs

These two programs were devised by Dr. P. Taylor (Department of Virology, University of Glasgow), and were run on the Digital PDP 11-34 computer.

PALIGN: compares two sequence files with a maximum of 2048 characters. This program uses the blocks that satisfy the minimum number of matches to obtain the best alignment and then align the remaining to the best. However, it has limitations and sometimes misses the match.

CINTHOM: creates a homology matrix plot between two sequence files .

2.17.3 UWGCG Programs

FIND : searches through sequence(s) for short sequence patterns. It is able to look through large data sets for any sequence patterns specified, recognise patterns with some symbols mismatched but not with gaps, and searches both strands of the sequence if necessary. Patterns may not be more than 41 characters long.

BESTFIT: finds the best region of similarity between two sequences, and inserts gaps to obtain the optimal alignment. The sequences can be of very

different lengths but the program cannot evaluate a surface of comparison larger than 10^6 base squared, with input sequences not more than 30,000 symbols long.

GAP : produces an optimal alignment between two sequences by inserting gaps in either one as necessary. It considers all possible alignments and gap positions, and creates the alignment with the largest number of matched bases and the fewest gaps.

REPEAT : finds repeats in sequences. It allows one to choose a minimum repeat window, stringency, a search range and then finds all the repeats with these parameters.

STEMLOOP: finds stems (inverted-repeats) in nucleic acid sequences. It allows one to choose a minimum stem length, maximum loop size and minimum bonds per stem. The stems found can be sorted by position, size (stem length), or quality (number of bonds).

FOLD : finds an 'optimal' secondary structure for an RNA molecule with minimum free energy.

Since the programs WORDSEARCH with SEGMENTS were used extensively throughout this project, they will be described in more detail.

WORDSEARCH tries to find places where one sequence is similar to any set of other sequences. The search finds diagonals in each comparison that have the largest number of common words. A word is any short sequence (n-mer) where n is preset to a constant, like 6 or 7, and it can be created from an alphabet consisting of the four letters G, A, T, and C. A diagonal is a path across a surface of comparison where X minus Y for every point is a constant. A series of dots along a diagonal represent a segment of similarity between two sequences. WORDSEARCH makes and sorts the scores of all the diagonals in the comparison, and shows a list of the N best diagonals, where N is pre-selected to be some finite number, like 25 or 100. WORDSEARCH is able to compare both strands of the query sequence to any set of sequences, and shows the specified number of best diagonals and the number of words on each of these diagonals. The best segments of similarity on or near the diagonals can be viewed with the program SEGMENTS.

The strategy used by WORDSEARCH/SEGMENTS is to use word comparison, to identify "regions of possible similarity" between a query sequence and some sets of sequences, and then to use optimal alignment to display the best segment of similarity in each segment. SEGMENTS uses a symbol comparison table, a gap

weight, and a gap length weight to find the best region of similarity between two sequences.

SEGMENTS uses symbol comparison values of 1.00 for each nucleotide match and -0.60 for every mismatch, to construct a path matrix that represents the entire surface of the comparison in a score at every position for the best possible alignment path to that point. Random alignments should have a path-value that averages about zero. The **gap weight** and **gap length weight** are user-variable penalties for the creation of a gap and for the number of nucleotides over which the gap extends, respectively. The best region has the highest **quality,** where for each alignment, the **quality** is equal to the sum of matches, minus 0.6 times the sum of mismatches, minus the gap weight times the sum of gaps, minus the gap length weight times total length of all the gaps.

Hence :

$$\text{Quality} = 1.0 \times \text{matches} - 0.60 \times \text{mismatches} - (\text{gap weight} \times \text{gap number}) - (\text{gap length weight} \times \text{total length of gaps})$$

The output of the program includes an area extending beyond the highest scoring region of a particular diagonal. In addition to the quality, which relates only to this highest scoring region, an important indicator of the comparison is the **ratio.** This is in effect the quality over the total length of the diagonal in the output.

Hence :

$$\text{Ratio} = \frac{\text{Quality}}{\text{Length of diagonal}}$$

An example of one of the output files in SEGMENTS that was used to display segments in the output file for WORDSEARCH is given in Figure 4.2.

CHAPTER 3

ANALYSIS OF ACTIN PSEUDOGENES

The first part of this work was focused on the actin-like DNA of genomic clones λmA82, λmA119 and λmA118, in order to establish whether they were functional genes or pseudogenes (the actin-like DNA of λmA36 was analysed by others as part of a different project in the same laboratory). Restriction analysis of these genomic clones was carried out in order to provide a basis for subcloning and sequencing, and to determine whether they represented different genomic regions.

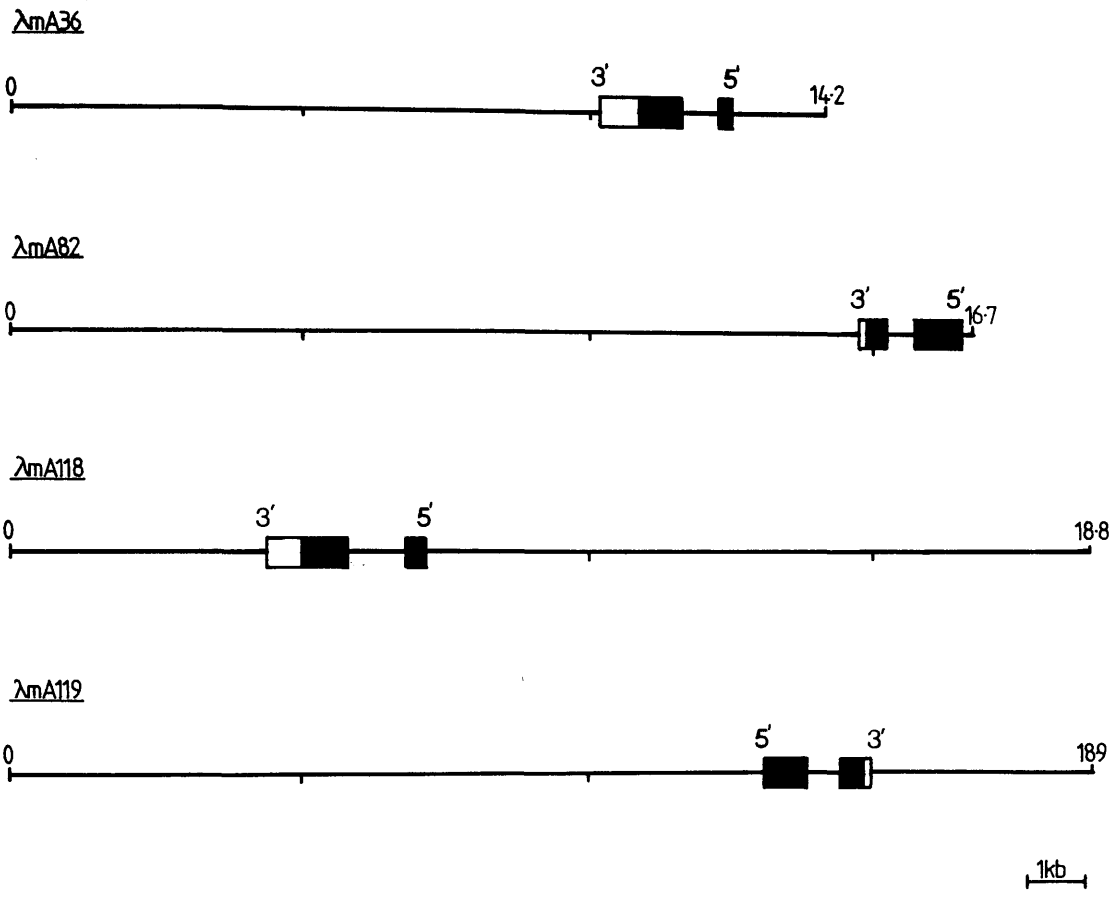
3.1 Restriction Analysis of the Genomic Clones

The objective of the restriction mapping was limited in the first instance to locating restriction sites near the actin-like region of the three genomic clones λmA82, λmA118 and λmA119.

The approach adopted was as follows. Single restriction digestion was performed on the genomic clones with suitable restriction enzymes that cleaved the DNA relatively infrequently. The fragments produced were hybridised to a ^{32}P-labelled actin DNA probe (PstI fragment from cDNA clone pmS3, containing most of the mouse skeletal muscle actin coding region, and 100 base pairs of the 3' non-coding region; Figure 2.3a) and fragments which contain all or part of the actin-like region were identified. The position of the actin-like regions had previously been determined by electron microscopic heteroduplex analysis (Figure 1.4; H. Delius, EMBL, Heidelberg), is shown in Figure 3.1. By combining this latter information with the above hybridisation results, a partial restriction map was constructed for the three genomic clones (Figure 3.2). Although this was incomplete it nevertheless provided an adequate basis for devising a subcloning strategy (see section 3.2)

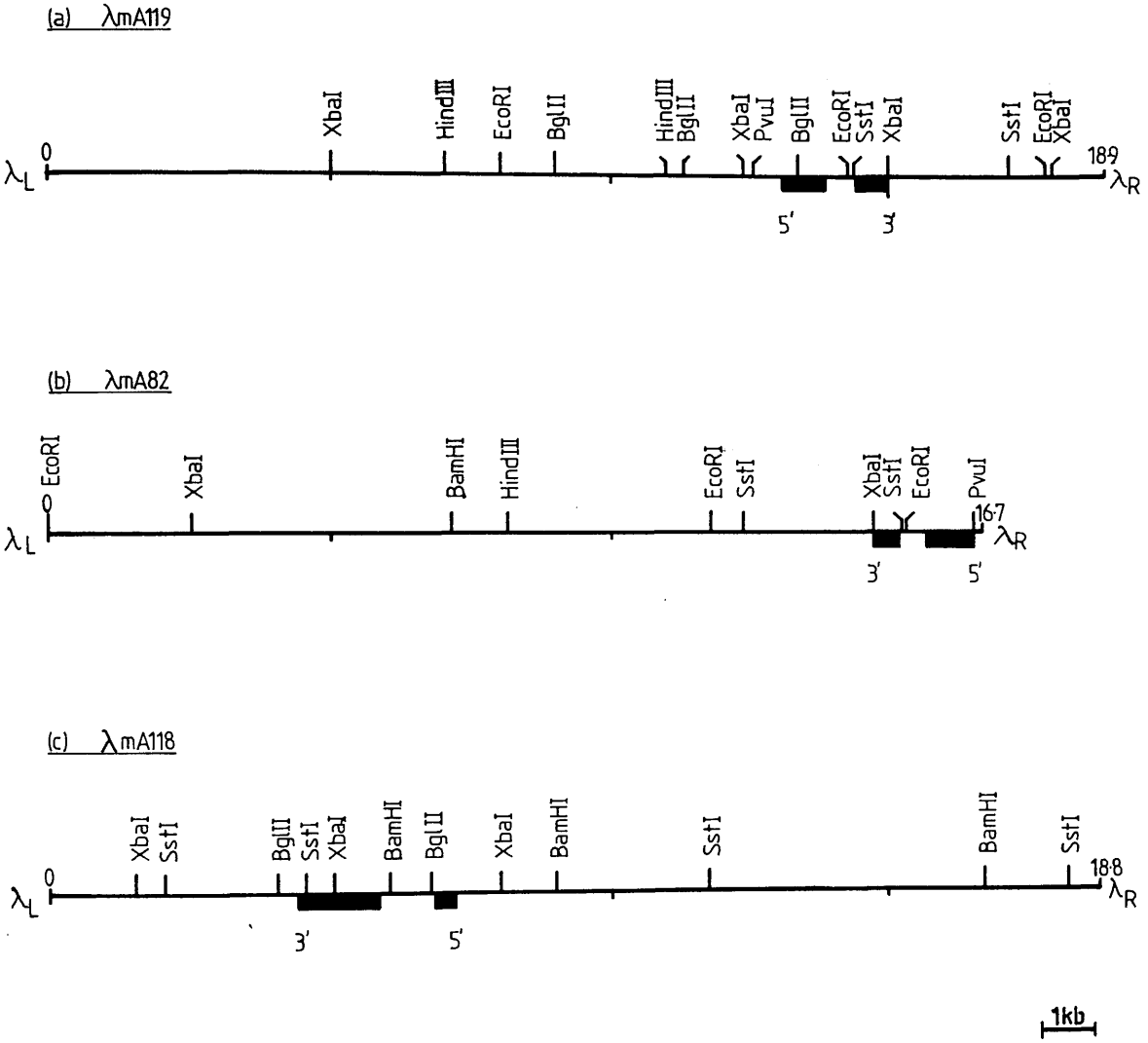
When the partial restriction maps of λmA119 and λmA82 were compared (Figure 3.2), it was observed that both clones contained some common

Figure 3.1 Physical maps of some mouse actin-like genomic clones



The position and orientation of the actin-like regions within each genomic clone deduced from the electron microscopic measurements of heteroduplexes with clones containing pseudogenes of known structure. The solid blocks represent the presumed positions of actin coding regions, and open blocks represent the presumed positions of 3' non-coding regions. In each case the pseudo-coding region is interrupted by extra DNA.

Figure 3.2 Partial restriction maps for clones λmA119, λmA82, and λmA118



Partial restriction maps are shown for the mouse DNA inserts of genomic clones λmA119, λmA82 , and λmA118. The solid blocks represent the presumed positions of the actin-like gene, and the vertical lines indicate the approximate positions of restriction sites. The restriction sites HindIII and XbaI are incomplete for (a) ; EcoRI and XbaI are incomplete for (b) ; and BamHI and BglII are incomplete for (c). λ_R and λ_L indicate the right-hand (9kb) and left-hand (20kb) arms of the lambda vector.

restriction sites (*eg.* SstI, EcoRI) at similar positions with respect to the actin-like region. This, together with the similar size and position of their insertions suggested that clones λmA119 and λmA82 possibly contained overlapping regions of DNA, but in opposite orientations (Figure 3.3). Different digestion patterns were observed when λmA119 and λmA82 were digested with BglII, XbaI, EcoRI, and SstI. However, when the two clones were double digested with BglII and XbaI, 1.4 kb fragments of identical mobility containing actin-like DNA were identified when the digest was blotted with a ^{32}P-labelled actin DNA probe (Figure 3.4a). In the sequenced uninterrupted γ-actin pseudogene, λmA19 (Leader *et al.*, 1985), there is a BglII site at amino-acid 84 and a XbaI site at amino-acid 374 , 0.9 kb away. Thus, when the 0.5 kb extra DNA is allowed for, it is evident that the 1.4 kb BglII-XbaI fragment from clones λmA119 and λmA82 could have been generated from cleavage at sites corresponding to those in λmA19. In addition when the two clones were double digested with BglII and SstI, 2.8 kb and 1.0 kb fragments of identical mobility containing actin-like DNA were identified (Figure 3.4b). This together with the results of other digestions (Figure 3.3), implied that the two clones λmA119 and λmA82 did indeed originate from the same genomic region. Therefore one only of these clones was taken for further analysis, λmA119 being chosen on the basis that it contained more flanking DNA 5' to the actin-like region.

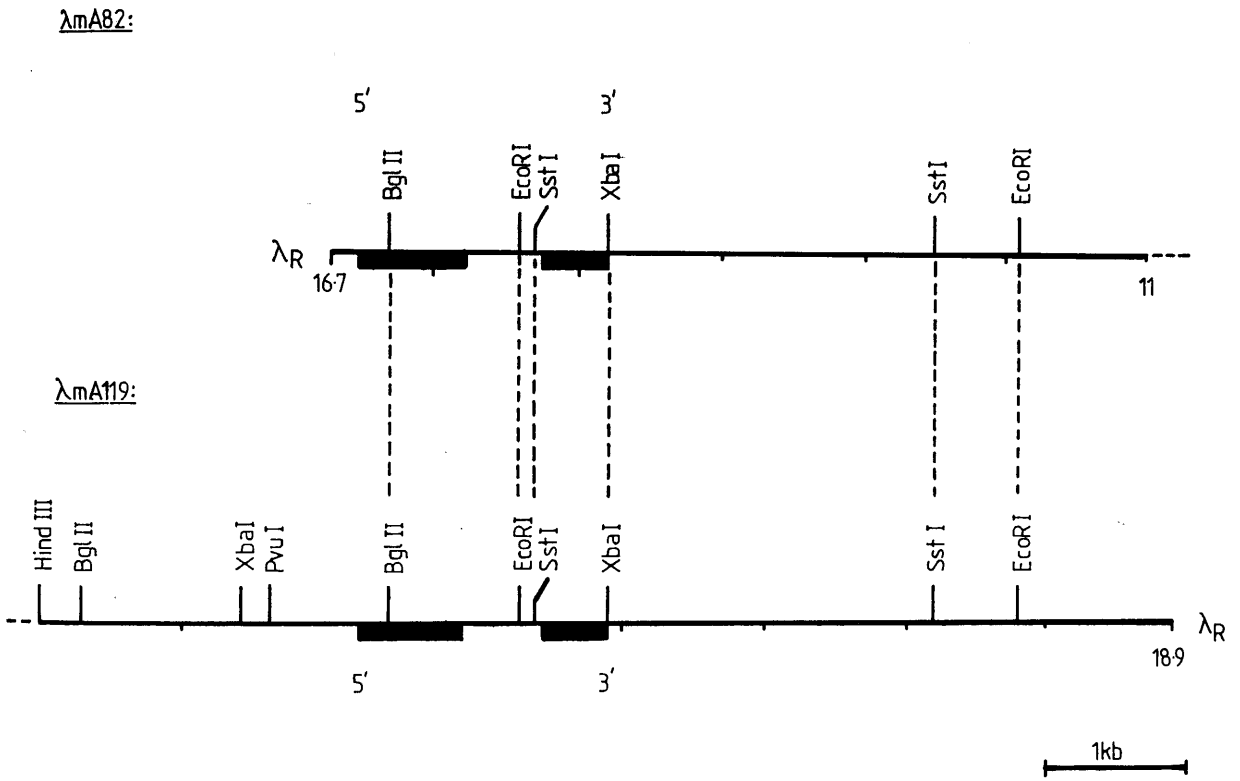
3.2 Subcloning Strategy

3.2.1 Subcloning of the Genomic Clone λmA118

On the basis of the partial restriction map of λmA118 in Figure 3.2, it was decided that the 2.6 kb XbaI fragment was likely to contain the actin coding DNA and was of a suitable size for subcloning. An XbaI subclone, was constructed (see section 2.10.6), and was designated 118Y1-1. Restriction analysis and hybridisation to the actin probe confirmed that it contained the desired 2.6 kb fragment.

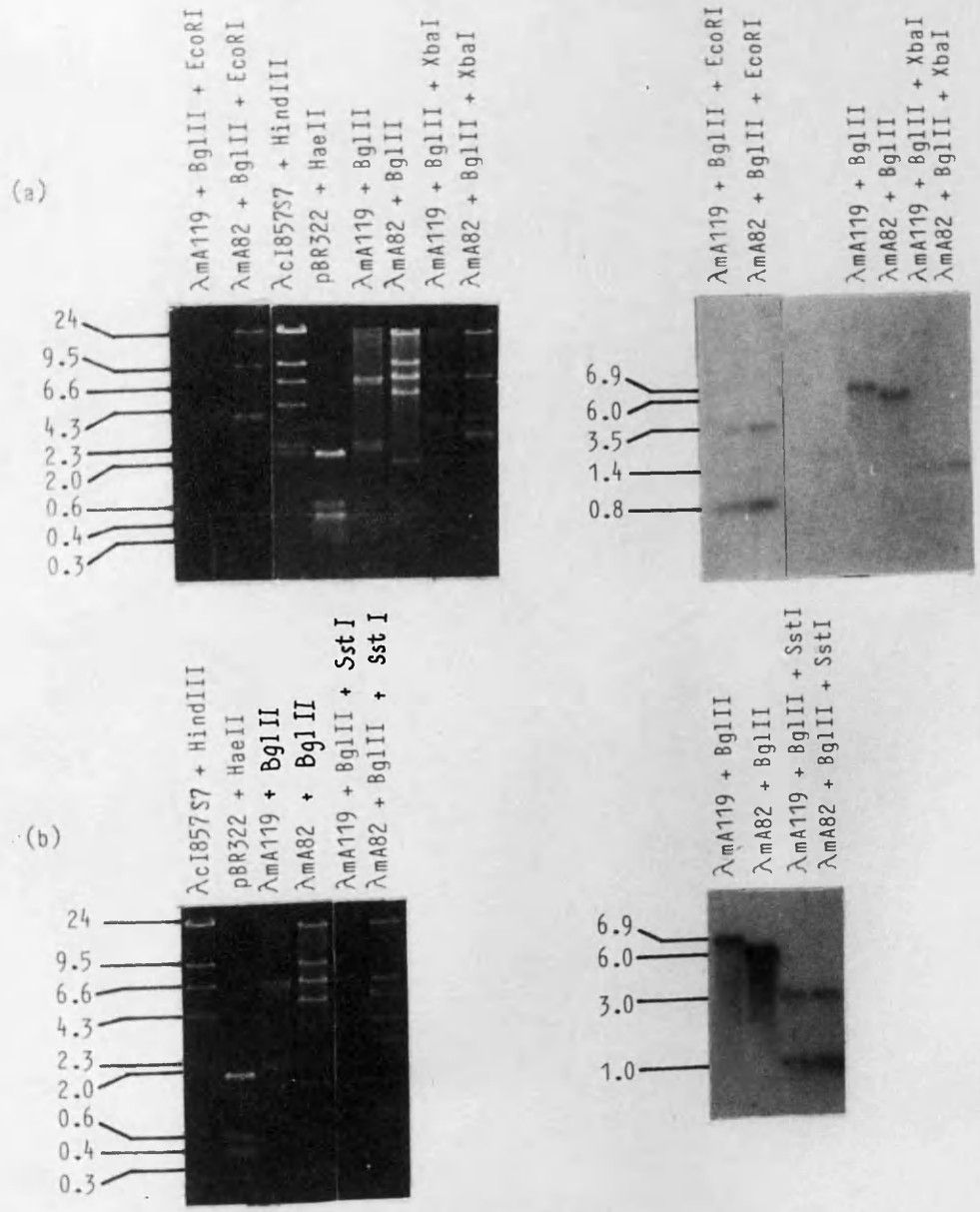
Restriction analysis was carried out on the subclone 118Y1-1 and a more detailed restriction map was constructed as shown in Figure 3.5. In order to assist the sequencing, further subclones from the original parent subclone 118Y1-1 were derived from the two internal PstI sites. A total of three subclones were constructed using : a 0.7 kb PstI-XbaI fragment ; a 1.1 kb XbaI-

Figure 3.3 Comparison of partial restriction maps between clones λmA82 and λmA119



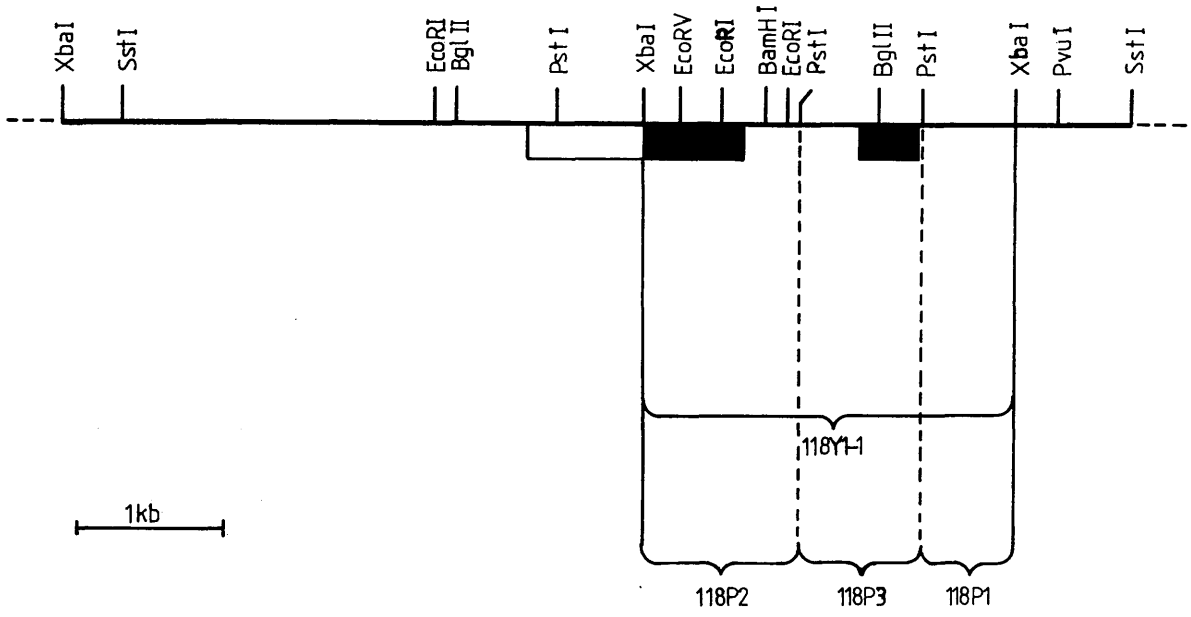
The partial restriction maps of genomic clones λmA119 and λmA82 are compared in the region of the actin-like genes. The solid blocks represent the presumed positions of the actin-like gene, the vertical lines indicate the approximate positions of restriction sites, and the vertical dotted lines indicate the alignment of restriction sites. λ_R indicates the right-hand (9kb) arm of the lambda vector.

Figure 3.4 Restriction digestions of clones λmA119 and λmA82 and hybridisation to radioactive-labelled actin probes



Double restriction digestions of clones λmA119 and λmA82 were performed, followed by hybridisation to a ^{32}P-labelled PstI fragment of actin cDNA clone pmS3 (Figure 2.3). (a) shows the digestion pattern of λmA119 and λmA82 digested with BglII, BglII and EcoRI, and BglII and XbaI; and an autoradiograph of the hybridisation. (b) shows the digestion pattern of λmA119 and λmA82 digested with BglII, and BglII and SstI; and an autoradiograph of the hybridisation.

Figure 3.5 Partial restriction map of clone λmA118 and subclones in the vicinity of the actin-like gene



A more detailed partial restriction map of a portion of genomic clone λmA118 is presented with reference to its derived subclones. The solid blocks represent the presumed positions of the actin coding regions, the open blocks represent the presumed positions of 3' non-coding regions, the vertical lines indicate the approximate positions of restriction sites. The derived subclones 118Y1-1, 118P1, 118P2, and 118P3 are indicated.

PstI fragment ; and a 0.9 kb PstI-PstI fragment of 118Y1-1, and were designated 118P1, 118P2, and 118P3 respectively (Figure 3.5).

3.2.2 Subcloning of the Genomic Clone λmA119

Three primary subclones were derived from λmA119 to encompass both coding and non-coding regions of the actin-like gene.

A subclone was derived from the 2.5 kb XbaI fragment of λmA119 (Figure 3.2; and section 2.10.6), and was designated 119X1-1. Restriction analysis and hybridisation to the actin probe confirmed that it contained the desired 2.5 kb fragment.

A further two subclones which hybridised to a γ-actin 3' non-coding probe (a ^{32}P-labelled XbaI-PstI fragment from a subclone of λmA19: Figure 2.3c) were derived from a 2.8 kb SstI fragment and a 3.1 kb XbaI fragment of λmA119, and were designated 119SS and 119X2-1, respectively (Figure 3.6).

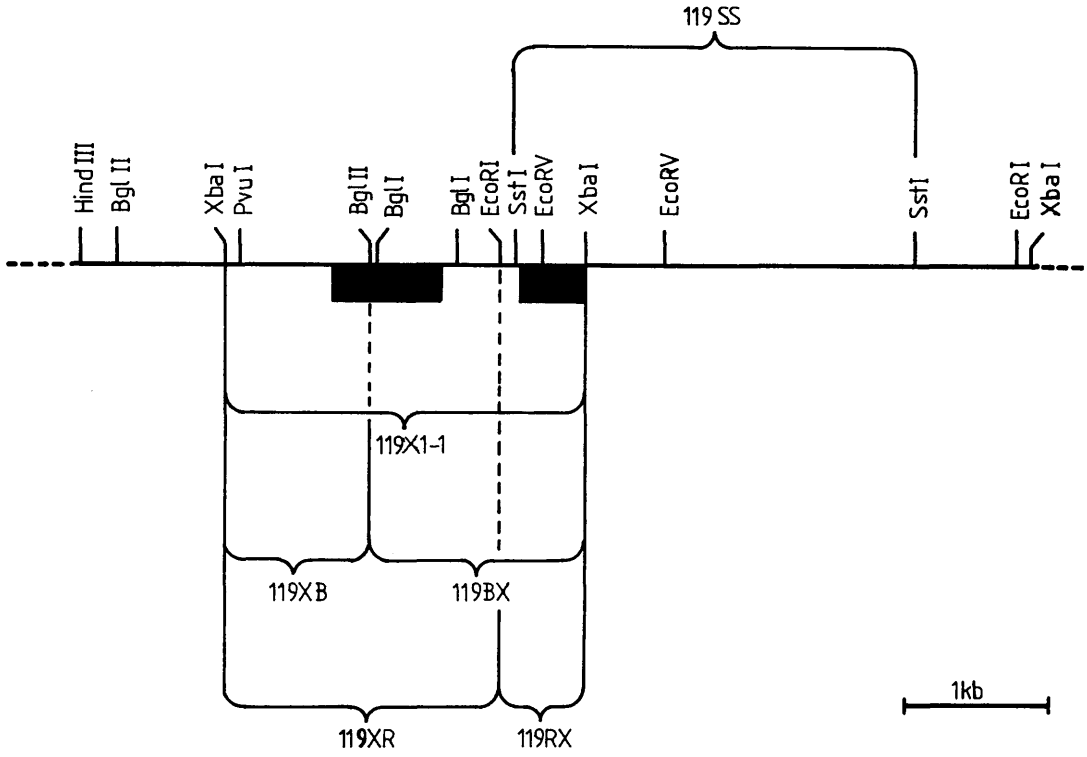
More detailed restriction analysis was performed on the subclone 119X1-1, allowing a further subcloning strategy to be devised (Figure 3.6). A total of four subclones were constructed using : a 1.0 kb XbaI-BglII fragment; a 1.5 kb BglII-XbaI fragment ; a 1.9 kb XbaI-EcoRI fragment ; and a 0.6 kb EcoRI-XbaI fragment of 119X1-1, and were designated 119XB, 119BX, 119XR and 119RX, respectively.

3.3 Sequencing

3.3.1 Sequencing of the Genomic Clone λmA118

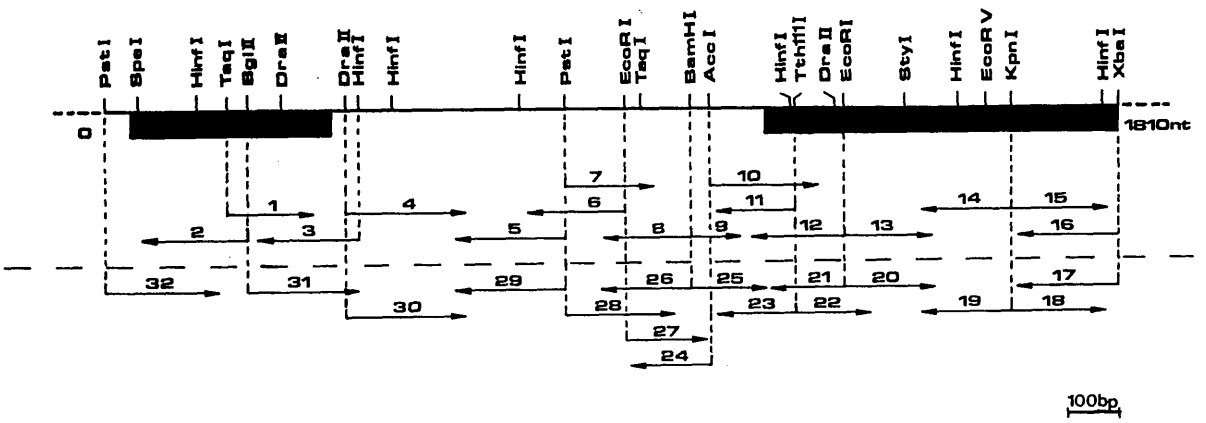
The sequencing strategy was based on the detailed partial restriction map of λmA118 and its derived subclones (Figure 3.5). Sequencing was performed by the chemical method of Maxam and Gilbert (see sections 2.11 and 2.12). The PstI and XbaI sites define the limits of the region sequenced, which included the coding portion of the actin-like region and the interrupted sequence (see section 4.1), as summarised in Figure 3.7. Sequencing of the actin region of clone λmA118 was as follows. Sequencing from the PstI, TaqI, and BglII sites (sequences number 1, 2, 31, and 32) allowed the region from Ile^{5} to Leu^{140} to be sequenced, while sequencing from the Tth111I, EcoRI, KpnI (Asp718), and XbaI sites (sequences number 11 to 23) allowed the completion of sequencing from

Figure 3.6 Partial restriction map of clone λmA119 and subclones in the vicinity of the actin-like gene



A more detailed partial restriction map of a portion of genomic clone λmA119 is presented with reference to its derived subclones. The solid blocks represent the presumed positions of the actin coding regions, the open blocks represent the presumed positions of 3' non-coding regions, the vertical lines indicate the approximate positions of restriction sites. The derived subclones 119X1-1, 119XB, 119BX, 119XR, and 119RX are indicated.

Figure 3.7 Sequencing strategy for the actin-like region and interrupted DNA of the genomic clone λmA118



The strategy for the determination of the nucleotide sequence in the region from the PstI site at the 5' end of the actin-like DNA to the 3' XbaI site in λmA118 is shown. Sequencing of the coding and non-coding strands with respect to the actin pseudogene are indicated by arrows below and above the broken line, respectively. Fragments are numbered sequentially for ease of reference. The arrows represent the portion of sequence read from a particular restriction site.

Leu^{140} to Phe^{374}.

The complete sequence was built up from a large number of overlapping fragments. Each part of the sequence was determined at least once, and 94% of the sequence was determined from both strands.

Compilation of the total sequence data gives the complete nucleotide sequence determined in the genomic clone λmA118, as shown in Figure 3.8. Both strands of the nucleotide sequence are shown, together with a number of restriction sites which were useful in the determination of the sequence. Comparison of this nucleotide sequence with other known sequences is made in subsequent sections.

3.3.2 Sequencing of the Genomic Clone λmA119

The sequencing strategy was based on the detailed partial restriction map of λmA119 and its derived subclones (Figure 3.6). Sequencing was performed by the chemical method of Maxam and Gilbert (see sections 2.11 and 2.12). The region sequenced was from approximately 260 nucleotides 5' of the BglII site to the XbaI site. This included the coding portion of the actin-like region and the interrupted sequence (see section 4.2), as summarised in Figure 3.9. Sequencing of the actin region of clone λmA119 was as follows. Sequencing from the BglII and HinfI sites (sequences number 1, 2, and 19 to 21) allowed the region from Met^{1} to Thr^{160} to be sequenced, while sequencing from the SstI and XbaI sites (sequences number 8 to 12) allowed the region from Phe^{265} to Phe^{374} to be sequenced. In order to identify further restriction sites suitable for sequencing, a 0.9 kb BglII-SstI fragment was isolated from subclone 119BX (see section 2.7.3), and subjected to extensive restriction analysis. The eventual identification of the restriction sites XhoII and StyI enabled further sequencing from these sites (sequences number 3 to 5, and 16 to 18) and allowed the completion of sequencing from the region Thr^{160} to Phe^{265} of the actin region.

A total of approximately 250 nucleotides were determined 3' to the stop codon at the XbaI site, although only 10 bases of the region 5' to actin-like DNA were covered. The complete sequence was built up from a large number of overlapping fragments. Each part of the sequence was determined at least once, and 97% of the sequence was determined from both strands.

Compilation of the total sequence data gives the complete nucleotide sequence determined in the genomic clone λmA119, as shown in Figure 3.10.

Figure 3.8 Nucleotide sequence of interrupted actin pseudogene determined in the genomic clone λmA118

```
        PstI
1       CTGCAGGCTA CACTGCGCTT CTTGCCGCTG CTCCATCGCC AATCAATCGC AATAGCCGCA   60
        GACGTCCGAT GTGACGCGAA GAACGGCGAC GAGGTAGCGG TTAGTTAGCG TTATCGGCGT

61      CTAGTCATTG ACAATGGCTC CGGCACGTCA ATGACAACGC CCTCAGGGCC ATGTTCCCTT  120
        GATCAGTAAC TGTTACCGAG GCCGTGCAGT TACTGTTGCG GGAGTCCCGG TACAAGGGAA

121     CCATCATAGG GCGCCCCCGA CACCAGGGTG TCTTGGTGGG CATTGGCCAG AAGGACTCCT  180
        GGTAGTATCC CGCGGGGGCT GTGGTCCCAC AGAACCACCC GTAACCGGTC TTCCTGAGGA
                                                          TaqI
181     ACGTGGGTGA TGAGGCCCAG AGCAAGAGGG GTATCCTGGC CCTGAAGTAC CCTGTCGAGC  240
        TGCACCCACT ACTCCGGGTC TCGTTCTCCC CATAGGACCG GGACTTCATG GGACAGCTCG
                                    BglII
241     ATGGCATTGT CACCAACTGG GACGACATGG AGAAGATCTG GCACCACACC TTCTACAATG  300
        TACCGTAACA GTGGTTGACC CTGCTGTACC TCTTCTAGAC CGTGGTGTGG AAGATGTTAC

301     AGCTGCGTGT GGCCCCTGAG GAGCACCCGG TGCTACTGAC CGAGGCCCCC CTGAACCCCA  360
        TCGACGCACA CCGGGGACTC CTCGTGGGCC ACGATGACTG GCTCCGGGGG GACTTGGGGT

361     AAGCTAACAG AGAGAAGATG ACGCAGATAA TGTTTGAACC CTTCAATACC CCAGCCTTGT  420
        TTCGATTGTC TCTCTTCTAC TGCGTCTATT ACAAACTTGG GAAGTTATGG GGTCGGAACA
                                                    DraII
421     ACGTCACCAT TCAGGTGGTG CTTGGTCACG GACTGGGGCT CTGTGGGCCC TCTTCGGTCT  480
        TGCAGTGGTA AGTCCACCAC GAACCAGTGC CTGACCCCGA GACACCCGGG AGAAGCCAGA
          HinfI
481     GCGGAATCAG AGTCTCAGAC AGATGGGCAT AGAGTGGGCG AGTGACAAAC AGACGTGACA  540
        CGCCTTAGTC TCAGAGTCTG TCTACCCGTA TCTCACCCGC TCACTGTTTG TCTGCACTGT

541     AGAGAACGTG TTGAATCTGA GTGTAATTTA TCAAATCCAG CATCAAACTT TTTATACAGA  600
        TCTCTTGCAC AACTTAGACT CACATTAAAT AGTTTAGGTC GTAGTTTGAA AAATATGTCT

601     ATAACAAGAA ACCAGGCGAA CACATCCGCT AAGTTACAGT GACACAAAAC AAAAGGAATG  660
        TATTGTTCTT TGGTCCGCTT GTGTAGGCGA TTCAATGTCA CTGTGTTTTG TTTTCCTTAC

661     CATACATCAA AAGATGGCGG GGACCAAGCT CATTACCACT AGAAGGAACA GGTGTAATGC  720
        GTATGTAGTT TTCTACCGCC CCTGGTTCGA GTAATGGTGA TCTTCCTTGT CCACATTACG

721     TAGTCTATTG TTAAACCCAC CACCAAGGGG TTCTTAGTAA ATGCCTGATT ATGCTGTTCC  780
        ATCAGATAAC AATTTGGGTG GTGGTTCCCC AAGAATCATT TACGGACTAA TACGACAAGG

781     TTTGGGCCTA GTGAAGAAAC CTGTCCAAGG GGGATTCCCT AACTCTTTCA TGGTTACCCC  840
        AAACCCGGAT CACTTCTTTG GACAGGTTCC CCCTAAGGGA TTGAGAAAGT ACCAATGGGG
                                                          PstI
841     ACCTATTTGC TAGGCCATTG TGTCCTAAGG CTACTGTCCT AAATAATCAC TCTGCAGACT  900
        TGGATAAACG ATCCGGTAAC ACAGGATTCC GATGACAGGA TTTATTAGTG AGACGTCTGA

901     AGCCCTGAGC TATTCTAGCT CCGTTCGGAG CACTGGGTGC TCCTCAGGGG CCACACACAC  960
        TCGGGACTCG ATAAGATCGA GGCAAGCCTC GTGACCCACG AGGAGTCCCC GGTGTGTGTG
                                                            EcoRI
961     GCTTCTCTAC TAGAAGTAAA TTTGAATGTT ACTGAATAGG TAACCTTCTC ACTGAATTCC 1020
        CGAAGAGATG ATCTTCATTT AAACTTACAA TGACTTATCC ATTGGAAGAG TGACTTAAGG

1021    CACTAAATTC CAAGCTCCTC GGCGTCGAGG ATTTTCTAGG ACATTGCAAC ACTGGCGAAG 1080
        GTGATTTAAG GTTCGAGGAG CCGCAGCTCC TAAAAGATCC TGTAACGTTG TGACCGCTTC
```

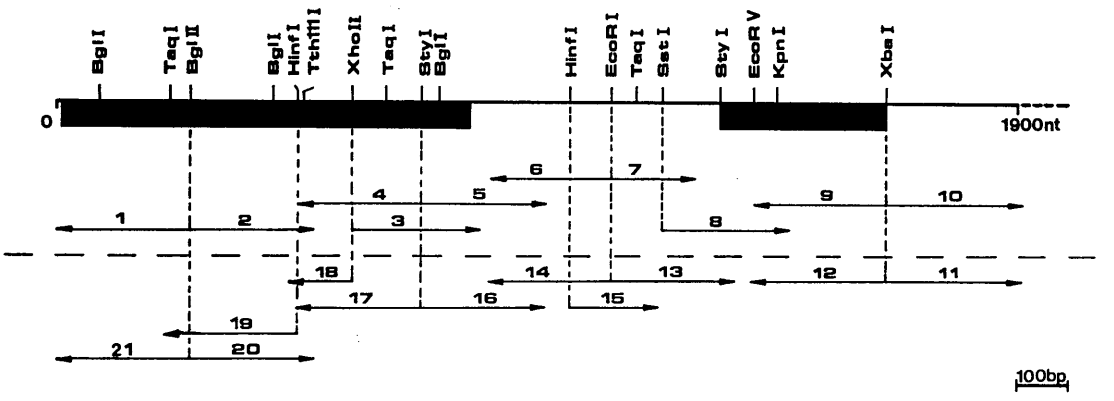
continued overleaf. . . .

continued. . . .

```
1081  GCTTAGCTAT GTCAAGCAAT CAAATCTTAA AGGCACTTAT AATAAAACAA TACTGAAAGA 1080
      CGAATCGATA CAGTTCGTTA GTTTAGAATT TCCGTGAATA TTATTTTGTT ATGACTTTCT
            BamHI                                          AccI
1141  GAGCACGTGG ATCCATACAC CAAACTAACA CGGGAAAAGG GTTTGAGTAT ACGGGCTATG 1200
      CTCGTGCACC TAGGTATGTG GTTTGATTGT GCCCTTTTCC CAAACTCATA TGCCCGATAC

1201  GGAATGCCAA GGTTCCAGGA GGCATAGTTT CCTTGAAACT CATTGCCTCG TGAGTGTTTC 1260
      CCTTACGGTT CCAAGGTCCT CCGTATCAAA GGAACTTTGA GTAACGGAGC ACTCACAAAG

1261  CAGGCCTCTT GGCCAGTCAA GCAGACTTCA CCGGAGTGGG CGTAGGAGGT GCTCTCCTTG 1320
      GTCCGGAGAA CCGGTCAGTT CGTCTGAAGT GGCCTCACCC GCATCCTCCA CGAGAGGAAC
                                                  Tth111I
1321  TATGTATCTG GGCGCACCAC TGGCATTGTC ATGGACTCTG GTGACGGGGT CACACACACA 1380
      ATACATAGAC CCGCGTGGTG ACCGTAACAG TACCTGAGAC CACTGCCCCA GTGTGTGTGT

1381  GTGGCCATCT ATGACAGCTA CACCCTTCCT CACGCCATCT TGTGTCTGGA CTTGGTTGGC 1440
      CACCGGTAGA TACTGTCGAT GTGGGAAGGA GTGCGGTAGA ACACAGACCT GAACCAACCG
                              EcoRI
1441  TAGGACCTGA CAGAGTACCT CATGAATTCC TTGACTGAAC GGGGCTACAG CTTTACCACC 1500
      ATCCTGGACT GTCTCATGGA GTACTTAAGG AACTGACTTG CCCCGATGTC GAAATGGTGG

1501  ACTGCTGAGA GGGAAATTGT GACAAGGAGA AGCTGTGCTA TGTTGCCCTG GATTTTGAGC 1550
      TGACGACTCT CCCTTTAACA CTGTTCCTCT TCGACACGAT ACAACGGGAC CTAAAACTCG

1561  AAGAAAAGGC TACTGCTGCA TCATCTTCCT CCTTGGAGAA GAGTTACCAG CTGCCCGATG 1620
      TTCTTTTCCG ATGACGACGT AGTAGAAGGA GGAACCTCTT CTCAATGGTC GACGGGCTAC

1621  GGCAGGTGAT CACCATTGGC AATGAGCGGT TCCGGTGTCC GGAGGCACTC TTCCAGCATT 1680
      CCGTCCACTA GTGGTAACCG TTACTCGCCA AGGCCACAGG CCTCCGTGAG AAGGTCGTAA

1681  CCTTCCTGGG CATGGAATCC TGTGGCATCT ACGAGACCAC CTTCAACTCC ATCATGAAGT 1740
      GGAAGGACCC GTACCTTAGG ACACCGTAGA TGCTCTGGTG GAAGTTGAGG TAGTACTTCA
                                                    KpnI
1741  GTGATGTGGA TATCTGCAAA GACCTGTATG CCAATACAGT GCTGTCCGGT GGTACCACCA 1800
      CACTACACCT ATAGACGTTT CTGGACATAC GGTTATGTCA AGACAGGCCA CCATGGTGGT

1801  TGTACCCAGG CATTGCTGAC AGGATGTAGA AGGAGATCAC AGCCCTAGCA CCCAGCACAA 1860
      ACATGGGTCC GTAACGACTG TCCTACATCT TCCTCTAGTG TCGGGATCGT GGGTCGTGTT

1861  TGAAGATTAA GATCATTGCT CCCCCTGAGC GCAAGTACTC AGTCTGGACC TGCGGCTCCA 1920
      ACTTCTAATT CTAGTAACGA GGGGGACTCG CGTTCATGAG TCAGACCTGG ACGCCGAGGT

1921  TCCTACCTCA CTGTCCACCT TCCAGCAGAT GTGGATCAGC AAGCAGGAGT ATGATGAGTC 1980
      AGGATGGAGT GACAGGTGGA AGGTCGTCTA CACCTAGTCG TTCGTCCTCA TACTACTCAG
                                  XbaI
1981  GGGCCCATCG TCCACCGCAA ATGCTTCTAG A  2011
      CCCGGGTAGC AGGTGGCGTT TACGAAGATC T
```

Both strands of the nucleotide sequence of the interrupted actin pseudogene in genomic clone λmA118 are shown, the PstI and XbaI sites (Figure 3.7) defining the limits of the region sequenced. Restriction sites are underlined above the appropriate nucleotides. The boxed sequences are direct repeats flanking the inserted sequence.

Figure 3.9 Sequencing strategy for the actin-like region and interrupted DNA of the genomic clone λmA119



The strategy for the determination of the nucleotide sequence in the region from the PstI site at the 5' of the BglII site of the actin-like region to the 3' XbaI site in λmA119 is shown. Sequencing of the coding and non-coding strands with respect to the actin pseudogene are indicated by arrows below and above the broken line, respectively. Fragments are numbered sequentially for ease of reference. The arrows represent the portion of sequence read from a particular restriction site.

Figure 3.10 Nucleotide sequence of interrupted actin pseudogene and flanking regions determined in the genomic clone λmA119

```
1    CGTCGCAATG GAAGAAGAAA TCGCCACACT CGTCATTGTC AATGGCTCCG GCATGTGCAA   60
     GCAGCGTTAC CTTCTTCTTT AGCGGTGTGA GCAGTAACAG TTACCGAGGC CGTACACGTT

61   AGCTGGCTTT GCTGGAGACG ACGCCCCCAG GGCCGTGTTC CCTTCCATCG TAGGGTGCCC  120
     TCGACCGAAA CGACCTCTGC TGCGGGGGTC CCGGCACAAG GGAAGGTAGC ATCCCACGGG

121  CCGACACCAG GACGTCATGG TGGGCATGGG CCAGAAAGAC TCGTATGTGG GTGACAAGGC  180
     GGCTGTGGTC CTGCAGTACC ACCCGTACCC GGTCTTTCTG AGCATACACC CACTGTTCCG

181  CCAGAGCAAG AGGGGTATCC TGACCCTGAA GTACCCTATC GAACACGGCA TTGTCACCAA  240
     GGTCTCGTTC TCCCCATAGG ACTGGGACTT CATGGGATAG CTTGTGCCGT AACAGTGGTT
                          BglII
241  CTGGGATGAC ATGGAGAAGA TCTGGCACCA CACCTTCTAC AATGAGCTGC GTGTGACCCC  300
     GACCCTACTG TACCTCTTCT AGACCGTGGT GTGGAAGATG TTACTCGACG CACACTGGGG

301  TGAGGAGCAC CCGGTGCTTC TGACCGAGGC CCCCCTGAAC CCCAAAGCTA ACAGAGAGAA  360
     ACTCCTCGTG GGCCACGAAG ACTGGCTCCG GGGGGACTTG GGGTTTCGAT TGTCTCTCTT

361  GATGACGCAG ATAATGTTTG AAACCTTCAA TACCCCAGCC ATGTACGTGG CCATTCAGGC  420
     CTACTGCGTC TATTACAAAC TTTGGAAGTT ATGGGGTCGG TACATGCACC GGTAAGTCCG
                                                        HinfI
421  GGTGCTGTCC TTGTATGCAT CTGGGTGCAC CACTGGCATT GTCATGGACT CTGGTGACGG  480
     CCACGACAGG AACATACGTA GACCCACGTG GTGACCGTAA CAGTACCTGA GACCACTGCC

481  GGTCACACAC ACAGTGCCCA TCTATGAGGG CTACGCCCTT CCCCATGCCG TCTTGCGTCT  540
     CCAGTGTGTG TGTCACGGGT AGATACTCCC GATGCGGGAA GGGGTACGGC AGAACGCAGA
                                      XhoII
541  GGACCTGGCT GGTCGGGTCC TGACAGACTA CCTCATGAAG ATCCTGACTG AACGGGGCTA  600
     CCTGGACCGA CCAGCCCAGG ACTGTCTGAT GGAGTACTTC TAGGACTGAC TTGCCCCGAT

601  CAGCTTTACC ACCACTGCTA AGAGGGAAAT TGTTCGAGAC ATAAAGGAGA AGCTGTGCTA  660
     GTCGAAATGG TGGTGACGAT TCTCCCTTTA ACAAGCTCTG TATTTCCTCT TCGACACGAT
                                                  StyI
661  TGTTGCCCTG TATTTTGAGC AAGAAATGGC TACTGCTACA TCATCTTCCT CCTTGGAGAA  720
     ACAACGGGAC ATAAAACTCG TTCTTTACCG ATGACGATGT AGTAGAAGGA GGAACCTCTT

721  GAGTTACGAG CTGCCCGATG GCAGGTTAT CACCATCGGC AATGAGCGGT TCCGGTGTCC  780
     CTCAATGCTC GACGGGCTAC CCGTCCAATA GTGGTAGCCG TTACTCGCCA AGGCCACAGG

781  AGAGGCACTC TTCCAGACTT CCTTCTGAAA GAAAGTGAAA TTTCAAGACC TGTAAGTCAT  840
     TCTCCGTGAG AAGGTCTGAA GGAAGACTTT CTTTCACTTT AAAGTTCTGG ACATTCAGTA

841  ATAAAGTACT CAGAAATTGC TGGCTGTTTG TGAGCCTAGA GGCGCCTGGG GCGAGAAAAG  900
     TATTTCATGA GTCTTTAACG ACCGACAAAC ACTCGGATCT CCGCGGACCC CGCTCTTTTC

901  AGAAAAACAA ACCTGGGTAT GCCTCGTAGT TAAAACATTC CTGGGAACAT CTTGACCATA  960
     TCTTTTTGTT TGGACCCATA CGGAGCATCA ATTTTGTAAG GACCCTTGTA GAACTGGTAT
                                          HinfI
961  AGATAAAGGG GACTGTGAAG ACATAGCAGG GCTATCTGAA CTGAGTCAAC AACTCACAGA 1020
     TCTATTTCCC CTGACACTTC TGTATCGTCC CGATAGACTT GACTCAGTTG TTGAGTGTCT
```

continued overleaf . . . . .

continued. . . .

```
1021 ACTCTGACAC CCTGCACGTA CATGTAATTT TTCTGTTAAT GTTTGAATAA GCCAATAGTG 1080
     TGAGACTGTG GGACGTGCAT GTACATTAAA AAGACAATTA CAAACTTATT CGGTTATCAC
                  EcoRI
1081 TGTCGCTATG CTGAATTCCA CACCCCTAAG CCCCTTACCC CATAAAACCC CCTAACTTTC 1140
     ACAGCGATAC GACTTAAGGT GTGGGGATTC GGGGAATGGG GTATTTTGGG GGATTGAAAG
                                                              SstI
1141 GAGCCTCGTG GCCGGCCATC CGTTATCTCC TGTGTGGGAT ACATGTCGGT CTGGAGCTCC 1200
     CTCGGAGCAC CGGCCGGTAG GCAATAGAGG ACACACCCTA TGTACAGCCA GACCTCGAGG

1201 GTAATTAAAC GTCCTCATGT AATTACAGCA AGATGGGTCC TCGTGTTTCT TTGGGTGCTC 1260
     CATTAATTTG CAGGAGTACA TTAATGTCGT TCTACCCAGG AGCACAAAGA AACCCACGAG

1261 TCACACTCCT GAGACTAGAG TGGGGGTCCC CAAAAGGGGT CTTACACTTC CTGGGCATGG 1320
     AGTGTGAGGA CTCTGATCTC ACCCCCAGGG GTTTTCCCCA GAATGTGAAG GACCCGTACC

1321 AGTCCTGTGG TATCCACGAG ATCACCTTCA ACTCCATCAT GAAGTGTGAT GTGGATATCC 1380
     TCAGGACACC ATAGGTGCTC TAGTGGAAGT TGAGGTAGTA CTTCACACTA CACCTATAGG

1381 GCAAAGACCT GTATGCCAAT ACAGTGCTGT CTGGTGGTAC CCACCATGTA CCCAGGCATT 1440
     CGTTTCTGGA CATACGGTTA TGTCACGACA GACCACCATG GGTGGTACAT GGGTCCGTAA

1441 GCTGACAGGA TGAAGAAGGA GATCACAACC CTAGCACCCA GCACAACGAA GATTAAGATC 1500
     CGACTGTCCT ACTTCTTCCT CTAGTGTTGG GATCGTGGGT CGTGTTGCTT CTAATTCTAG

1501 ATTGCTCCCC CTGAGCGCAA GTACTCAGTC TGGTTCTGTG GCTCCATTCT GGCCTCACTG 1560
     TAACGAGGGG GACTCGCGTT CATGAGTCAG ACCAAGACAC CGAGGTAAGA CCGGAGTGAC

1561 TCCACCTTCC AGCAGATGTG GATCAGCAAG CAGGAGTATG ATGAGTTGGG CCCCTCTATC 1620
     AGGTGGAAGG TCGTCTACAC CTAGTCGTTC GTCCTCATAC TACTCAACCC GGGGAGATAG
                       XbaI
1621 ATCCACATCA AATGCTTCTA GATGGACCGT GGCAGGTGCC AAGCATCTGC TGCATGAGCC 1680
     TAGGTGTAGT TTACGAAGAT CTACCTGGCA CCGTCCACGG TTCGTAGACG ACGTACTCGG

1681 GATATTGAAG TATTGATTTG CCCTGGCAAA TGTACACACC TCATGCTAGC CTCATGAAAC 1740
     CTATAACTTC ATAACTAAAC GGGACCGTTT ACATGTGTGG AGTACGATCG GAGTACTTTG

1741 TGGAATAAGT CCCCCCCCCC TTTCCTTTTT ATTTTTTATT CACTTAACAT CCCAGACACA 1800
     ACCTTATTCA GGGGGGGGGG AAAGGAAAAA TAAAAAATAA GTGAATTGTA GGGTCTGTGT

1801 GCCCCCGCCC CTTTTCAGAG TTCCTCCTTT ACAAGTTCCT TCCACCATTC CCTTTTCTCT 1860
     CGGGGGCGGG GAAAAGTCTC AAGGAGGAAA TGTTCAAGGA AGGTGGTAAG GGAAAAGAGA

1861 TTGTCTCTTA GAAATGGGAC CCGTTTGTGT ACCACTTACC  1900
     AACAGAGAAT CTTTACCCTG GGCAAACACA TGGTGAATGG
```

Both strands of the nucleotide sequence of interrupted actin pseudogene and flanking region determined in the genomic clone λmA119 are shown. The BglII and XbaI sites, near the limits of the region sequenced are those of Figure 3.9. Other restriction sites are underlined above the appropriate nucleotides. The boxed sequences are direct repeats flanking the inserted sequence.

Both strands of the nucleotide sequence are shown, together with a number of restriction sites which were useful in the determination of the sequence. Comparison of this nucleotide sequence with other known sequences is discussed in subsequent sections.

3.4 Analysis of Actin-like Amino-acid Sequences

3.4.1 The Actin-like Sequence in Clone λmA118

The portion of the nucleotide sequence of clone λmA118 corresponding to the actin-like gene is shown in Figure 3.11. This was related to the coding sequence of an actin-like gene over an area from amino-acid 5 to a stop codon following amino-acid 374. Examination of the predicted amino-acid sequence shows that λmA118 most closely resembles a gene for a cytoplasmic isoform of actin (Vandekerckhove and Weber, 1979a). Of the 22 residues unique to cytoplasmic actins, all are found in the predicted sequence in clone λmA118 (represented by the underlined residues in Figure 3.11), except for two amino-acid residues at positions 16 and 17, and the latter position is part of a deletion.

There are four amino-acids at the N-terminal end of the sequence which differentiate the cytoplasmic actin ß and γ isoforms (Vandekerckhove and Weber, 1979a). These are amino-acid position 2 (ß = Asp, γ = Glu), position 3 (ß = Asp, γ = Glu), position 4 (ß = Asp, γ = Glu), and position 10 (ß = Val, γ = Ile). However, the actin-like sequence in clone λmA118 only starts from the Ile at codon position 5, therefore only the amino-acid at position 10 could be used to identify the isoform. As this corresponds to Ile (bold in Figure 3.11), the predicted amino-acid sequence resembles that of the γ, rather than the ß isoform.

The actin-like gene of λmA118 bears some of the hallmarks of a processed pseudogene. There are 33 differences in the predicted amino-acid sequence to that of γ-actin (represented by residues in bold italics in Figure 3.11), including stop codons rather than Arg and Gln, at positions 183 and 313 respectively. In addition, there are four deletions of nucleotides, these being at codon positions 16 to 24, 209 to 213, 345 to 347, and 365 to 368. Furthermore, the actin-like sequence is not interrupted by the introns anticipated for a mammalian actin. Although it is not yet known whether the coding gene for γ-actin has introns, the genes for the four mammalian actin isoforms so far

Figure 3.11 Nucleotide sequence and amino-acid translation of the γ-actin pseudogene in genomic clone λmA118

```
                                                                    5
                                                                   Ile Ala
1     CTGCAGGCTACACTGCGCTTCTTGCCGCTGCTCCATCGCCAATCAATCGCA ATA GCC     57
                    10                          16          24
      Ala Leu Val Ile Asp Asn Gly Ser Gly Thr          Asp Asn Ala
58    GCA CTA GTC ATT GAC AAT GGC TCC GGC ACG  TCAAT  GAC AAC GCC    101
                    30                                  40
      Leu Arg Ala Met Phe Pro Ser Ile Ile Gly Arg Pro Arg His Gln
102   CTC AGG GCC ATG TTC CCT TCC ATC ATA GGG CGC CCC CGA CAC CAG    146
                                        50
      Gly Val Leu Val Gly Ile Gly Gln Lys Asp Ser Tyr Val Gly Asp
147   GGT GTC TTG GTG GGC ATT GGC CAG AAG GAC TCC TAC GTG GGT GAT    191
                    60                                  70
      Glu Ala Gln Ser Lys Arg Gly Ile Leu Ala Leu Lys Tyr Pro Val
192   GAG GCC CAG AGC AAG AGG GGT ATC CTG GCC CTG AAG TAC CCT GTC    236
                                        80
      Glu His Gly Ile Val Thr Asn Trp Asp Asp Met Glu Lys Ile Trp
237   GAG CAT GGC ATT GTC ACC AAC TGG GAC GAC ATG GAG AAG ATC TGG    281
                    90                                  100
      His His Thr Phe Tyr Asn Glu Leu Arg Val Ala Pro Glu Glu His
282   CAC CAC ACC TTC TAC AAT GAG CTG CGT GTG GCC CCT GAG GAG CAC    326
                                        110
      Pro Val Leu Leu Thr Glu Ala Pro Leu Asn Pro Lys Ala Asn Arg
327   CCG GTG CTA CTG ACC GAG GCC CCC CTG AAC CCC AAA GCT AAC AGA    371
                    120                                 130
      Glu Lys Met Thr Gln Ile Met Phe Glu Pro Phe Asn Thr Pro Ala
372   GAG AAG ATG ACG CAG ATA ATG TTT GAA CCC TTC AAT ACC CCA GCC    416
                                        140
      Leu Tyr Val Thr Ile Gln Val Val Leu Ser Leu Tyr Val Ser Gly
417   TTG TAC GTC ACC ATT CAG GTG GTG CTC TCC TTG TAT GTA TCT GGG   1332
                                  442 ⇑ 1313
                    150                                 160
      Arg Thr Thr Gly Ile Val Met Asp Ser Gly Asp Gly Val Thr His
1333  CGC ACC ACT GGC ATT GTC ATG GAC TCT GGT GAC GGG GTC ACA CAC   1377
                                        170
      Thr Val Ala Ile Tyr Asp Ser Tyr Thr Leu Pro His Ala Ile Leu
1378  ACA GTG GCC ATC TAT GAC AGC TAC ACC CTT CCT CAC GCC ATC TTG   1422
                    180                                 190
      Cys Leu Asp Leu Val Gly End Asp Leu Thr Glu Tyr Leu Met Asn
1423  TGT CTG GAC TTG GTT GGC TAG GAC CTG ACA GAG TAC CTC ATG AAT   1467
                                        200
      Ser Leu Thr Glu Arg Gly Tyr Ser Phe Thr Thr Thr Ala Glu Arg
1468  TCC TTG ACT GAA CGG GGC TAC AGC TTT ACC ACC ACT GCT GAG AGG   1512
```

continued overleaf . . . . .

continued....

```
                209      213                                220
      Glu Ile Val        Lys Glu Lys Leu Cys Tyr Val Ala Leu Asp Phe
1513  GAA ATT GTG  AC AAG GAG AAG CTG TGC TAT GTT GCC CTG GAT TTT  1556
                                230                234a
      Glu Gln Glu Lys Ala Thr Ala Ala Ser Ser Ser Ser Leu Glu Lys
1557  GAG CAA GAA AAG GCT ACT GCT GCA TCA TCT TCC TCC TTG GAG AAG  1601
                240                                      250
      Ser Tyr Gln Leu Pro Asp Gly Gln Val Ile Thr Ile Gly Asn Glu
1602  AGT TAC CAG CTG CCC GAT GGG CAG GTG ATC ACC ATT GGC AAT GAG  1646
                                260
      Arg Phe Arg Cys Pro Glu Ala Leu Phe Gln His Ser Phe Leu Gly
1647  CGG TTC CGG TGT CCG GAG GCA CTC TTC CAG CAT TCC TTC CTG GGC  1691
                270                                      280
      Met Glu Ser Cys Gly Ile Tyr Glu Thr Thr Phe Asn Ser Ile Met
1692  ATG GAA TCC TGT GGC ATC TAC GAG ACC ACC TTC AAC TCC ATC ATG  1736
                                290
      Lys Cys Asp Val Asp Ile Cys Lys Asp Leu Tyr Ala Asn Thr Val
1737  AAG TGT GAT GTG GAT ATC TGC AAA GAC CTG TAT GCC AAT ACA GTG  1781
                300                                      310
      Leu Ser Gly Gly Thr Thr Met Tyr Pro Gly Ile Ala Asp Arg Met
1782  CTG TCC GGT GGT ACC ACC ATG TAC CCA GGC ATT GCT GAC AGG ATG  1826
                                320
      End Lys Glu Ile Thr Ala Leu Ala Pro Ser Thr Met Lys Ile Lys
1827  TAG AAG GAG ATC ACA GCC CTA GCA CCC AGC ACA ATG AAG ATT AAG  1871
                330                                      340
      Ile Ile Ala Pro Pro Glu Arg Lys Tyr Ser Val Trp Thr Cys Gly
1872  ATC ATT GCT CCC CCT GAG CGC AAG TAC TCA GTC TGG ACC TGC GGC  1916
                345      347       350
      Ser Ile Leu        Ser Leu Ser Thr Phe Gln Gln Met Trp Ile Ser
1917  TCC ATC CTA CC  TCA CTG TCC ACC TTC CAG CAG ATG TGG ATC AGC  1960
                360                365      368      370
      Lys Gln Glu Tyr Asp Glu Ser Gly      Ile Val His Arg Lys Cys
1961  AAG CAG GAG TAT GAT GAG TCG GGC  CC ATC GTC CAC CGC AAA TGC  2004
      374
      Phe End
2005  TTC TAG A  2011
```

The figure shows the complete γ-like actin coding sequence from clone λmA118. Numbering of amino-acids is as in Vandekerckhove and Weber (1979a).

Underline = residues specific for cytoplasmic actins.

Bold = residues specific for cytoplasmic γ-actin.

Bold italics = difference from amino-acid sequence of mouse γ-actin.

The arrow between nucleotides 442 and 1313 indicates the position of the inserted sequence. The boxed sequences are direct repeats flanking the inserted sequence.

characterised all have introns at amino-acid positions 41, 267 and 327 (as well as at other positions specific for different isoforms).

Although it is possible, in principle, that the inserted sequence in λmA118 might be an intron (discussed in section 5.2.3i), the lack of any other introns, including those that have been conserved in mammalian actins, is consistent with λmA118 being a pseudogene.

Most processed pseudogenes contain DNA copies of the whole of the mRNA, including the 5' untranslated region. However, λmA118 only contains actin-like sequence from amino-acid 5 (Figure 3.12). Another γ-actin-like pseudogene, λmA19, is truncated at position 7 (Leader *et al.*, 1985). This was shown by a preceding sequence (target-site direct repeat) that was repeated after the poly A tail. In the case of λmA118 the 3' untranslated region has not yet been sequenced, so that it is unclear whether the pseudogene arose from a truncated transcript, or from a full-length transcript, the 5' portion of which was subsequently deleted.

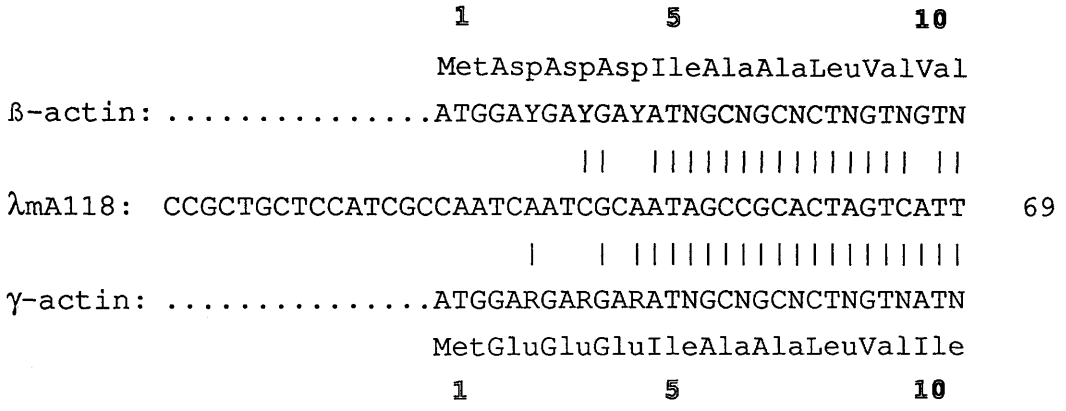
3.4.2 The Actin-like Sequence in Clone λmA119

The portion of the nucleotide sequence of clone λmA119 corresponding to the actin-like gene is shown in Figure 3.13. This resembled the coding sequence of an actin-like gene from an initiating Met codon to a termination codon after amino-acid position 375. Examination of the predicted amino-acid sequence shows that λmA119 also most closely resembles the gene for a cytoplasmic isoform of actin (Vandekerckhove and Weber, 1979a). Of the 22 residues unique to cytoplasmic actins, all of them corresponded to those of the predicted sequence in clone λmA119 (represented by the underlined residues in Figure 3.13), except for Leu at amino-acid position 364.

Of the four amino-acids at the N-terminal end of the sequence which distinguish the ß and γ isoforms of cytoplasmic actin (Vandekerckhove and Weber, 1979a), all of these (Glu^2, Glu^3, Glu^4, and Ile^{10}) in the predicted amino-acid sequence of clone λmA119 correspond to those of the γ-isoform (indicated by residues in bold in Figure 3.13). Thus the sequence of the actin-like gene in λmA119 resembles that of the γ, rather than the ß isoform.

The actin-like gene of λmA119 also has characteristics of a processed pseudogene. It differs in predicted amino-acid sequence from the γ-actin at 21 positions (indicated by residues in bold italics in Figure 3.13). The predicted amino-acid sequence does not have the potential to encode a full actin-like

Figure 3.12 Comparison of nucleotides encoding the N-terminal sequences of actins with corresponding region in clone λmA118



The nucleotide sequences that could encode the known amino-acid sequences of the N-terminal positions of mouse ß- and γ-cytoplasmic actins are aligned to the corresponding region of the processed actin pseudogene in λmA118, numbered as in Figure 3.8. R = purine, Y = pyrimidine and N = unspecified nucleotide. Vertical lines between nucleotides indicate identity.

Figure 3.13 Nucleotide sequence and amino-acid translation of the γ-actin pseudogene in genomic clone λmA119

```
                  1                                                  10
             Met Glu Glu Glu Ile Ala Thr Leu Val Ile Val Asn Gly
1    CGTCGCA ATG GAA GAA GAA ATC GCC ACA CTC GTC ATT GTC AAT GGC    46
                                 20
     Ser Gly Met Cys Lys Ala Gly Phe Ala Gly Asp Asp Ala Pro Arg
47   TCC GGC ATG TGC AAA GCT GGC TTT GCT GGA GAC GAC GCC CCC AGG    91
        30                                                  40
     Ala Val Phe Pro Ser Ile Val Gly Cys Pro Arg His Gln Asp Val
92   GCC GTG TTC CCT TCC ATC GTA GGG TGC CCC CGA CAC CAG GAC GTC   136
                                 50
     Met Val Gly Met Gly Gln Lys Asp Ser Tyr Val Gly Asp Lys Ala
137  ATG GTG GGC ATG GGC CAG AAA GAC TCG TAT GTG GGT GAC AAG GCC   181
        60                                                  70
     Gln Ser Lys Arg Gly Ile Leu Thr Leu Lys Tyr Pro Ile Glu His
182  CAG AGC AAG AGG GGT ATC CTG ACC CTG AAG TAC CCT ATC GAA CAC   226
                                 80
     Gly Ile Val Thr Asn Trp Asp Asp Met Glu Lys Ile Trp His His
227  GGC ATT GTC ACC AAC TGG GAT GAC ATG GAG AAG ATC TGG CAC CAC   271
        90                                                 100
     Thr Phe Tyr Asn Glu Leu Arg Val Thr Pro Glu Glu His Pro Val
272  ACC TTC TAC AAT GAG CTG CGT GTG ACC CCT GAG GAG CAC CCG GTG   316
                                110
     Leu Leu Thr Glu Ala Pro Leu Asn Pro Lys Ala Asn Arg Glu Lys
317  CTT CTG ACC GAG GCC CCC CTG AAC CCC AAA GCT AAC AGA GAG AAG   361
        120                                                130
     Met Thr Gln Ile Met Phe Glu Thr Phe Asn Thr Pro Ala Met Tyr
362  ATG ACG CAG ATA ATG TTT GAA ACC TTC AAT ACC CCA GCC ATG TAC   406
                                140
     Val Ala Ile Gln Ala Val Leu Ser Leu Tyr Ala Ser Gly Cys Thr
407  GTG GCC ATT CAG GCG GTG CTG TCC TTG TAT GCA TCT GGG TGC ACC   451
        150                                                160
     Thr Gly Ile Val Met Asp Ser Gly Asp Gly Val Thr His Thr Val
452  ACT GGC ATT GTC ATG GAC TCT GGT GAC GGG GTC ACA CAC ACA GTG   496
                                170
     Pro Ile Tyr Glu Gly Tyr Ala Leu Pro His Ala Val Leu Arg Leu
497  CCC ATC TAT GAG GGC TAC GCC CTT CCC CAT GCC GTC TTG CGT CTG   541
        180                                                190
     Asp Leu Ala Gly Arg Val Leu Thr Asp Tyr Leu Met Lys Ile Leu
542  GAC CTG GCT GGT CGG GTC CTG ACA GAC TAC CTC ATG AAG ATC CTG   586
                                200
     Thr Glu Arg Gly Tyr Ser Phe Thr Thr Thr Ala Lys Arg Glu Ile
587  ACT GAA CGG GGC TAC AGC TTT ACC ACC ACT GCT AAG AGG GAA ATT   631
        210                                                220
     Val Arg Asp Ile Lys Glu Lys Leu Cys Tyr Val Ala Leu Tyr Phe
632  GTT CGA GAC ATA AAG GAG AAG CTG TGC TAT GTT GCC CTG TAT TTT   676
```

continued overleaf .....

continued....

```
                                        230                     234a
      Glu Gln Glu Met Ala Thr Ala Thr Ser Ser Ser Ser Leu Glu Lys
677   GAG CAA GAA ATG GCT ACT GCT ACA TCA TCT TCC TCC TTG GAG AAG  721
               240                                      250
      Ser Tyr Glu Leu Pro Asp Gly Gln Val Ile Thr Ile Gly Asn Glu
722   AGT TAC GAG CTG CCC GAT GGG CAG GTT ATC ACC ATC GGC AAT GAG  766
                                    260                 265
      Arg Phe Arg Cys Pro Glu Ala Leu Phe Gln Thr Ser Phe Leu Gly
767   CGG TTC CGG TGT CCA GAG GCA CTC TTC CAG ACT TCC TTC CTG GGC 1316
                                                  805 ⇑ 1310
               270                                      280
      Met Glu Ser Cys Gly Ile His Glu Ile Thr Phe Asn Ser Ile Met
1317  ATG GAG TCC TGT GGT ATC CAC GAG ATC ACC TTC AAC TCC ATC ATG 1361
                                    290
      Lys Cys Asp Val Asp Ile Arg Lys Asp Leu Tyr Ala Asn Thr Val
1362  AAG TGT GAT GTG GAT ATC CGC AAA GAC CTG TAT GCC AAT ACA GTG 1406
            300       302       303                       310
      Leu Ser Gly Gly Thr       Thr Met Tyr Pro Gly Ile Ala Asp Arg
1407  CTG TCT GGT GGT ACC  C    ACC ATG TAC CCA GGC ATT GCT GAC AGG 1449
                                        320
      Met Lys Lys Glu Ile Thr Thr Leu Ala Pro Ser Thr Thr Lys Ile
1450  ATG AAG AAG GAG ATC ACA ACC CTA GCA CCC AGC ACA ACG AAG ATT 1494
                  330                                   340
      Lys Ile Ile Ala Pro Pro Glu Arg Lys Tyr Ser Val Trp Phe Cys
1495  AAG ATC ATT GCT CCC CCT GAG CGC AAG TAC TCA GTC TGG TTC TGT 1539
                                        350
      Gly Ser Ile Leu Ala Ser Leu Ser Thr Phe Gln Gln Met Trp Ile
1540  GGC TCC ATT CTG GCC TCA CTG TCC ACC TTC CAG CAG ATG TGG ATC 1584
                  360                                   370
      Ser Lys Gln Glu Tyr Asp Glu Leu Gly Pro Ser Ile Ile His Ile
1585  AGC AAG CAG GAG TAT GAT GAG TTG GGC CCC TCT ATC ATC CAC ATC 1629
            374
      Lys Cys Phe End
1630  AAA TGC TTC TAG ATGGACCGTGGCAGGTGCCAAGCATCTGCTGCATGAGCCGATA 1684

1685  TTGAAGTATTGATTTGCCCTGGCAAATGTACACACCTCATGCTAGCCTCATGAAACTGG 1743

1744  AATAAGTCCCCCCCCCCTTTCCTTTTTATTTTTTATTCACTTAACATCCCAGACACAGC 1802

1803  CCCCGCCCCTTTTCAGAGTTCCTCCTTTACAAGTTCCTTCCACCATTCCCTTTTCTCTT 1861

1862  TGTCTCTTAGAAATGGGACCCGTTTGTGTACCACTTACC 1900
```

The figure shows the complete γ-like actin coding sequence from clone λmA119. Numbering of amino-acids is as in Vandekerckhove and Weber (1979a).

Underline = residues specific for cytoplasmic actins.

Bold = residues specific for cytoplasmic γ-actin.

Bold italics = difference from amino-acid sequence of mouse γ-actin.

The arrow between nucleotides 805 and 1310 indicates the position of the inserted sequence. The boxed sequences are direct repeats flanking the inserted sequence.

protein as it is interrupted by an out-of-phase nucleotide between amino-acid positions 302 and 303. Furthermore, the actin-like sequence is not interrupted by the introns that are conserved in known mammalian actin isoforms (discussed in section 3.4.1, above), although there is a sequence interrupting it, the nature of which is discussed in Chapter 5.

The 3' untranslated region of λmA119 also resembles that of mouse γ-actin, both the previously published γ-actin pseudogene λmA19 (Leader *et al.*, 1985), and the as yet unpublished partial mouse γ-actin cDNA sequence (Peter and Leader, personal communication). However, unlike λmA19, this relatedness only extends to nucleotide 1751, afterwhich the sequences diverge (Figure 3.14).

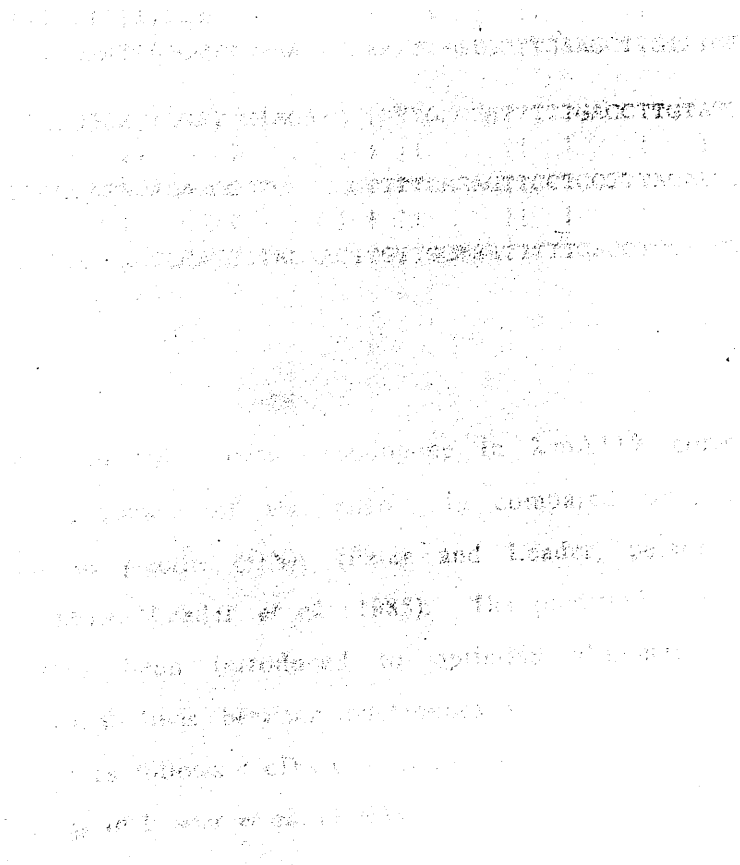


Figure 3.14 Comparison of γ-actin 3' untranslated sequences with clone λmA119

```
                End
cDNA:           TAGATGGACTGA-GCAGGTGCCAGGCATCTGCTGCATGAGCTGATATTGA   1035
                |||||||||| | ||||||||| |||||||| ||||||||||||||||
λmA119:         TAGATGGACCGTGGCAGGTGCCAAGCATCTGCCGCATGAGCTGATATTGA   1688
                |||||||||    ||||||||| |||||||| ||||||||||||||||
λmA19:          TAGATGGAC-TGAGCAGGTGCCAGGCATCTGCTGCATGAGCTGATATTGA   1400

cDNA:           AGTATCAATTTGCCCTGGCAAATGTACACACCTCATGCTAGCCTCATGAA   1085
                |||||  ||||||||||||||||||||||||||||||||||||||||||
λmA119:         AGTATTGATTTGCCCTGGCAAATGTACACACCTCATGCTAGCCTCATGAA   1738
                |||||  ||||||||||||||||||| |||||||||||||||||||||
λmA19:          AGTATCGATTTGCCCTGGCAAATGTATACACCTCATGCTAGCCTCATGAA   1450

cDNA            ACTGGAATAAGCCTTTGAAAAGAAATTTAGTCCTTGAAGCTTGTATCTGA   1135
                |||||||||||  |              |  |  ||     ||   |
λmA119:         ACTGGAATAAGTCCCCCCCCCCTTTCCTTTTTATTTTTTATTCACTTAAC   1788
                |||||||||||||              |  |  ||     ||   |
λmA19:          ACTGGAATAAGCCTTTGAAAAGAAATTT-GTCCTTGAAGCTTGTATCTGA   1499

cDNA:           TATCAGCACTGGATCGTAGAACTTGTTGCTGATTTTTGACCTTGTATTCA   1185
                   |  ||  ·  |  |  | | ||     ||  |    |   | |
λmA119:         ATCCCAGACACAGCCCCCGCCCCTTTTCAGAGTTCCTCCTTTACAAGTTC   1838
                   |  ||     |  |  | | ||     ||  |    |   | |
λmA19:          TATCAGCACTGGATCGTAGAACTTGTTGCTGATTTTTGACCTTGTATTCA   1549
```

The region of the γ-actin pseudogene in λmA119 corresponding to the 3'untranslated sequence of the mRNA is compared with the corresponding regions in mouse γ-actin cDNA (Peter and Leader, personal communication), and clone λmA19 (Leader *et al.*, 1985). The predicted stop codons are boxed. Gaps that have been introduced to optimise alignment are indicated by hyphens. Vertical lines between nucleotides indicate identity. The numbering of sequences are as follows : cDNA is as in Figure 5.3; λmA119 is as in Figure 3.10; λmA19 is as in Leader *et al.*, (1985).

CHAPTER 4

ANALYSIS OF INSERTED SEQUENCES IN ACTIN-LIKE GENES

The electron microscopic heteroduplexes formed between the actin regions of λmA36, λmA118, and λmA119 and the reference clones (Figure 1.4), demonstrated that the DNA of each of these clones was interrupted by a single-stranded loop. The genomic regions corresponding to these loops were subjected to structural analysis in order to characterise them further.

4.1 Analysis of the Inserted Sequence in Clone λmA118

4.1.1 Nucleotide Sequence of the Inserted Sequence in Clone λmA118

Figure 3.5 shows a detailed partial restriction map of clone λmA118 in the vicinity of the actin-like gene. Further subclones containing the inserted sequence in this gene were derived using an internal PstI site in the original parent subclone 118Y1-1, and were designated 118P2 and 118P3 (section 3.2.1). In order to further investigate the nature of the inserted sequence, the nucleotide sequence of the regions containing the inserted sequence in subclones 118P2 and 118P3 was determined.

The inserted sequence in clone λmA118 was sequenced as follows. Sequencing from the PstI, EcoRI, BamHI, and AccI sites (Figure 3.7 ; sequences number 5 to 10, and 24 to 29) allowed the determination of a total of 600 nucleotides of the inserted sequence. In order to identify further restriction sites for sequencing, a 0.6 kb BglII-PstI fragment was isolated from subclone 118P3 (see section 2.7.3), and subjected to extensive restriction analysis. The eventual identification of the restriction sites DraII and HinfI enabled further sequencing from these sites (sequences number 3, 4, and 30), and allowed the completion of the determination of this inserted sequence.

The complete nucleotide sequence of the inserted sequence in clone λmA118 is shown in Figure 4.1. The start of the inserted sequence begins at nucleotide number 443 and ends at nucleotide number 1307, with a total length of 865 nucleotides.

The inserted sequence starts after the second base of the codon for Leu^{140} of the γ-actin-like gene of λmA118. The actin-like gene appears to resume at the third nucleotide of the codon for Ala^{138}; this particular nucleotide, plus those of Val^{139} and the first two of Leu^{140} being repetitions of nucleotides preceding the start of the inserted sequence. The inserted sequence is therefore flanked by a short direct repeat of 6 base pairs of actin sequence, indicating that it was inserted at a staggered break. The inserted sequence in clone λmA118 was therefore designated IE 118 (inserted element 118).

4.1.2 Computer Analysis of IE 118

The nucleotide sequence of IE 118 was subjected to analysis on the VAX cluster at EMBL, Heidelberg, and compared with sequences in the GenBank and EMBL nucleotide sequence databases using the programs, WORDSEARCH together with SEGMENTS (see section 2.17.3), of the UWGCG sequence analysis software package (Devereux *et al.*, 1984).

The searches of the GenBank (release 40, consisting a total of 6379 sequences) and EMBL (release 9, consisting a total of 6396 sequences) nucleotide sequence data banks revealed no sequence with extensive homology to IE 118. An example of one of the 20 best matches is shown in Figure 4.2. The alignment with the sequence ecotgy1 (*E.coli* Tyr-tRNA-1 sequence from GenBank) has a low 'quality' (17.2) and the ratio (0.273) is not outstanding. This is not surprising in view of its bacterial nature. The highest quality recorded was from the sequence alignment with ptglb1.mbl, which gave a value of 72.3, but had an extremely poor ratio of 0.096 (this was a chimpanzee beta-globin sequence from EMBL). Other sequences contributing to the best diagonals included sequence humhbb (human beta-globin sequence from GenBank), with quality of 49.0 and ratio of 0.090 ; sequence yscg3pdc. (yeast glyceraldehyde-3-phosphate dehydrogenase sequence from GenBank), with quality of 41.0 and ratio of 0.090 ; and sequence musafp (mouse alpha-foetoprotein sequence from GenBank), with quality of 38.7 and ratio of 0.050. The relatively poor quality and ratio values for these matches together with visual inspection suggested that in no case did one of the 20 best matches represent biologically significant

Figure 4.1 Nucleotide sequence of IE 118

```
                    138139140
                      aValLe                                    DraII
437  ..............GGTGCTTGGTCACGGACTGGGGCTCTGTGGGCCCTCTTCGGTCT  480
     ..............CCACGAACCAGTGCCTGACCCCGAGACACCCGGGAGAAGCCAGA
        HinfI HinfI
481  GCGGAATCAGAGTCTCAGACAGATGGGCATAGAGTGGGCGAGTGACAAACAGACGTGACA  540
     CGCCTTAGTCTCAGAGTCTGTCTACCCGTATCTCACCCGCTCACTGTTTGTCTGCACTGT
                  HinfI
541  AGAGAACGTGTTGAATCTGAGTGTAATTTATCAAATCCAGCATCAAACTTTTTATACAGA  600
     TCTCTTGCACAACTTAGACTCACATTAAATAGTTTAGGTCGTAGTTTGAAAAATATGTCT

601  ATAACAAGAAACCAGGCGAACACATCCGCTAAGTTACAGTGACACAAAACAAAAGGAATG  660
     TATTGTTCTTTGGTCCGCTTGTGTAGGCGATTCAATGTCACTGTGTTTTGTTTTCCTTAC

661  CATACATCAAAAGATGGCGGGGACCAAGCTCATTACCACTAGAAGGAACAGGTGTAATGC  720
     GTATGTAGTTTTCTACCGCCCCTGGTTCGAGTAATGGTGATCTTCCTTGTCCACATTACG

721  TAGTCTATTGTTAAACCCACCACCAAGGGGTTCTTAGTAAATGCCTGATTATGCTGTTCC  780
     ATCAGATAACAATTTGGGTGGTGGTTCCCCAAGAATCATTTACGGACTAATACGACAAGG
                                HinfI
781  TTTGGGCCTAGTGAAGAAACCTGTCCAAGGGGGATTCCCTAACTCTTTCATGGTTACCCC  840
     AAACCCGGATCACTTCTTTGGACAGGTTCCCCCTAAGGGATTGAGAAAGTACCAATGGGG
                                                          PstI
841  ACCTATTTGCTAGGCCATTGTGTCCTAAGGCTACTGTCCTAAATAATCACTCTGCAGACT  900
     TGGATAAACGATCCGGTAACACAGGATTCCGATGACAGGATTTATTAGTGAGACGTCTGA

901  AGCCCTGAGCTATTCTAGCTCCGTTCGGAGCACTGGGTGCTCCTCAGGGGCCACACACAC  960
     TCGGGACTCGATAAGATCGAGGCAAGCCTCGTGACCCACGAGGAGTCCCCGGTGTGTGTG
                                                          EcoRI
961  GCTTCTCTACTAGAAGTAAATTTGAATGTTACTGAATAGGTAACCTTCTCACTGAATTCC 1020
     CGAAGAGATGATCTTCATTTAAACTTACAATGACTTATCCATTGGAAGAGTGACTTAAGG
                                TaqI
1021 CACTAAATTCCAAGCTCCTCGGCGTCGAGGATTTTCTAGGACATTGCAACACTGGCGAAG 1080
     GTGATTTAAGGTTCGAGGAGCCGCAGCTCCTAAAAGATCCTGTAACGTTGTGACCGCTTC

1081 GCTTAGCTATGTCAAGCAATCAAATCTTAAAGGCACTTATAATAAAACAATACTGAAAGA 1080
     CGAATCGATACAGTTCGTTAGTTTAGAATTTCCGTGAATATTATTTTGTTATGACTTTCT
            BamHI                                        AccI
1141 GAGCACGTGGATCCATACACCAAACTAACACGGGAAAAGGGTTTGAGTATACGGGCTATG 1200
     CTCGTGCACCTAGGTATGTGGTTTGATTGTGCCCTTTTCCCAAACTCATATGCCCGATAC

1201 GGAATGCCAAGGTTCCAGGAGGCATAGTTTCCTTGAAACTCATTGCCTCGTGAGTGTTTC 1260
     CCTTACGGTTCCAAGGTCCTCCGTATCAAAGGAACTTTGAGTAACGGAGCACTCACAAAG

1261 CAGGCCTCTTGGCCAGTCAAGCAGACTTCACCGGAGTGGGCGTAGGAGGTGCT..... 1313
     GTCCGGAGAACCGGTCAGTTCGTCTGAAGTGGCCTCACCCGCATCCTCCACGA.....
                                          aValLeu
                                        138139140
```

The nucleotide sequence of both strands of the inserted sequence in genomic clone λmA118 is shown. Restriction sites used in sequencing are indicated, as are the amino-acid equivalents of the flanking nucleotides of the actin processed pseudogene. The boxed sequences are direct repeats flanking the inserted sequence. The numbering of the sequence is as in Figure 3.8.

Figure 4.2 Example of output of WORDSEARCH/SEGMENTS on IE 118

```
SEGMENTS from : INS118.SEC  8-JUL-86  22:43
WORDSEARCH of : disk$users:[lehrach.david]ins118.rft;1 check: 5608
from: 1   to:865
ASSEMBLE  14-JUN-85  12:30
Symbols:   1 to : 865 from: ma118.rft   ck: 3255.   811 to: 1675
TO:SEARCHDATA:GENBANK.SST  Files:1  Sequences:6379  Total-length:5516947
Word-size:7  Words:1333409  Diagonals:1135702  Total-diagonals:22056806
Integral-width: 3  Alphabet: 4  List-size: 20    8-JUL-86  22:43

AvMatch: 1.00  AvMisMatch: -0.60   GapWeight: 3.50  LengthWeight: 0.10

     ins118.rft          check: 5608  from: 222  to: 865
     genbank.sst  entry: 318  check: 318  from: 1  to: 1949
ENTRY: 318 SEARCHSET of : ecotgy1.  check:318  from: 1  to:1949
disk$users:[pubdata.genbank_40.uwgcg.][bacterial]ecotgy1.
Gaps: 1  Quality: 17.2  Ratio: 0.273  Words: 9  Width: 3  Limits: +/-4

                .         .         .         .         .
224    ATCAAAAGATGGCGGGGACCA.AGCTCATTACCACTAGAAGGAACAGGTG    272
       |||||||||||||||||   |||  | |  ||| ||     |  |   ||||
  2    atcaaaagatggcggatgccattgatgcttatcaacctgactacgtggtg     51
                .
273    TAATGCTAGTCTAT    286
              ||| |||
 52    ctggcgaagtatat     65
```

The example shown is part of the output of comparison program, WORDSEARCH/SEGMENTS (see section 2.17.3), to compare the sequence of IE 118 (filename: ins118.rft) to the GenBank (release 40) nucleotide sequence database (filename: GENBANK.SST). The GenBank contains a total number of 6379 sequences, totalling 5516947 nucleotides in length. The word-size chosen is 7, in a total number of 1333409 words. There are 1135702 diagonals which represent segments of similarity, in a total number of 22056806 diagonals. Each match scores a value of 1.00, while each mismatch scores -0.60. The gap weight is 3.50, and the length weight is 0.10. The output is for one of the 20 best diagonals, that of IE 118 compared with the sequence ecotgy1 (this represented an *E.coli* Tyr-tRNA-1 sequence from GenBank), a quality of 17.2 and a ratio of 0.273 was scored. This is derived from the definitions given in Section 2.17.3 as follows :

$$\text{Quality} = 37 - (0.60 \times 27) - (3.50 \times 1) - (0.10 \times 1) = 17.2$$

$$\text{Ratio} = 17.2/64 = 0.273$$

homology. Nor did the identity of any of the sequences give grounds for thinking otherwise.

4.1.3 Genomic Southern Blotting of IE 118

In order to investigate the occurrence of sequences related to IE 118 in the mouse genome, Southern blotting of digests of mouse genomic DNA was performed.

Mouse liver DNA from the inbred strain Balb/C was isolated (see section 2.15.1), digested with the restriction enzymes EcoRI, HindIII, and BamHI, subjected to gel electrophoresis in 1% agarose, and transferred to nitrocellulose filters (see section 2.9.1). Two ^{32}P-labelled DNA probes from different parts of IE 118 (Figure 3.7) were prepared. These were from a 0.43 kb DraII-PstI fragment (designated 118**a**), and a 0.25 kb PstI-BamHI (designated 118**b**) respectively (see section 2.9.2), and hybridised to the restricted genomic DNA attached to the filter (see section 2.9.3).

Figure 4.3 shows the genomic Southern blots of mouse DNA against the two probes from IE 118. It can be seen that a large number of hybridising bands were obtained, indicating that IE 118 is repeated in the mouse genome. Both probes 118**a** and 118**b** gave a similar hybridisation pattern with the digested genomic DNA, indicating that they occur together in these multiple sequences.

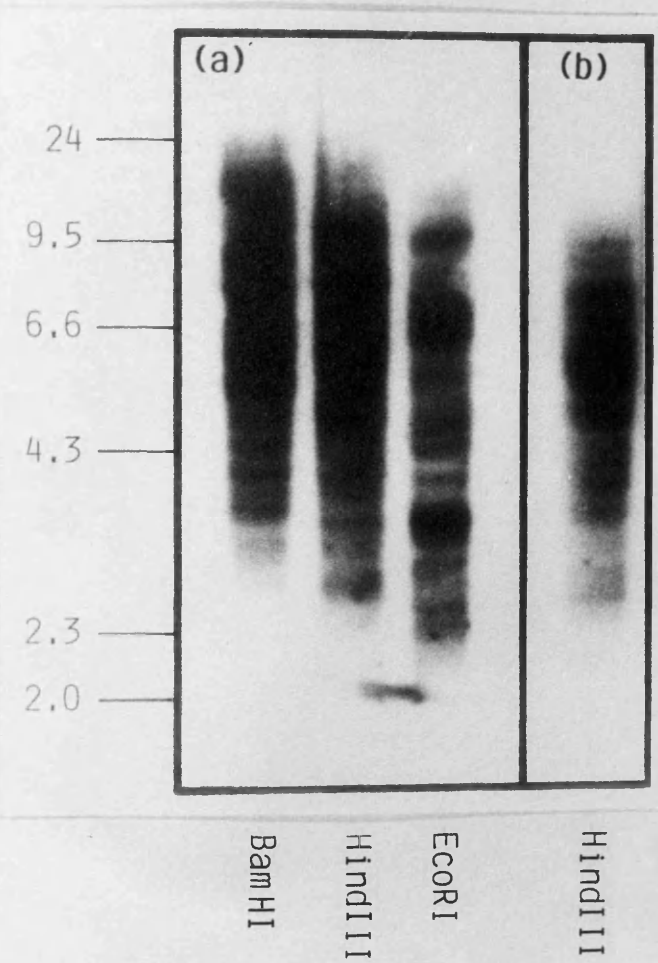
4.1.4 Estimation of Copy Number of IE 118 by Plaque Hybridisation

In order to obtain an estimate of the copy number of IE 118 in the mouse genome, a bacteriophage lambda mouse genomic library was screened with the IE 118 probes.

A DBA/2J mouse genomic lambda library (kindly provided by Dr A. M. Frishauf) was used to infect the bacterial host *E.coli* Y1090, producing approximately 1,000 plaques per plate, and these were transferred to nitrocellulose filters (see section 2.16.1). Two ^{32}P-labelled probes, 118**a** and 118**b**, were prepared from IE 118 as in section 4.1.3, and hybridised to the plaques attached to the filters (see section 2.16.2).

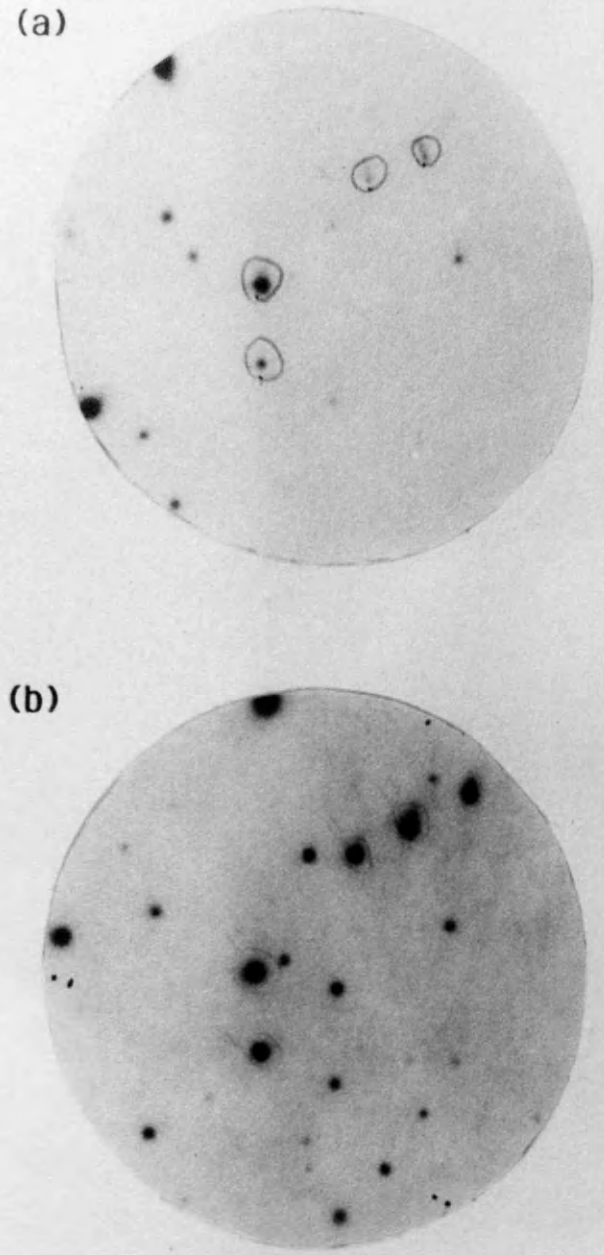
Figure 4.4 shows the plaques of the mouse genomic lambda library which hybridised to one or both probes from IE 118. It can be seen that 118**b**

Figure 4.3 Genomic Southern blot of mouse DNA hybridised to probes from IE 118



Mouse liver DNA (Balb/C strain) was digested with restriction enzymes EcoRI, HindIII, and BamHI as indicated, subjected to electrophoresis, transferred to nitrocellulose filters, and hybridised to ^{32}P-labelled fragments of (a) DraII-PstI, designated 118a, and (b) PstI-BamHI, designated 118b from IE 118.

Figure 4.4 Hybridisation of probes from IE 118 to plaques of a recombinant lambda mouse genomic library



A DBA/2J mouse genomic library in bacteriophage lambda was screened with ^{32}P-labelled fragments of (a) DraII-PstI, designated 118a, and (b) PstI-BamHI, designated 118b, from IE 118. Autoradiography of the same plate (number 3) hybridised with the two probes are presented (see Table 4.1).

hybridised to the majority of plaques to which 118a hybridised, although there were significantly more plaques to which only 118b hybridised than those to which only 118a hybridised. The possible implications of this result are discussed in section 5.2.3ii. The frequencies of positive plaque hybridisation of the two probes 118a and 118b to the lambda library is shown in Table 4.1. Probes 118a and 118b hybridised with a total of 42 and 78 plaques, respectively, out of a total number of 8056 plaques.

The copy number of IE 118 sequences in the mouse genome is calculated as follows :

Let n = sum of the hybridising plaques, and t = total number of plaques (= 8056). Then there are n copies of IE in t recombinant plaques containing mouse genomic DNA. The average size of the λ insert in the mouse genomic library is assumed to be 15kb (± 5kb), so there are n copies of IE in t X 15kb of the mouse genome. Assuming the size of the mouse haploid genome to be approximately 3 X 10^6kb,

then there are :

$$n \times \frac{3 \times 10^6 \text{ kb}}{t \times 15 \text{ kb}} \quad \text{copies of IE per mouse genome}$$

or :

$$\text{Genomic Copy No. of IE} = \frac{\text{sum of positive plaques}}{\text{total number of plaques}} \times \frac{\text{size of mouse genome}}{\text{average size of } \lambda \text{ insert}}$$

Hence :

$$\text{Copy number of IE 118a per mouse haploid genome} = \frac{42}{8056} \times \frac{3 \times 10^6}{15} = 1040$$

$$\text{Copy number of IE 118b per mouse haploid genome} = \frac{78}{8056} \times \frac{3 \times 10^6}{15} = 1940$$

4.2 Analysis of the Inserted Sequence in Clone λmA119

4.2.1 Nucleotide Sequence of the Inserted Sequence in Clone λmA119

Figure 3.6 shows a detailed partial restriction map of clone λmA119 in the vicinity of the actin-like gene containing the inserted sequence. Subclones

Table 4.1 Frequency of plaque hybridisation with probes from different inserted elements

Plate number	Plaques/plate	Plaques hybridising to IE probes			
		IE 118a	IE 118b	IE 119c	IE 36d
I	1275	5	9	12	13
II	1446	11	11	10	11
III	1357	9	16	22	11
IV	1331	6	14	17	12
V	1330	6	12	17	14
VI	1317	5	16	14	14
Total	8056	42	78	92	75

The number of the positive plaques hybridised to the probes IE118a, IE118b, IE119c, and IE36d per plate of the bacteriophage lambda mouse genomic library (DBA/2J) is presented.

containing the inserted sequence were derived from its internal EcoRI and SstI sites, and were designated 119XR and 119RX, and 119SS, respectively (see section 3.2.2). In order to further investigate the nature of the inserted sequence, the nucleotide sequence of the regions containing the inserted sequence in subclones 119XR, 119RX, and 119SS were determined.

Sequencing of the inserted sequence in clone λmA119 was as follows. Sequencing from the EcoRI and SstI sites (Figure 3.9; sequences number 6 to 8, 13, and 14) allowed determination of a total of 450 nucleotides of the inserted sequence. In order to identify further restriction sites suitable for sequencing, a 0.9kb BglII-SstI fragment was isolated from subclone 119BX (see section 2.7.3), and subjected to extensive restriction analysis. The eventual identification of the restriction sites XhoII and StyI enabled further sequencing from these sites (sequences number 3 to 5, and 16 to 18), and allowed the completion of the determination of this inserted sequence.

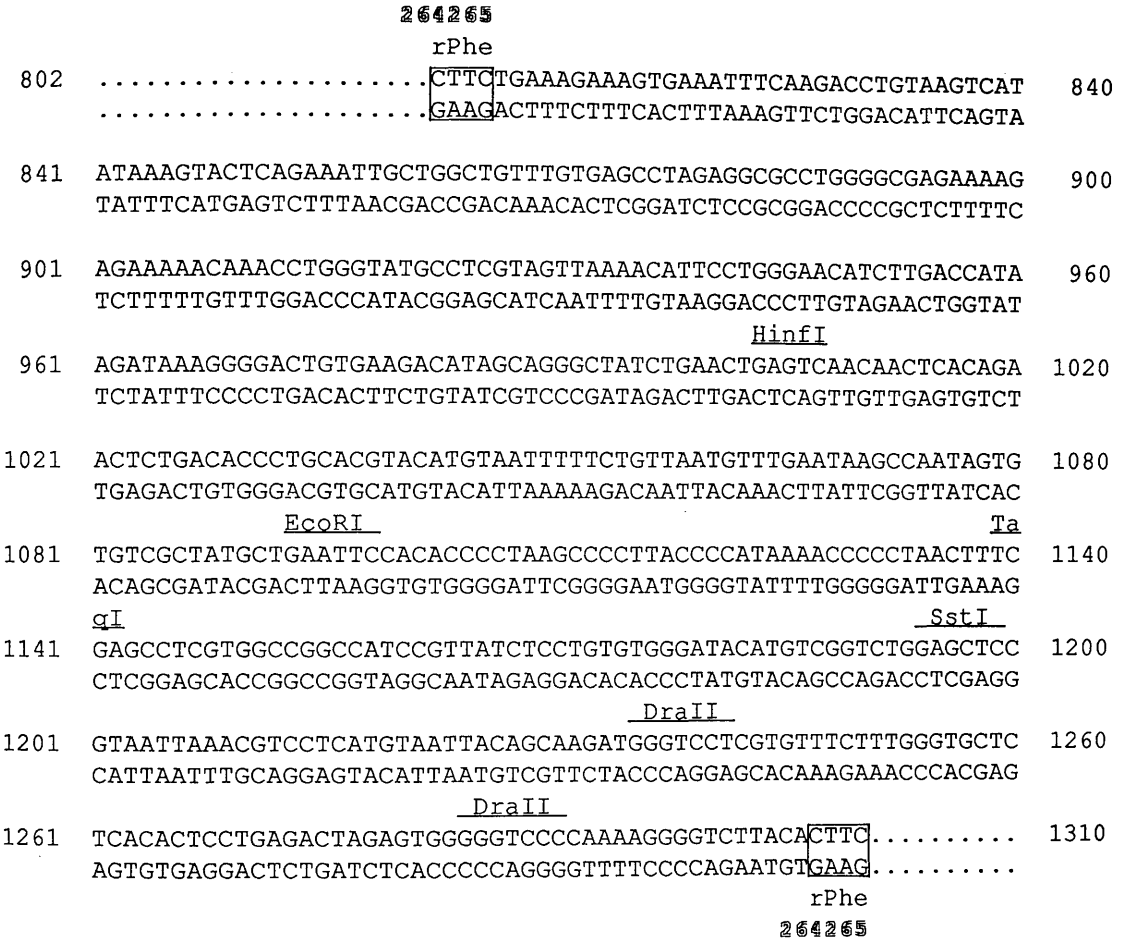
The complete nucleotide sequence of the inserted sequence in clone λmA119 is presented in Figure 4.5. The inserted sequence begins at nucleotide number 806 and ends at nucleotide number 1306, with a total length of 501 base pairs. The inserted sequence begins following the nucleotides encoding Phe^{265} of the γ-actin-like gene of λmA119. The actin-like gene appears to resume at the third nucleotide of Ser^{264} ; this particular nucleotide, and those of Phe^{265} being repetitions of nucleotides preceding the start of the inserted sequence. The inserted sequence is therefore flanked by a short direct repeat of 4 base pairs of actin sequence, indicating that it was inserted at a staggered break. The inserted sequence of λmA119 was therefore designated IE 119 (inserted element 119).

4.2.2 Computer Analysis of IE 119

The sequence of IE 119 was subjected to analysis on the VAX cluster at EMBL, Heidelberg, and compared with sequences in the GenBank and EMBL nucleotide sequence databases using the programs, WORDSEARCH together with SEGMENTS (section 2.17.3 and section 4.1.2) of the UWGCG sequence software package (Devereux *et al.*, 1984).

Computer searching of the nucleotide sequence databases revealed that IE 119 was homologous to a sequence MS57 (Propst and Vande Woude, 1984), which had been observed to have features similar to retroviral long terminal repeats. Detailed comparisons will be made in the Discussion.

Figure 4.5 Nucleotide sequence of IE 119



The nucleotide sequence of both strands of the inserted sequence in genomic clone λmA119 is shown. Restriction sites used in sequencing are indicated, as are the amino-acid equivalents of the flanking nucleotides of the actin processed pseudogene. The boxed sequences are direct repeats flanking the inserted sequence. The numbering of the sequence is as in Figure 3.10.

4.2.3 Genomic Southern Blotting of IE 119

In order to investigate the occurrence of sequences related to IE 119 in the mouse genome, Southern blotting of digests of mouse genomic DNA was performed.

Mouse liver DNA from the inbred strain Balb/C was isolated (see section 2.15.1), digested with restriction enzymes EcoRI, HindIII, and BamHI, subjected to gel electrophoresis in 1% agarose, and transferred to nitrocellulose filters (see section 2.9.1). A ^{32}P-labelled DNA probe from IE 119 (Figure 3.9) was prepared from a 0.48 kb StyI-SstI fragment (designated 119c), which contained 80 nucleotides of the actin-like sequence, as computer analysis (program CUTSIT, section 2.17.2) showed that there were no suitable restriction sites to allow a fragment of reasonable size to be isolated uniquely from IE 119. The labelled probe was hybridised to the restricted genomic DNA attached to the filter (see section 2.9.3).

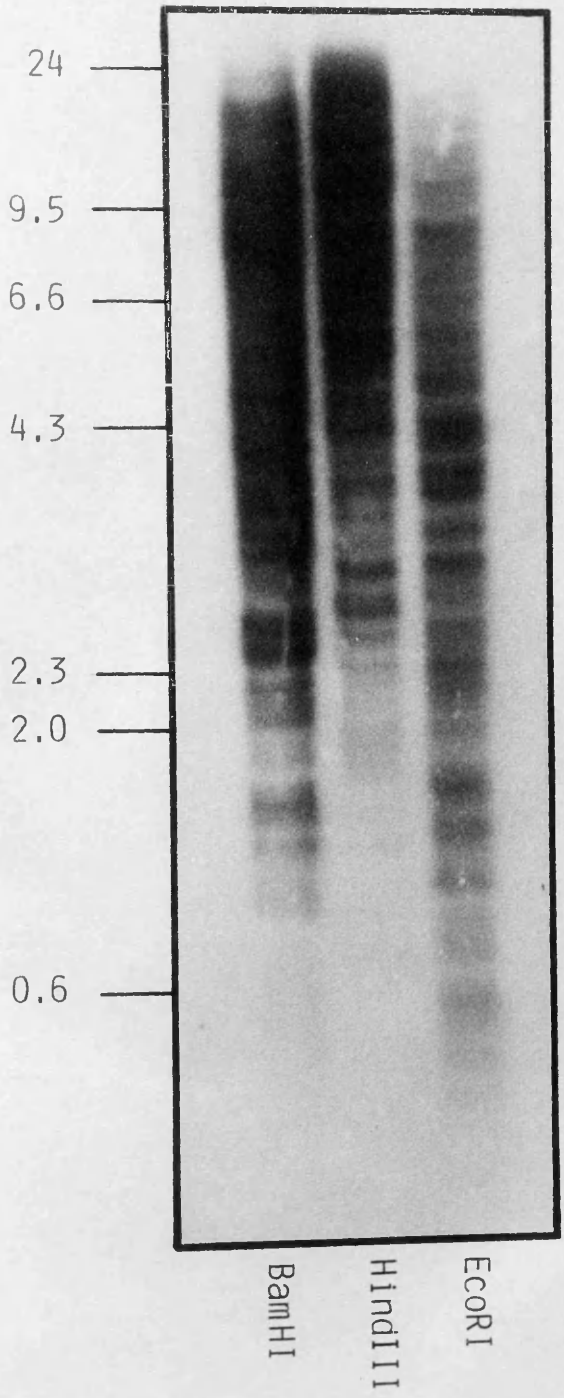
Figure 4.6 shows the genomic Southern blots of mouse DNA hybridised to the probe from IE 119. It can be seen that a large number of hybridising bands were obtained. This was far more than the 10 to 20 copies seen with actin probes (Minty *et al.*, 1983). Thus it is clear that IE 119 is repeated in the mouse genome.

4.2.4 Estimation of Copy Number of IE 119 by Plaque Hybridisation

In order to obtain an estimate of the copy number of IE 119 in the mouse genome, a bacteriophage lambda mouse genomic library (DBA/2J) was screened (as in section 4.1.4) with the IE 119 probe. A ^{32}P-labelled probe 119c was prepared from IE 119 as in section 4.2.3, and hybridised to the plaques attached to the filters (see section 2.16.2).

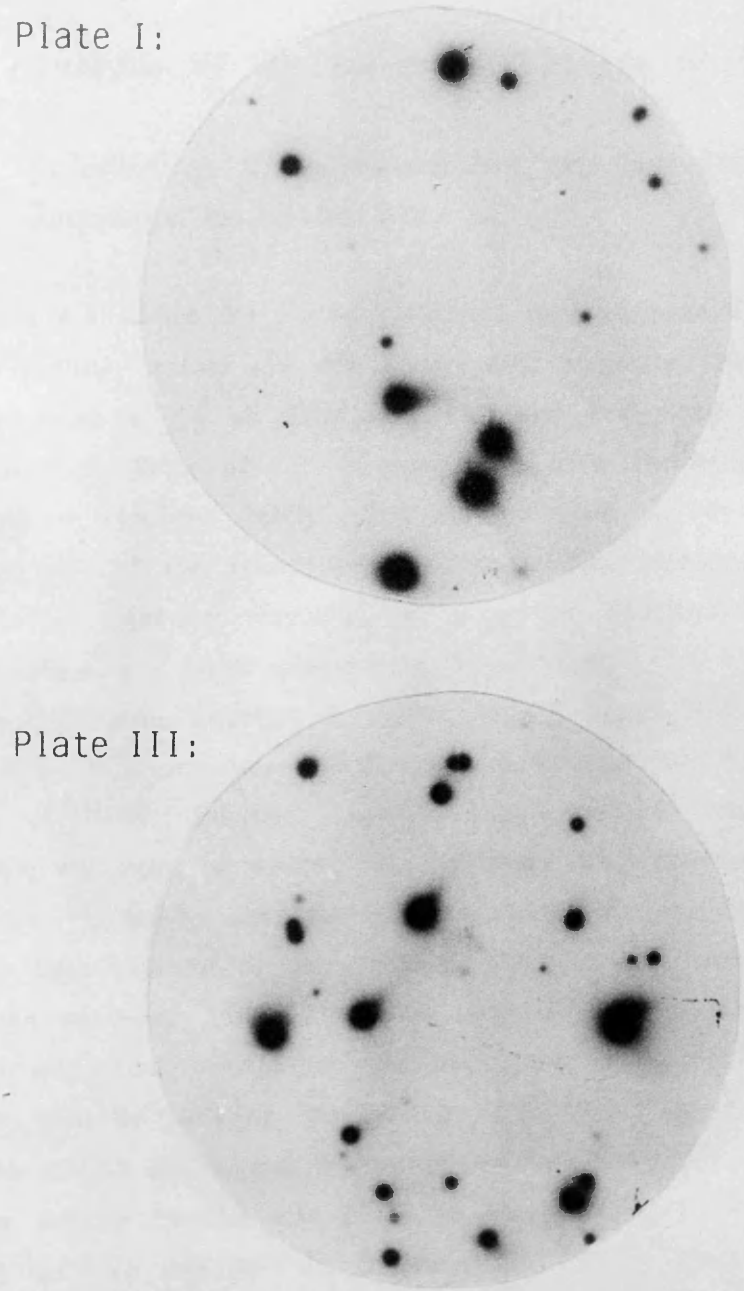
Figure 4.7 shows the plaques of the mouse genomic lambda library which hybridised to the probe from IE 119. The frequencies of the positive plaque hybridisation between the probe 119c and the lambda library is shown in Table 4.1. Probe 119c hybridised with a total of 92 plaques, out of a total number of 8056 plaques. The copy number of IE 119 sequences in the mouse genome is calculated by the equation given in section 4.1.4.

Figure 4.6 Genomic Southern blot of mouse DNA hybridised to a probe from IE 119



Mouse liver DNA (Balb/C strain) was digested with restriction enzymes EcoRI, HindIII, and BamHI as indicated, subjected to electrophoresis, transferred to nitrocellulose filters, and hybridised to a ^{32}P-labelled StyI-SstI fragment, designated 119c, from IE 119.

Figure 4.7 Hybridisation of a probe from IE 119 to plaques of a recombinant lambda mouse genomic library



A DBA/2J mouse genomic library in bacteriophage lambda was screened with a ^{32}P-labelled StyI-SstI fragment, designated 119c, from IE 119. Two such plates I and III, containing positive hybridisation plaques are presented (see Table 4.1).

Thus :

$$\text{Copy number of IE 119 per mouse haploid genome} = \frac{92}{8506} \times \frac{3 \times 10^6}{15} = 2280$$

4.3 Analysis of the Inserted Sequence in Clone λmA36

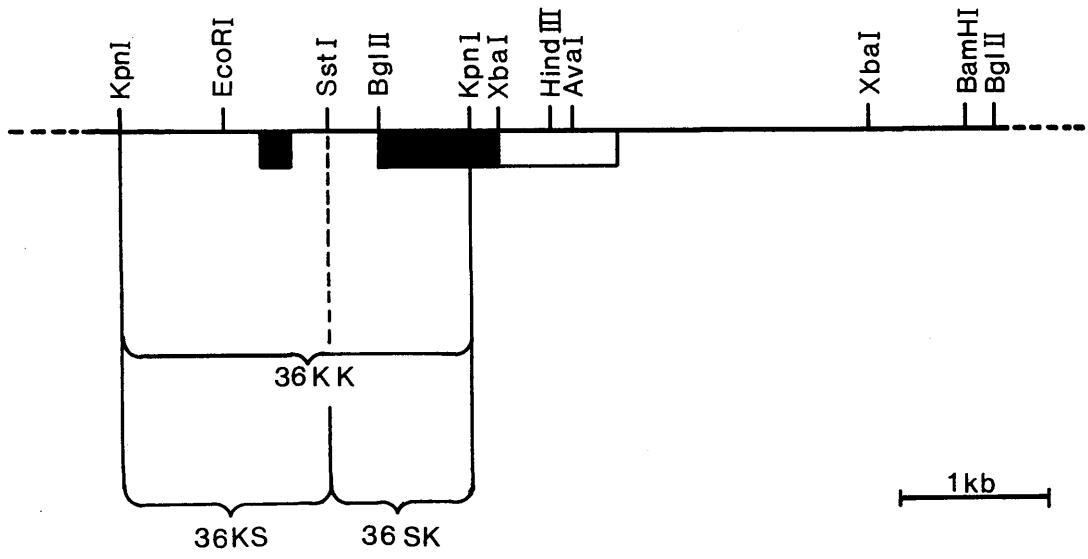
4.3.1 Subcloning and Sequencing of the Inserted Sequence in Clone λmA36

Figure 4.8 shows the partial restriction map of λmA36 in the vicinity of the actin coding region and the interrupting sequence. An original subclone, constructed from a 2.4 kb KpnI fragment, and designated 36KK was obtained from Miss C.E. Begg of this department. Two further subclones from the original parent subclone 36KK, were derived from an internal SstI site, each containing part of the interrupting sequence. The subclones were constructed using a 1.4 kb KpnI-SstI fragment and a 1.0 kb SstI-KpnI fragment, and were designated 36KS and 36SK respectively (Figure 4.8).

The sequencing strategy is summarised in Figure 4.9, from which it can be seen that the region was finally found to contain four SstI sites, rather than the one originally assumed. This together with other factors, caused considerable difficulty in determining the sequence. Sequences number 1 to 5 was carried out in the subclone 36SK without any problems. Sequencing from polylinker sites adjacent to the flanking SstI site and from the internal EcoRI site in the subclone 36KS was more difficult because there was no known secondary restriction site in between these (see section 2.11.3). This problem was dealt with by isolating the 0.8 kb EcoRI-SstI fragment from the subclone 36KS (see section 2.7.3), and performing detailed restriction analysis on it. The restriction enzyme Fnu4HI was found to cleave the 0.8 kb fragment into two, producing 0.15 kb and 0.65 kb fragments. However when sequence number 6 was performed from the SstI (polylinker EcoRI) site on the subclone 36KS, a further two SstI sites were identified 40 base pairs apart. Thus, together with the previously identified SstI site, a total of three SstI sites were observed. A similar sequence (number 7) was carried out on the opposite strand and confirmed this finding.

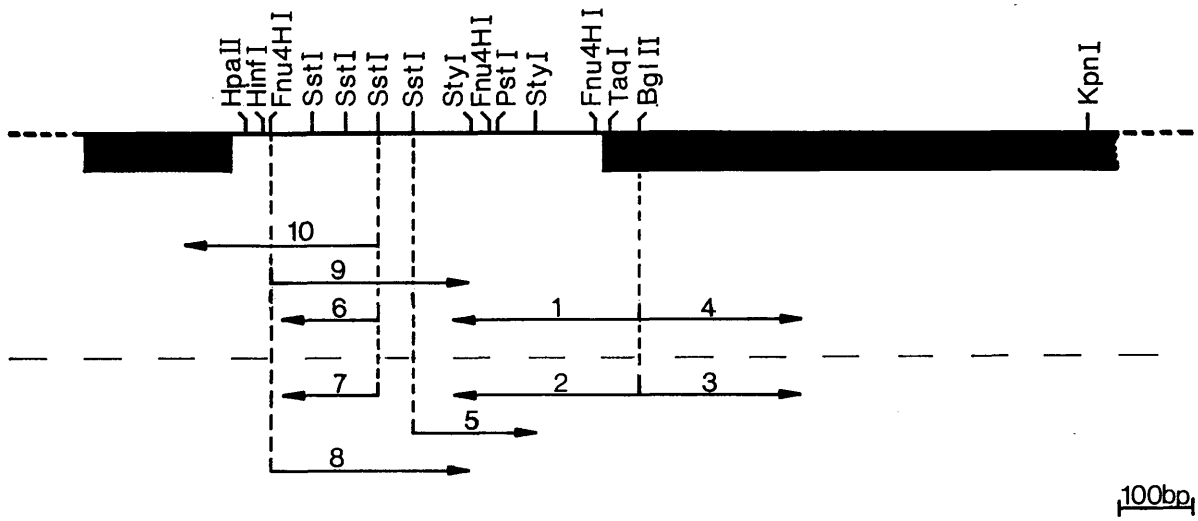
At this point it was felt necessary to check the relationship between the

Figure 4.8 Partial restriction map of clone λmA36 and its subclones in the vicinity of the actin-like gene



A partial restriction map of a portion of genomic clone λmA36 is presented with reference to its derived subclones. The solid blocks represent the presumed positions of actin coding regions, the open blocks represent the presumed position of 3' non-coding regions, and the vertical lines indicate the approximate positions of the restriction sites. The derived subclones 36KK, 36KS, and 36SK are indicated.

Figure 4.9 Sequencing strategy for the interrupted DNA of the genomic clone λmA36



The strategy for determining the nucleotide sequence in the region from the Fnu4HI site at the 5' of the interrupted DNA to the 3' BglII site in the actin pseudocoding area of clone λmA36 is shown. Sequencing of the coding and non-coding strands are indicated by arrows below and above the broken line, respectively. Fragments are numbered for ease of reference. The arrows represent the portion of sequence read from a particular restriction site.

40 base pair 'SstI duplication' in 36KS and the situation in λmA36. Further restriction analysis of the original 36KK subclone was therefore performed as this clone encompassed the whole of the inserted sequence and flanking regions. This demonstrated the presence of a small SstI fragment of approximately 40 base pairs. However it was not possible to determine whether there were multiple copies of the sequence, and if so, how many. If there were indeed multiple copies in λmA36 (and in 36KK) the subclone 36KS might not have contained all of these. Subsequent work was therefore carried out on the original subclone 36KK. This was not at all easy, however, as the insert of 36KK was 2.4 kb in length, and there was a lack of suitable infrequent restriction sites that produced 5' protruding ends for sequencing by the method of Maxam and Gilbert, while there were so many common restriction sites within this region that mapping was not easily accomplished.

A 1.6kb PstI fragment was isolated from the 36KK subclone and extensive restriction analysis was carried out on this isolated fragment. No useful restriction sites for sequencing were found, except for the known Fnu4HI site. Sequences number 8 and 9 were carried out by initially isolating the 1.6 kb PstI fragment from subclone 36KK, restricted with Fnu4HI, labelling with a selection of α^{32}P-dNTPs and γ^{32}P-ATP at the 5' protruding end of Fnu4HI, and performing the sequencing reaction (see section 2.12). This established that there were in fact four internal SstI sites in 36KK, rather than the one originally identified by mapping, or the three that were previously found in 36KS. The sequence to the left of the Fnu4HI site, covering the region towards the 5' end of the inserted sequence, could not be obtained by the above method, though a selection of radioactive labels were used. The reason for this was never discovered.

A further attempt to obtain the sequence of the region between the Fnu4HI site and the start of the inserted sequence was made by constructing smaller subclones derived from 36KS but with fewer SstI sites. Thus it was hoped that a single sequencing gel would allow reading along a shorter distance than in 36KS to the start of the inserted sequence. The smaller subclones were constructed by completely digesting 36KS with SstI, separating the bulk of the clone from the 40 base pair SstI fragments by gel electrophoresis, and religating the remainder of the subclone. The smaller subclones were made successfully, but detailed restriction analysis on the 0.7 kb EcoRI-SstI insert did not detect suitable restriction sites for secondary cleavage in the sequencing reaction.

Because of this a totally different strategy was employed, that of the Sanger Chain Termination method (see sections 2.13 and 2.14). A 0.7 kb EcoRI fragment of 36KS (one of the EcoRI sites was provided by the vector pUC18) was ligated into the EcoRI sites of the vectors M13mp18 and M13mp19 (Figure 2.5), transformed into *E.coli* JM109 'competent' cells, single-stranded templates were prepared, and the sequence determined (sequence number 10; Figure 4.9). A total of 250 bases pairs were determined from sequence number 10 which enabled the complete covering of the inserted sequence in λmA36.

The complete nucleotide sequence of the inserted sequence in clone λmA36 is shown in Figure 4.10. The start of the inserted sequence is numbered 1, and the end numbered 500, with a total length of 500 base pairs.

The inserted sequence starts after the nucleotides encoding Ile^{71} of the γ-actin-like gene of λmA36. The actin-like gene appears to resume at nucleotides encoding Pro^{70} ; these three nucleotides, and those of Ile^{71} being repetitions of nucleotides preceding the start of the inserted sequence. The inserted sequence is therefore flanked by a short direct repeat of 6 base pairs of actin sequence, indicating that it was inserted into a staggered break. The inserted sequence of λmA36 was therefore designated IE 36 (inserted element 36).

4.3.2 Computer Analysis of IE 36

The sequence of IE 36 was subjected to analysis on the VAX cluster at EMBL, Heidelberg, and compared with sequences in the GenBank and EMBL nucleotide sequence databases using the programs, WORDSEARCH together with SEGMENTS (see sections 2.17.3 and 4.1.2) of the UWGCG sequence software package (Devereux *et al.*, 1984).

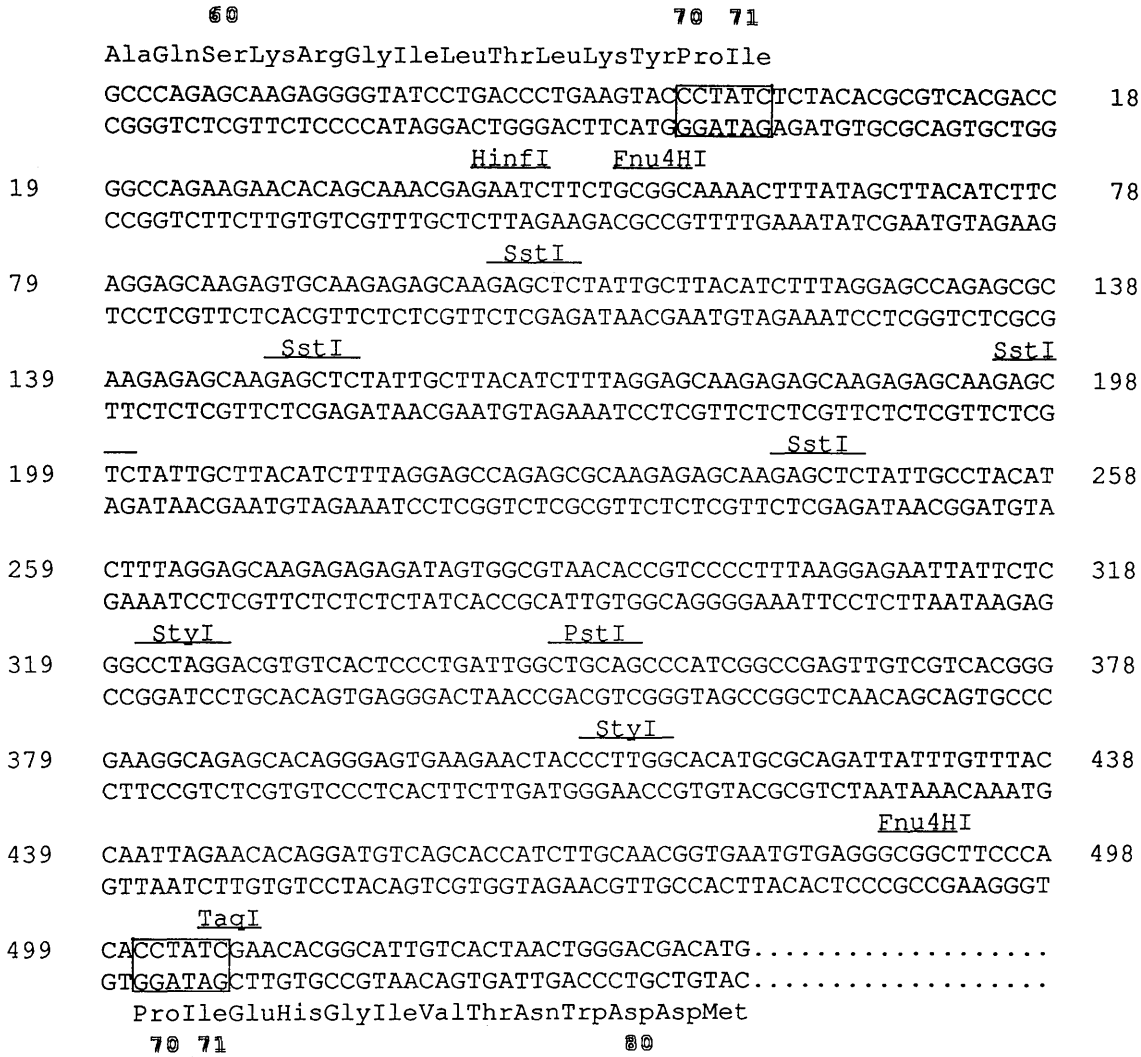
Computer searching of the nucleotide sequence databases revealed that IE 36 was related to the long terminal repeat of the retroviral-like mouse intracisternal A-particles. Detailed comparisons will be made in the Discussion.

4.3.3 Genomic Southern Blotting of IE 36

In order to investigate the occurrence of sequences related to IE 36 in the mouse genome, Southern blotting of digests of mouse genomic DNA was performed.

Mouse liver DNA from the inbred strain Balb/C was isolated (see section 2.15.1), digested with restriction enzymes EcoRI, HindIII, and BamHI, subjected

Figure 4.10 Nucleotide sequence of IE 36



The nucleotide sequence of both strands of the inserted sequence in genomic clone λmA36 is shown. Restriction sites used in sequencing are indicated, as are the amino-acid equivalents of the flanking nucleotides of the actin processed pseudogene. The boxed sequences are direct repeats flanking the inserted sequence.

to gel electrophoresis in 1% agarose, and transferred to nitrocellulose filters (see section 2.9.1). A ^{32}P-labelled DNA probe from IE 36 (Figure 4.9) was prepared from a 0.31 kb HinfI-PstI fragment (designated 36d), and hybridised to the restricted genomic DNA attached to the filter (see section 2.9.3).

Figure 4.11 shows the genomic Southern blots of mouse DNA hybridised to the probe from IE 36. It can be seen that a large number of hybridising bands were obtained, indicating that IE 36 is repeated in the mouse genome.

4.3.4 Estimation of Copy Number of IE 36 by Plaque Hybridisation

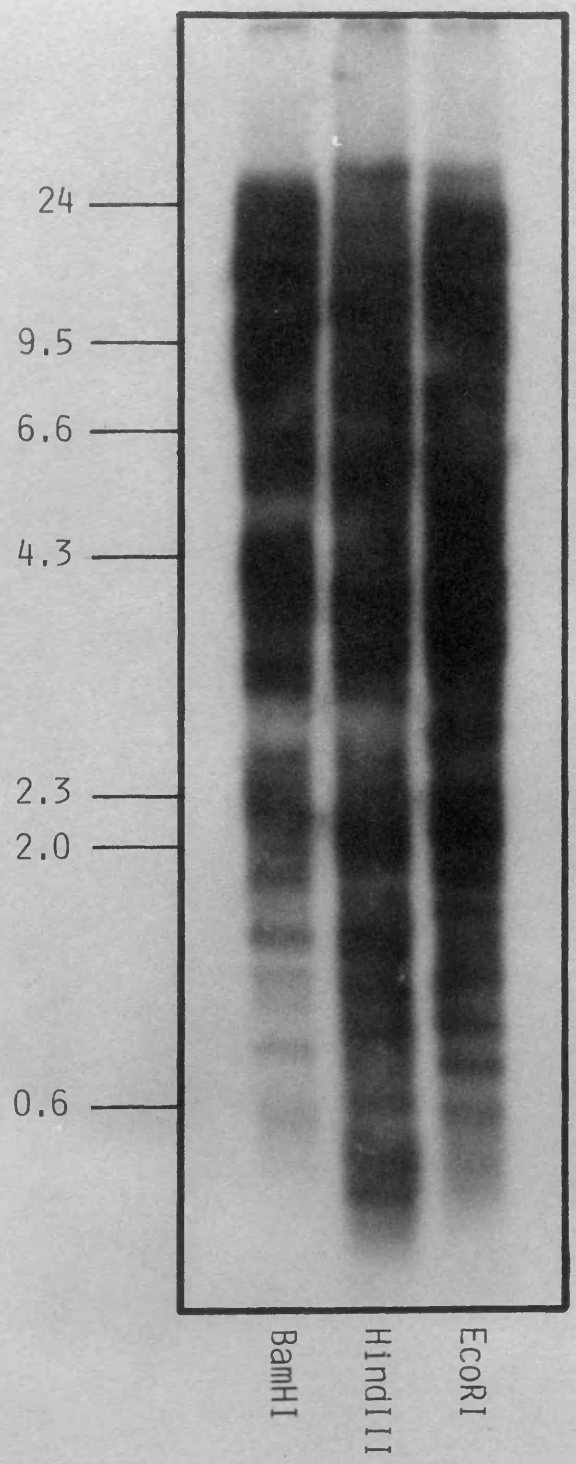
In order to obtain an estimate of the copy number of IE 36 in the mouse genome, a bacteriophage lambda mouse genomic library (DBA/2J) was screened (as in section 4.1.4) with the IE 36 probe. A ^{32}P-labelled probe 36d was prepared from IE 36 as in section 4.3.3, and hybridised to the plaques attached to the filters (see section 2.16.2).

Figure 4.12 shows the plaques of the mouse genomic lambda library which hybridised to the probe from IE 36. The frequencies of the positive plaque hybridisation between the probe 36d and the lambda library is shown in Table 4.1. Probe 36d hybridised with a total of 75 plaques, out of a total number of 8056 plaques. The copy number of IE 36 sequences in the mouse genome is calculated by the equation given in section 4.1.4.

Thus :

$$\text{Copy number of IE 36 per mouse haploid genome} = \frac{75}{8506} \times \frac{3 \times 10^6}{15} = 1860$$

Figure 4.11 Genomic Southern blot of mouse DNA hybridised to a probe from IE 36



Mouse liver DNA (Balb/C strain) was digested with restriction enzymes EcoRI, HindIII, and BamHI as indicated, subjected to electrophoresis, transferred to nitrocellulose filters, and hybridised to a ^{32}P-labelled HinfI-PstI fragment, designated 36d, from IE 36.

Figure 4.12 Hybridisation of a probe from IE 36 to plaques of a recombinant lambda mouse genomic library

Plate I:

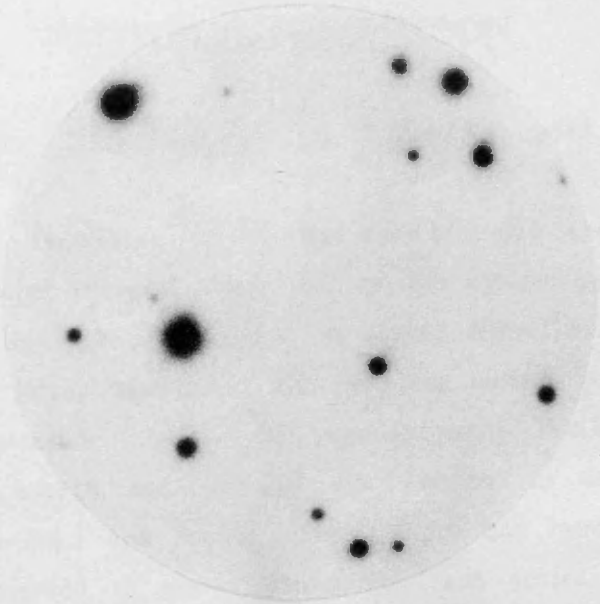
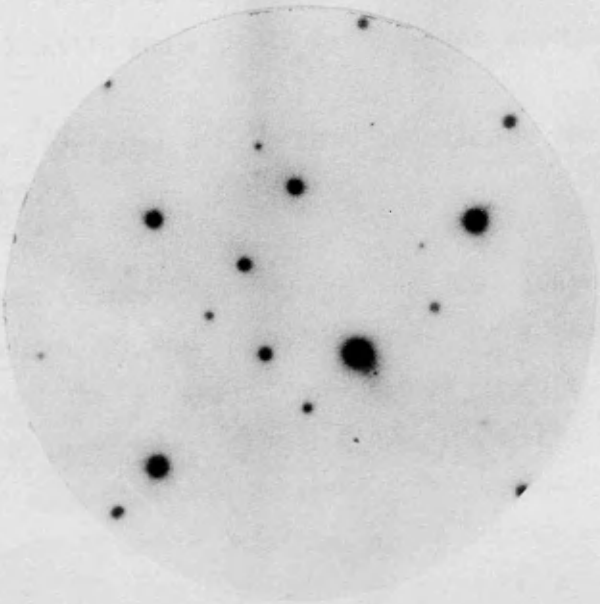


Plate VI:



A DBA/2J mouse genomic library in bacteriophage lambda was screened with a ^{32}P-labelled HinfI-PstI fragment, designated 36d, from IE 36. Two such plates I and VI, containing positive hybridisation plaques are presented (see Table 4.1).

CHAPTER 5

GENERAL DISCUSSION

5.1 Actin-like Pseudogenes in λmA118 and λmA119

The actin-like sequences of clones λmA118 and λmA119 have been determined, and both of these resemble that of the cytoplasmic γ-actin isoform. However, the actin-like DNA sequences in clones λmA118 and λmA119 were observed to possess many mutational changes that clearly indicate them to be functionless, and establish them to be pseudogenes. Furthermore, the actin-like sequences in λmA118, and λmA119 were shown to be uninterrupted by introns (see sections 3.4.1 and 3.4.2). The lack of any introns, including those that have been conserved in mammalian actins, and others which are specific for different isoforms (Carroll *et al.*, 1986; Chang *et al.*, 1984; Hamada *et al.*, 1982; Ng *et al.*, 1985; Bergsma *et al.*, 1985; Foran *et al.*, 1985), is consistent with the pseudogenes in λmA118 and λmA119 being of the processed type.

5.1.1 Possible Origins of Actin-like Genes in λmA118 and λmA119

Processed pseudogenes are thought to be generated from reverse transcripts of the mRNA, as they normally contain DNA copies of the whole of the mRNA, including the 3' and 5' untranslated regions. Target-site direct repeats flanking the pseudogene clearly define its extent. As mentioned in section 3.4.1, the actin-like coding amino-acid sequence in clone λmA118 appears only to extend in the 5' direction to the residue at position 5 (Figure 3.11 and Figure 3.12), and thus lacks the region corresponding to the 5' end of the mRNA. This divergence from a typical processed pseudogene could have been caused by a deletion after the presumed reverse transcript was integrated into the genome. An alternative explanation is that clone λmA118 may have originated from either a 5'-truncated mRNA, or from an incomplete or partially degraded reverse transcript of a full length mRNA. However, since the

sequence of the 3' untranslated region of clone λmA118 has not been determined, thus preventing the identification of its presumed flanking direct repeat, it is uncertain which explanation is correct. A human cytoplasmic γ-actin processed pseudogene has recently been sequenced and shows no sign of 5' truncation (Leube and Gallwitz, 1986). However, two examples of such 5' truncation have been reported : in the mouse γ-actin processed pseudogene in clone λmA19 (MγA-ψ1; Leader *et al.*, 1985), where the actin-like coding amino-acid sequence extended up to the Ala at position 7 ; and in the processed pseudogene derived from mouse cellular tumor antigen p53, where at least 80 nucleotides are missing from a long 5'-untranslated region (Zakut-Houri *et al.*, 1983). In both these cases the presence of one member of the direct repeat immediately adjacent to the truncated 5' end indicates that these pseudogenes are derived from incomplete or partially degraded transcripts. However, reports have suggested that the 5' truncation in certain other processed pseudogenes may be a result of insertion of a transposon into the 5' end of these (Shimida *et al.*, 1984; Scarpulla, 1984). In the case of λmA118, the 40 nucleotides sequenced preceding the start of the pseudogene do not correspond to any sequence in the EMBL or GenBank databases, thus ruling out a possible 5' insertion of known mobile mouse sequences such as B1 or L1Md. In other cases, the processed pseudogenes appear to be derived from aberrant transcripts generated by faulty splicing or by initiation down-stream from the normal cap site. Examples of these are the human immunoglobulin lambda light chain (Hollis *et al.*, 1982), and the human immunoglobulin epsilon heavy chain (Battey *et al.*, 1982; Ueda *et al.*, 1982) processed pseudogenes. These processed pseudogenes are unusual in being derived from mRNA not normally expressed in the germ line. Their derivation from aberrant transcripts may well be related to this fact, and there is no reason to expect that the pseudogene in λmA118 will resemble these.

In the case of clone λmA119, the actin-like coding amino-acid sequence extended at least to the initiating Met codon at position 1 (Figure 3.13). As the sequence of the 5' untranslated region of mouse γ-actin mRNA has not yet been determined, it is not possible to say how much further the processed pseudogene in λmA119 extends. The 3' untranslated region also resembles that of mouse γ-actin (see section 3.4.2 and Figure 3.14), but only extends to nucleotide 1750 (*ie* for 108 out of approximately 700 nucleotides), after which the sequence diverges from that of γ-actin mRNA. No poly A tail or identifiable 3' flanking direct repeat to sequences near the 5' end of the pseudogene are

evident. Nor do the 3' sequences bear any relationship to known retroposons. It therefore seems probable that this 3' divergence of the actin-like gene in clone λmA119 from the γ-actin mRNA is the result of a deletion of the 3' untranslated region after the presumed reverse transcript was integrated into the genome.

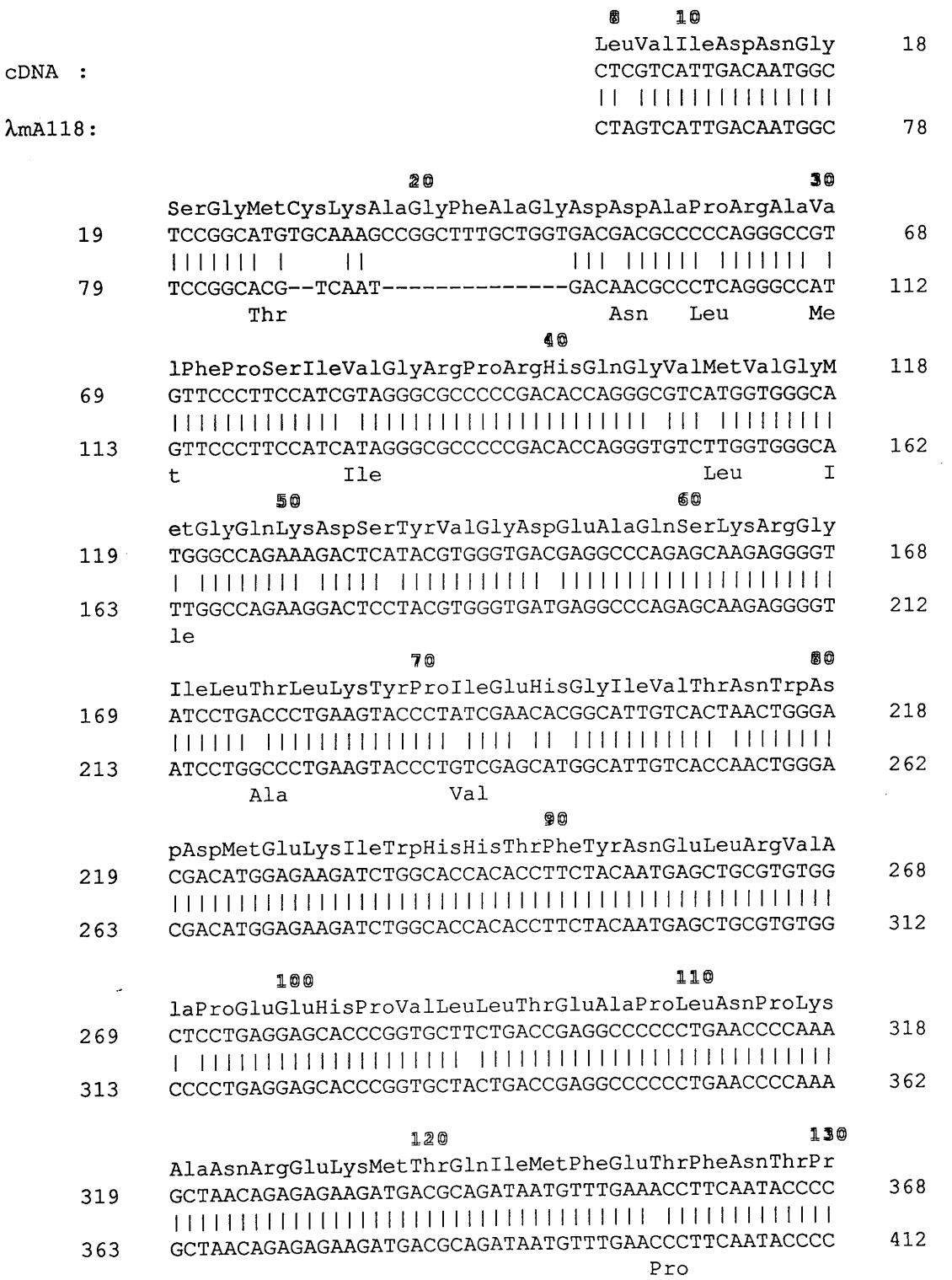
5.1.2 Evolution of Actin-like Genes λmA118 and λmA119

In order to determine the nature of the processed pseudogenes λmA118 and λmA119, and their relatedness to the functional actin gene, a comparison between their γ-actin region and that of the γ-actin cDNA was made. The nucleotide sequences of the actin-like DNA in clones λmA118 and λmA119 were compared with the partial sequence of a mouse γ-actin cDNA containing the region from amino-acid 8 to 375 (Peter and Leader, unpublished; Figure 5.1 and Figure 5.2). The comparison was made by using the computer program BESTFIT (see section 2.17.3), where alignments depended on the introduction of gaps for maximal homology.

The actin-like coding sequence of λmA118 showed a high degree of homology of 94.0% to the cDNA (Figure 5.1). There were 67 base changes, of which 26 were at silent sites, 35 altered the amino-acid residue, and 6 were base deletions. Comparison of the nucleotide sequence in the actin-like coding region in λmA119 showed a homology of 96.1% to the cDNA (Figure 5.2). There were 43 base changes, of which 20 were at silent sites, 22 altered the amino-acid residue, and 1 was a base insertion. The differences of these processed pseudogenes (as well as that of the previously published pseudogene of λmA19) from the cDNA are shown in Figure 5.3.

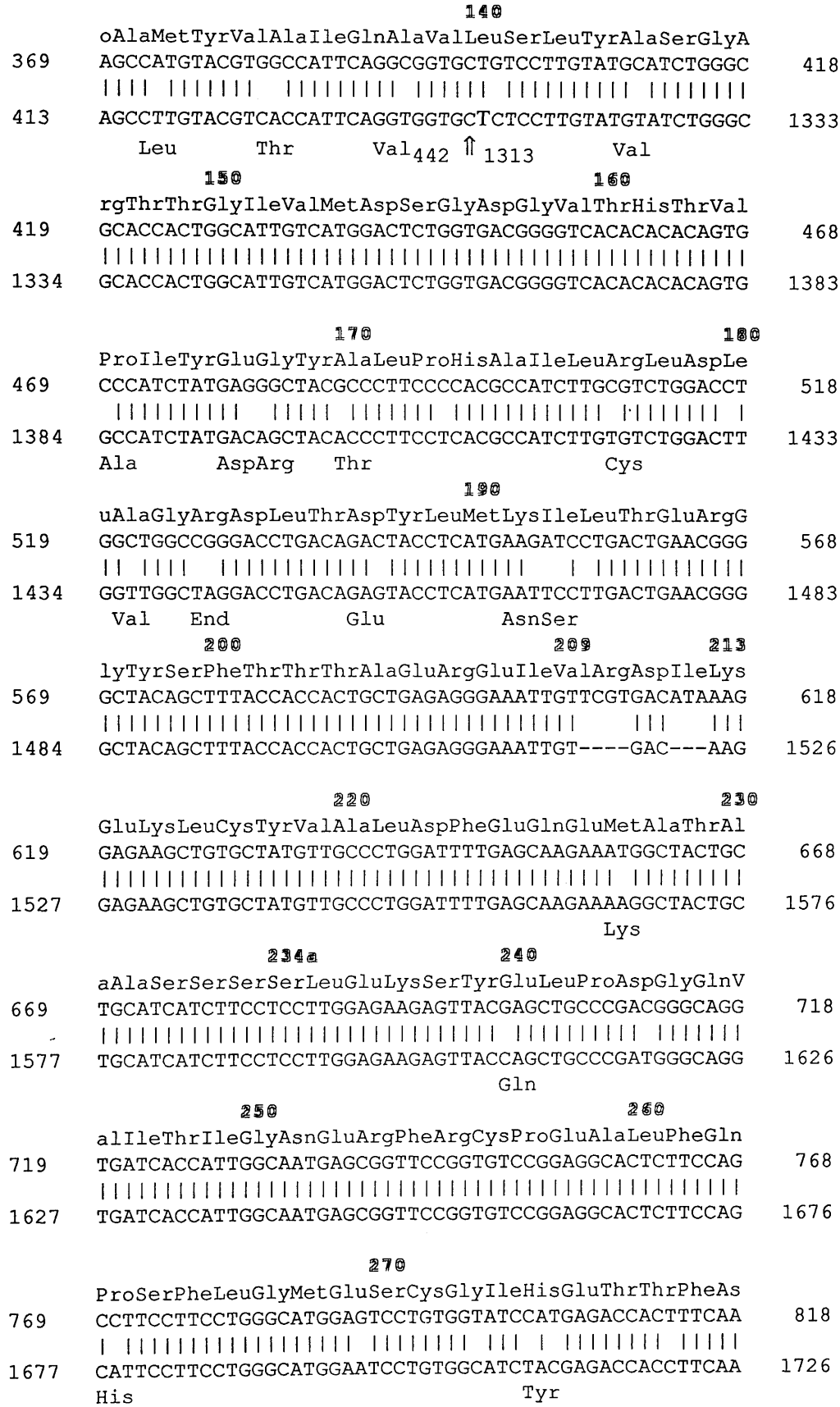
A numerical estimate of the divergence time of pseudogenes can be obtained from their percentage divergences from the functional gene, assuming that, being inactive since their formation, changes accumulate in processed pseudogenes at a constant rate in all positions, free from any selection. The percentage divergence can be related to evolutionary time using the value of 0.7 for the unit evolutionary period (UEP), the time in millions of years required for the fixation of 1% change between two sequences (Perler *et al.*, 1980). However, to estimate the time of divergence of a pseudogene from the active γ-actin gene, allowance must be made for the fact that effectively only 24.4% of the nucleotides in the coding region of this gene can undergo (neutral) mutation as the amino-acid sequence has been absolutely conserved.

Figure 5.1 A comparison between actin-like sequences of clone λmA118 and the partial sequence of mouse γ-actin cDNA



continued overleaf. . . .

continued. . . .



continued overleaf. . . .

continued. . . .

```
          280                                290
         nSerIleMetLysCysAspValAspIleArgLysAspLeuTyrAlaAsnT
819      CTCCATCATGAAGTGTGATGTGGATATCCGCAAAGACCTGTATGCCAATA    868
         ||||||||||||||||||||||||||| ||||||||||||||||||||
1727     CTCCATCATGAAGTGTGATGTGGATATCTGCAAAGACCTGTATGCCAATA   1776
                                    Cys
                   300                              310
         hrValLeuSerGlyGlyThrThrMetTyrProGlyIleAlaAspArgMet
869      CAGTGCTGTCTGGTGGTACCACCATGTACCCAGGCATTGCTGACAGGATG    918
         |||||||||| ||||||||||||||||||||||||||||||||||||||
1777     CAGTGCTGTCCGGTGGTACCACCATGTACCCAGGCATTGCTGACAGGATG   1826

                              320
         GlnLysGluIleThrAlaLeuAlaProSerThrMetLysIlelysIleIl
919      CAGAAGGAGATCACAGCCCTAGCACCTAGCACGATGAAGATTAAGATCAT    968
          |||||||||||||||||||||||| ||||| ||||||||||||||||
1827     TAGAAGGAGATCACAGCCCTAGCACCCAGCACAATGAAGATTAAGATCAT   1876
         End
          330                                340
         eAlaProProGluArgLysTyrSerValTrpIleGlyGlySerIleLeuA
969      TGCTCCCCCTGAGCGCAAGTACTCAGTCTGGATCGGTGGCTCCATTCTGG   1018
         |||||||||||||||||||||||||||||| | | |||||||| ||
1877     TGCTCCCCCTGAGCGCAAGTACTCAGTCTGGACCTGCGGCTCCATCCT-A   1925
                                       ThrCys             T
                   350                              360
         laSerLeuSerThrPheGlnGLnMetTrpIleSerLysGlnGluTyrAsp
1019     CCTCACTGTCCACCTTCCAGCAGATGTGGATCAGCAAGCAGGAGTATGAT   1068
         ||||||||||||||||||||||||||||||||||||||||||||||||||
1926     CCTCACTGTCCACCTTCCAGCAGATGTGGATCAGCAAGCAGGAGTATGAT   1975
         hr
                              370          374
         GluSerGlyProSerIleValHisArgLysCysPheEnd
1069     GAGTCAGGCCCCTCCATCGTCCACCGCAAATGCTTCTAG    1107
         ||||| |||    |||||||||||||||||||||||||
1976     GAGTCGGGC----CCATCGTCCACCGCAAATGCTTCTAG    2011
```

The nucleotide sequences that correspond to the predicted γ-like-actin in clone λmA118 is aligned to the corresponding regions of the partial sequence of mouse γ-actin cDNA (Peter and Leader, personal communication). Vertical lines between nucleotides indicate identity, and hyphens indicate deletions in sequence. The replacement amino-acids are listed below the sequence in question. The inserted sequence is not shown, but the arrow between nucleotides 442 and 1313 indicate its position.

Figure 5.2 A comparison between actin-like sequences of clone λmA119 and the partial sequence of mouse γ-actin cDNA

```
                                                  8    10
                                                LeuValIleAspAsnGly
cDNA:                                           CTCGTCATTGACAATGGC      18
                                                |||||||||| |||||||
λmA119:                                         CTCGTCATTGTCAATGGC      46
                                                      Val
                          20                                  30
                    SerGlyMetCysLysAlaGlyPheAlaGlyAspAspAlaProArgAlaVa
          19        TCCGGCATGTGCAAAGCCGGCTTTGCTGGTGACGACGCCCCCAGGGCCGT      68
                    |||||||||||||||||| |||||||||| |||||||||||||||||||
          47        TCCGGCATGTGCAAAGCTGGCTTTGCTGGAGACGACGCCCCCAGGGCCGT      96

                                        40
                    lPheProSerIleValGlyArgProArgHisGlnGlyValMetValGlyM
          69        GTTCCCTTCCATCGTAGGGCGCCCCCGACACCAGGGCGTCATGGTGGGCA      118
                    |||||||||||||||||| |||||||||||||| ||||||||||||||
          97        GTTCCCTTCCATCGTAGGGTGCCCCCGACACCAGGACGTCATGGTGGGCA      146
                                Cys             Asp
                     50                                  60
                    etGlyGlnLysAspSerTyrValGlyAspGluAlaGlnSerLysArgGly
          119       TGGGCCAGAAAGACTCATACGTGGGTGACGAGGCCCAGAGCAAGAGGGGT      168
                    |||||||||||||||| || ||||||||| ||||||||||||||||||||
          147       TGGGCCAGAAAGACTCGTATGTGGGTGACAAGGCCCAGAGCAAGAGGGGT      196
                                        Lys
                          70                                  80
                    IleLeuThrLeuLysTyrProIleGluHisGlyIleValThrAsnTrpAs
          169       ATCCTGACCCTGAAGTACCCTATCGAACACGGCATTGTCACTAACTGGGA      218
                    |||||||||||||||||||||||||||||||||||||||||||| ||||||||
          197       ATCCTGACCCTGAAGTACCCTATCGAACACGGCATTGTCACCAACTGGGA      246

                                        90
                    pAspMetGluLysIleTrpHisHisThrPheTyrAsnGluLeuArgValA
          219       CGACATGGAGAAGATCTGGCACCACACCTTCTACAATGAGCTGCGTGTGG      268
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
          247       TGACATGGAGAAGATCTGGCACCACACCTTCTACAATGAGCTGCGTGTGA      296
                                                                    T
                     100                                 110
                    alProGluGluHisProValLeuLeuThrGluAlaProLeuAsnProLys
          269       CTCCTGAGGAGCACCCGGTGCTTCTGACCGAGGCCCCCCTGAACCCCAAA      318
                    | ||||||||||||||||||||||||||||||||||||||||||||||||
          297       CCCCTGAGGAGCACCCGGTGCTTCTGACCGAGGCCCCCCTGAACCCCAAA      346
                    hr
                          120                                 130
                    AlaAsnArgGluLysMetThrGlnIleMetPheGluThrPheAsnThrPr
          319       GCTAACAGAGAGAAGATGACGCAGATAATGTTTGAAACCTTCAATACCCC      368
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
          347       GCTAACAGAGAGAAGATGACGCAGATAATGTTTGAAACCTTCAATACCCC      396
```

continued overleaf. . . .

continued. . . .

```
                                      140
        oAlaMetTyrValAlaIleGluAlaValLeuSerLeuTyrAlaSerGlyA
369     AGCCATGTACGTGGCCATTCAGGCGGTGCTGTCCTTGTATGCATCTGGGC     418
        ||||||||||||||||||||||||||||||||||||||||||||||||||
397     AGCCATGTACGTGGCCATTCAGGCGGTGCTGTCCTTGTATGCATCTGGGT     446
                                                        C
              150                                 160
        rgThrThrGlyIleValMetAspSerGlyAspGlyValThrHisThrVal
419     GCACCACTGGCATTGTCATGGACTCTGGTGACGGGGTCACACACACAGTG     468
        ||||||||||||||||||||||||||||||||||||||||||||||||||
447     GCACCACTGGCATTGTCATGGACTCTGGTGACGGGGTCACACACACAGTG     496
        ys
                          170                             180
        ProIleTyrGluGlyTyrAlaLeuProHisAlaIleLeuArgLeuAspLe
469     CCCATCTATGAGGGCTACGCCCTTCCCCACGCCATCTTGCGTCTGGACCT     518
        |||||||||||||||||||||||||||| ||| ||||||||||||||||
497     CCCATCTATGAGGGCTACGCCCTTCCCCATGCCGTCTTGCGTCTGGACCT     546
                                        Val
                                    190
        uAlaGlyArgAspLeuThrAspTyrLeuMetLysIleLeuThrGluArgG
519     GGCTGGCCGGGACCTGACAGACTACCTCATGAAGATCCTGACTGAACGGG     568
        |||||| |||| ||||||||||||||||||||||||||||||||||||||
547     GGCTGGTCGGGTCCTGACAGACTACCTCATGAAGATCCTGACTGAACGGG     596
              Val
            200                                 210
        lyTyrSerPheThrThrThrAlaGluArgGluIleValArgAspIleLys
569     GCTACAGCTTTACCACCACTGCTGAGAGGGAAATTGTTCGTGACATAAAG     618
        |||||||||||||||||||||| |||||||||||||||| |||||||||
597     GCTACAGCTTTACCACCACTGCTAAGAGGGAAATTGTTCGAGACATAAAG     646
                              Lys
                          220                             230
        GluLysLeuCysTyrValAlaLeuAspPheGluGluGluMetAlaThrAl
619     GAGAAGCTGTGCTATGTTGCCCTGGATTTTGAGCAAGAAATGGCTACTGC     668
        |||||||||||||||||||||||| ||||||||||||||||||||||||
647     GAGAAGCTGTGCTATGTTGCCCTGTATTTTGAGCAAGAAATGGCTACTGC     696
                                  Tyr
                      234a                240
        aAlaSerSerSerSerLeuGluLysSerTyrGluLeuProAspGlyGlnV
669     TGCATCATCTTCCTCCTTGGAGAAGAGTTACGAGCTGCCCGACGGGCAGG     718
        | |||||||||||||||||||||||||||||||||||||||| |||||||
697     TACATCATCTTCCTCCTTGGAGAAGAGTTACGAGCTGCCCGATGGGCAGG     746
        Thr
                  250                                 260
        alIleThrIleGlyAsnGluArgPheArgCysProGluAlaLeuPheGln
719     TGATCACCATTGGCAATGAGCGGTTCCGGTGTCCGGAGGCACTCTTCCAG     768
        | ||||||||| |||||||||||||||||||| ||||||||||||||||
747     TTATCACCATCGGCAATGAGCGGTTCCGGTGTCCAGAGGCACTCTTCCAG     796

                                  270
        ProSerPheLeuGlyMetGluSerCysGlyIleHisGluThrThrPheAs
769     CCTTCCTTCCTGGGCATGGAGTCCTGTGGTATCCATGAGACCACTTTCAA     818
        |||||||||||||||||||||||||||||||| |||| ||| |||||
797     ACTTCCTTCCTGGGCATGGAGTCCTGTGGTATCCACGAGATCACCTTCAA    1351
        Thr 805 ⇑ 1310                                  Ile
```

continued overleaf. . .

continued. . . .

```
          280                                290
        nSerIleMetLysCysAspValAspIleArgLysAspLeuTyrAlaAsnT
819     CTCCATCATGAAGTGTGATGTGGATATCCGCAAAGACCTGTATGCCAATA    868
        ||||||||||||||||||||||||||||||||||||||||||||||||||
1352    CTCCATCATGAAGTGTGATGTGGATATCCGCAAAGACCTGTATGCCAATA   1401

                  300                                310
        hrValLeuSerGlyGlyT hrThrMetTyrProGlyIleAlaAspArgMe
869     CAGTGCTGTCTGGTGGTA*CCACCATGTACCCAGGCATTGCTGACAGGAT    917
        |||||||||||||||||| ||||||||||||||||||||||||||||||
1402    CAGTGCTGTCTGGTGGTACCCACCATGTACCCAGGCATTGCTGACAGGAT   1451

                            320
        tGlnLysGluIleThrAlaLeuAlaProSerThrMetLysIleLysIleI
918     GCAGAAGGAGATCACAGCCCTAGCACCTAGCACGATGAAGATTAAGATCA    967
        | ||||||||||||||| |||||||||| ||||| | ||||||||||||||
1452    GAAGAAGGAGATCACAACCCTAGCACCCAGCACAACGAAGATTAAGATCA   1501
         Lys             Thr                  Thr
          330                               340
        leAlaProProGluArgLysTyrSerValTrpIleGlyGlySerIleLeu
968     TTGCTCCCCCTGAGCGCAAGTACTCAGTCTGGATCGGTGGCTCCATTCTG   1017
        |||||||||||||||||||||||||||||| || ||||||||||||||
1502    TTGCTCCCCCTGAGCGCAAGTACTCAGTCTGGTTCTGTGGCTCCATTCTG   1551
                                       PheCys
                  350                                360
        AlaSerLeuSerThrPheGlnGlnMetTrpIleSerLysGlnGluTyrAs
1018    GCCTCACTGTCCACCTTCCAGCAGATGTGGATCAGCAAGCAGGAGTATGA   1067
        ||||||||||||||||||||||||||||||||||||||||||||||||||
1552    GCCTCACTGTCCACCTTCCAGCAGATGTGGATCAGCAAGCAGGAGTATGA   1601

                               370          374
        pGluSerGlyProSerIleValHisArgLysCysPheEnd
1068    TGAGTCAGGCCCCTCCATCGTCCACCGCAAATGCTTCTAGATGGACTGA*   1116
        |||||  |||||||| ||| |||||  |||||||||||||||||| |
1602    TGAGTTGGGCCCCTCTATCATCCACATCAAATGCTTCTAGATGGACCGTG   1651
             Leu            Ile   Ile

1117    GCAGGTGCCAGGCATCTGCTGCATGAGCTGATATTGAAGTATCAATTTGC   1166
        |||||||||| ||||||||  ||||||||||||||||||||  ||||||
1652    GCAGGTGCCAAGCATCTGCCGCATGAGCTGATATTGAAGTATTGATTTGC   1701

1167    CCTGGCAAATGTACACACCTCATGCTAGCCTCATGAAACTGGAATAAG     1214
        ||||||||||||||||||||||||||||||||||||||||||||||||
1702    CCTGGCAAATGTACACACCTCATGCTAGCCTCATGAAACTGGAATAAG     1749
```

The nucleotide sequences that correspond to the predicted γ-like-actin in clone λmA119 is aligned to the corresponding regions of the partial sequence of mouse γ-actin cDNA (Peter and Leader, personal communication). Vertical lines between nucleotides indicate identity. The replacement amino-acids are listed below the sequence in question. The inserted sequence is not shown, but the arrow between nuclectides 805 and 1310 indicate its position. Gaps introduced in the cDNA sequence for maximal homology are indicated by asterisks.

Figure 5.3 Comparison between actin-like sequences of clones λmA19, λmA118, and λmA119 and the partial sequence of mouse γ-actin cDNA

```
                                                  8     10
                                                 LeuValIleAspAsnGly
cDNA:                                            CTCGTCATTGACAATGGC      18
λmA19:                                           ..................
λmA118:                                          ..A................
λmA119:                                          ..........T.......
                                20                                30
          SerGlyMetCysLysAlaGlyPheAlaGlyAspAspAlaProArgAlaVa
cDNA:     TCCGGCATGTGCAAAGCCGGCTTTGCTGGTGACGACGCCCCCAGGGCCGT      68
λmA19:    ................................T.................
λmA118:   .......C.--TC..T--------------...A......T.......A.
λmA119:   .................T...........A.....................
                                      40
          lPheProSerIleValGlyArgProArgHisGlnGlyValMetValGlyM
cDNA:     GTTCCCTTCCATCGTAGGGCGCCCCCGACACCAGGGCGTCATGGTGGGCA     118
λmA19:    ................................................
λmA118:   .............A.....................T...T.........
λmA119:   ...................T..............A...............
                  50                                60
          etGlyGlnLysAspSerTyrValGlyAspGluAlaGlnSerLysArgGly
cDNA:     TGGGCCAGAAAGACTCATACGTGGGTGACGAGGCCCAGAGCAAGAGGGGT     168
λmA19:    ................G...............................
λmA118:   .T........G.....C...........T....................
λmA119:   ................G..T.........A....................
                            70                          80
          IleLeuThrLeuLysTyrProIleGluHisGlyIleValThrAsnTrpAs
cDNA :    ATCCTGACCCTGAAGTACCCTATCGAACACGGCATTGTCACTAACTGGGA     218
λmA19:    ........................................C........
λmA118:   ......G.............G....G..T...........C........
λmA119:   ........................................C........
                                      90
          pAspMetGluLysIleTrpHisHisThrPheTyrAsnGluLeuArgValA
cDNA :    CGACATGGAGAAGATCTGGCACCACACCTTCTACAATGAGCTGCGTGTGG     268
λmA19:    ................................................
λmA118:   ................................................
λmA119:   T...............................................A
                  100                               110
          laProGluGluHisProValLeuLeuThrGluAlaProLeuAsnProLys
cDNA:     CTCCTGAGGAGCACCCGGTGCTTCTGACCGAGGCCCCCCTGAACCCCAAA     318
λmA19:    ..............T.................................
λmA118:   .C.................A..........................
λmA119:   .C................................................
```

continued overleaf. . . .

continued. . . .

```
                                120                                 130
          AlaAsnArgGluLysMetThrGlnIleMetPheGluThrPheAsnThrPr
cDNA:     GCTAACAGAGAGAAGATGACGCAGATAATGTTTGAAACCTTCAATACCCC    368
λmA19:    ..................................................
λmA118:   ..................................C.................
λmA119:   ..................................................
                                    140
          oAlaMetTyrValAlaIleGlnAlaValLeuSerLeuTyrAlaSerGlyA
cDNA:     AGCCATGTACGTGGCCATTCAGGCGGTGCTGTCCTTGTATGCATCTGGGC    418
λmA19:    .........A..............AA..................A......
λmA118:   ....T.......CA.........T......C...........T........
λmA119:   ..................................................T
                  150                               160
          rgThrThrGlyIleValMetAspSerGlyAspGlyValThrHisThrVal
cDNA:     GCACCACTGGCATTGTCATGGACTCTGGTGACGGGGTCACACACACAGTG    468
λmA19:    ..................................................
λmA118:   ..................................................
λmA119:   ..................................................
                                170                                 180
          ProIleTyrGluGlyTyrAlaLeuProHisAlaIleLeuArgLeuAspLe
cDNA:     CCCATCTATGAGGGCTACGCCCTTCCCCACGCCATCTTGCGTCTGGACCT    518
λmA19:    ..................................................
λmA118:   G..........CA.....A.......T............T........T.
λmA119:   ............................T...G.................
                                    190
          uAlaGlyArgAspLeuThrAspTyrLeuMetLysIleLeuThrGluArgG
cDNA:     GGCTGGCCGGGACCTGACAGACTACCTCATGAAGATCCTGACTGAACGGG    568
λmA19:    ........C..............G..............C...........
λmA118:   ..T....TA............G...........TTC.T............
λmA119:   ......T....T..............................------...
                  200                               210
          lyTyrSerPheThrThrThrAlaGluArgGluIleValArgAspIleLys
cDNA:     GCTACAGCTTTACCACCACTGCTGAGAGGGAAATTGTTCGTGACATAAAG    618
λmA19:    ..................................................
λmA118:   ..................................----.....---...
λmA119:   ......................A................A.........
                                220                                 230
          GluLysLeuCysTyrValAlaLeuAspPheGluGlnGluMetAlaThrAl
cDNA:     GAGAAGCTGTGCTATGTTGCCCTGGATTTTGAGCAAGAAATGGCTACTGC    668
λmA19:    ..................................................
λmA118:   ....................................................A.........
λmA119:   ......................T...............................
                        234a                240
          aAlaSerSerSerSerLeuGluLysSerTyrGluLeuProAspGlyGlnV
cDNA:     TGCATCATCTTCCTCCTTGGAGAAGAGTTACGAGCTGCCCGACGGGCAGG    718
λmA19:    ............................T...........T.......
λmA118:   ..............................C..........T.......
λmA119:   .A..................................................T.......
```

continued overleaf. . . .

continued. . . .

```
                        250                                     260
            alIleThrIleGlyAsnGluArgPheArgCysProGluAlaLeuPheGln
cDNA:       TGATCACCATTGGCAATGAGCGGTTCCGGTGTCCGGAGGCACTCTTCCAG    768
λmA19:      ..........C.....................................
λmA118:     ..........C.....................................
λmA119:     .T........C.....................A...............
                              270
            ProSerPheLeuGlyMetGluSerCysGlyIleHisGluThrThrPheAs
cDNA:       CCTTCCTTCCTGGGCATGGAGTCCTGTGGTATCCATGAGACCACTTTCAA    818
λmA19:      ..........C.....................................
λmA118:     .A.................A........C...T.C........C.....
λmA119:     A.................................C....T...C.....
             280                                290
            nSerIleMetLysCysAspValAspIleArgLysAspLeuTyrAlaAsnT
cDNA:       CTCCATCATGAAGTGTGATGTGGATATCCGCAAAGACCTGTATGCCAATA    868
λmA19:      ..........C.....................................
λmA118:     ..........C.....................................
λmA119:     ..........C.....................................
                        300                                     310
            hrValLeuSerGlyGlyThrThrMetTyrProGlyIleAlaAspArgMet
cDNA:       CAGTGCTGTCTGGTGGTACCACCATGTACCCAGGCATTGCTGACAGGATG    918
λmA19:      ..........C.....................................
λmA118:     ..........C.....................................
λmA119:     ..........C.....................................
                              320
            GlnLysGluIleThrAlaLeuAlaProSerThrMetLysIleLysIleIl
cDNA:       CAGAAGGAGATCACAGCCCTAGCACCTAGCACGATGAAGATTAAGATCAT    968
λmA19:      ..........C.....................C.....A..........
λmA118:     T.........C.....................C.....A..........
λmA119:     A.....................A.........C.....A.C........
             330                                340
            eAlaProProGluArgLysTyrSerValTrpIleGlyGlySerIleLeuA
cDNA:       TGCTCCCCCTGAGCGCAAGTACTCAGTCTGGATCGGTGGCTCCATTCTGG   1018
λmA19:      ..........C.....................................
λmA118:     ..........C.....................C.T.C........C..-A
λmA119:     ..........C.....................T..T............
                        350                                     360
            laSerLeuSerThrPheGlnGlnMetTrpIleSerLysGlnGluTyrAsp
cDNA:       CCTCACTGTCCACCTTCCAGCAGATGTGGATCAGCAAGCAGGAGTATGAT   1068
λmA19:      ..........C.....................................
λmA118:     ..........C.....................................
λmA119:     ..........C.....................................
                              370         374
            GluSerGlyProSerIleValHisArgLysCysPheEnd
cDNA:       GAGTCAGGCCCCTCCATCGTCCACCGCAAATGCTTCTAGATGGACTGAGC   1118
λmA19:      .....G..........................................
λmA118:     .....G...----.....................//////////
λmA119:     ....TG........T...A.....AT..................C.G..
```

continued overleaf. . . .

continued. . . .

```
cDNA:      AGGTGCCAGGCATCTGCTGCATGAGCTGATATTGAAGTATCAATTTGCCC   1168
λmA19:     .....................................G........
λmA118:    //////////////////////////////////////////////////
λmA119:    ........A........C.....................TG........

cDNA:      TGGCAAATGTACACACCTCATGCTAGCCTCATGAAACTGGAATAAGCCTT   1218
λmA19:     ...........T.....................................
λmA118:    //////////////////////////////////////////////////
λmA119:    ..............................................TCCC

cDNA:      TGAAAAGAAATTTAGTCCTTGAAGCTTGTATCTGATATCAGCACTGGATC   1268
λmA19:     ............-.....................................
λmA118:    //////////////////////////////////////////////////
λmA119:    CCCCCCCTTTCC.TT.TA..TTTTA..CAC.TAACATC.CAG..ACAGC.

cDNA:      GTAGAACTTGTTGCTGATTTTTGACCTTGTATTCAAGTTAACTGCTCCCT   1318
λmA19:     .................................................
λmA118:    //////////////////////////////////////////////////
λmA119:    CCC.CC.C.T..CAGAG..CC.CCTT.ACA.G.TCCT.CC..CAT.....

cDNA:      TGGTATATGTTTAATACCCTGTGCATATCTTGATTTCTCCTTAGTTCATG   1368
λmA19:     .................................................
λmA118:    //////////////////////////////////////////////////
λmA119:    .TTCTCT.TG.CTC.TAGAAA..GGACC.G.TTG.GTA..ACTTAC.

cDNA:      TGGCTCGGTCACTTGGGGCTGGGGAGAGCACGCTGTAGATGAGAAAGCCC   1418
λmA19:     ....................A.............G.GC......A...
λmA118:    //////////////////////////////////////////////////

cDNA:      CAGCCTGGTTGATCTCTGTGAGCACCACTGAGTGATCTGTGCAGGGTATT   1468
λmA19:     ........................T..G..............A....
λmA118:    //////////////////////////////////////////////////
```

The nucleotide sequences that correspond to the predicted γ-like-actin in clones λmA19 (Leader *etal.*, 1985), λmA118, and λmA119 are aligned to the corresponding regions of the partial sequence of mouse γ-actin cDNA (Peter and Leader, personal communication). Dots below the cDNA sequence indicate identity, and hyphens indicate deletions in the sequence. Strokes below the cDNA sequence indicate sequence not determined. The positions of insertions are not shown.

(The calculation of the percentage of possible neutral mutations in the γ-actin gene sequence, was according to Leader *et al.*, 1986b). Correction for this fact gives a value of 1.13 for the UEP. Therefore assuming neutral drift for the actin-like sequences in clone λmA118 and λmA119 since their formation from the active γ-actin gene, it can be estimated that this event occurred approximately 6.8 million years ago in λmA118 (6.0% with a UEP of 1.13) and 4.4 million years ago in λmA119 (3.9% with a UEP of 1.13). The times of formation of these actin processed pseudogenes are relatively recent compared with the divergence of the two lines leading to rat and mouse, which is assumed to have occurred 15 million years ago (Alonso *et al.*, 1986). The evolutionary time of formation of the actin-like sequence in clone λmA118 is significantly longer ago than that of clone λmA119, and that of clone λmA19 (1.9 million years; 1.7% divergence with a UEP of 1.13).

The above calculation assumes a similar rate for the accumulation of mutations in neutral positions in a functional gene and pseudogene. However the comparison shown in Figure 5.3 allows one to identify the mutations in the functional gene since the divergence of the younger pseudogenes, λmA119 and λmA19. Differences from the cDNA common to all three pseudogenes are statistically most likely to have occurred in the functional gene since the origin of λmA19, and differences common only to λmA118 and λmA119 are more likely to have occurred in the functional gene since the origin of λmA119 but before the origin of λmA19. There are 6 in total in the former case and a total of 2 in the latter case. All these changes have taken place in silent positions of the cDNA sequence. The occurrence of 6 functional gene changes out of a total of 19 base changes for the comparison with λmA19, and 8 functional gene changes out of a total of 43 base changes in λmA119 are approximately consistent with the proportion of total neutral positions (24.4%) found in the cDNA sequence. This is in contrast with the results of others who found that globin pseudogenes evolve faster at neutral positions than do functional genes (Miyata and Yasunaga, 1981; Miyata and Hayashida, 1981; Li *et al.*, 1981).

In order to determine whether the actin-like sequences in λmA118 and λmA119 have evolved at a neutral rate, as assumed for the calculations above, the R/S (replacement changes / silent changes) ratios were calculated. The replacement changes in a functional coding sequence are more likely to be detrimental, and therefore be selected against, than silent changes. As a consequence, the R/S ratio allows one to discriminate between functional genes and pseudogenes. Pseudogenes are in general expected to have 2.5 to 3.0 times

as many R as S changes because in this case these are not detrimental (Czelusniak *et al.*, 1982). The R/S ratio of the γ-actin-like genes in λmA118 and λmA119 were determined to be 1.6 and 1.15 respectively, both ratios seeming inappropriate for processed pseudogenes evolving under neutral selection. Because the γ-actin amino-acid sequence is totally conserved (*ie* no R changes in the functional gene) the values of Czelusniak *et al* (1982) are too high. Having identified the changes in the functional gene occurring since λmA119 originated it is possible to eliminate these and identify the R/S changes in the pseudogene itself where the 24.4% of silent positions predicts an R/S ratio of 3.0 for a pseudogene. In fact the corrected number of 23 replacement changes and 12 silent changes gives a R/S ratio of 1.9 for λmA119 (a similar value of 2.0 is obtained for λmA19). With 6.8 million years of divergence from the functional gene, λmA118 is the 'oldest' mouse γ-actin pseudogene identified to date. Therefore it is not possible to estimate the number of changes the functional gene had acquired since the formation of λmA118. However, eliminating changes that the functional gene had acquired since the origin of λmA119, the corrected values of 41 replacement changes to 18 silent changes give rise to a R/S ratio of 2.3. This might suggest that during part of its existence, this gene was evolving under a selective pressure for a protein coding sequence. However, this seems unlikely as most processed pseudogenes are thought to become inactive as soon as they are inserted into the genome. Therefore other factors were sought as possible explanations for these deviations of the R/S ratio from that expected.

Bulmer (1986) analysed mutations in a number of processed pseudogenes and pointed out that the frequency of transition in CG doublets in vertebrates is ten times that expected on a random basis. This can be attributed to the high frequency of methylated cytosine in this doublet, and the correspondingly high level of deamination of this 5-methyl cytosine (mC) into thymine (Coulondre *et al.*, 1978; Bird, 1980; Razin and Riggs, 1980). Thus the doublet CG is converted via mCG to TG and its complement CA. The reduction in frequency of occurrence of the doublet CG is indeed often accompanied by corresponding increases in the doublets TG and CA (Setlow, 1976; Russell *et al.*, 1976). Examination of the functional γ-actin gene (cDNA) in Figure 5.3 allowed the identification of 40 such CG doublets with the possibility of methylation and then deamination into the doublets TG and CA. In λmA119, there have been 9 of these changes, where 4 were replacement changes and 5 silent changes. If one chooses to discount the CG transitions, one can obtain a corrected R/S ratio for

λmA119 of 2.7 (with 19 replacement changes to 7 silent changes). In λmA118, a total of 10 base changes can be accounted for by deamination of CG doublets, consisting of 5 replacement and 5 silent changes, and allowing a corrected R/S ratio of 2.8 (with 36 replacement changes to 13 silent changes). Thus it would seem most likely that the actin-like genes in λmA118 and λmA119 have been inactive since their origin, the bias of CG to TG transitions in the silent position accounting for the deviation of R/S ratio from that expected for a pseudogene. The explanation for this bias is, however, unclear.

5.2 Insertion Elements IE 36, IE 119 and IE 118

5.2.1 Analysis of IE 36

The nucleotide sequence of IE 36 indicated that it is related to a solo long terminal repeat (LTR) of the retroviral-like mouse intracisternal A-particle (IAP), as illustrated by the sequence comparison shown in Figure 5.4a. A duplication of the target-site (6 base pairs of actin DNA coding for Pro^{70} and Ile^{71}) flanking IE 36 indicates that IE 36 arose following an insertion at a staggered break in the actin processed-pseudogene. This mode of insertion is typical of retroviruses and transposable elements and is also consistent with other IAP insertions which generate 6 base-pair target-site duplications. As the LTRs of IAPs do not themselves contain the genetic information for retrotransposition, it is assumed that the original insertion was of a complete IAP gene, the LTRs of which subsequently underwent unequal crossing-over.

The solo IAP LTR of IE 36 is 500 base pairs in length. This includes a duplicated region, not found in other IAPs and discussed below. Ignoring this duplication, in comparison with the most related IAP LTR nucleotide sequence in the EMBL and GenBank databases, 5'rc-mos (Canaani *et al.*, 1983), IE 36 shows a sequence homology of 87% (Figure 5.4a). Several conserved sequence motifs which are essential for transcriptional regulation have been used to subdivide LTRs into three functional domains, U3-R-U5 (Temin, 1981). The sequence CCAAT (CAAT box) usually occurs in the U3 region 75 base pairs 5' to R ; the sequence $G/_C T/_A A T T/_A T/_A AAG$ (Goldberg-Hogness box, TATA) usually occurs 23 base pairs before R. These sequences are thought to be important for transcriptional promotion (Breathnach and Chambon, 1981). The R region always starts at the capping nucleotide, G, and ends with the polyA addition site,

continued . . . .

Figure 5.4b Comparision of members of 46 base pair repeat in IE 36

```
228   CTCTTGCTC----------------CTAAAGATGTAGGCAATAGAG   257
                                          *
258   CTCTTGCTCTCTTGCGCTCTGGCTCCTAAAGATGTAAGCAATAGAG   303
                    *     *
304   CTCTTGCTCTCTTGCTCTCTTGCTCCTAAAGATGTAAGCAATAGAG   349
                    *     *
350   CTCTTGCTCTCTTGCGCTCTGGCTCCTAAAGATGTAAGCAATAGAG   395
                    *     *       *             *   *
396   CTCTTGCTCTCTTGCACTCTTGCTCCTGAAGATGTAAGCTATAAAG   441
```

(a) Comparision of IE 36 with the related IAP LTR, 5'rc-mos (Canaani *et al.*, 1983) is shown. Vertical lines between sequences indicate identity, target site direct repeats are boxed, and sequences of possible functional importance are underlined. Gaps introduced to optimise alignment are indicated by hyphens. The amino-acid equivalents of the opposite strand of the flanking nucleotides of the actin processed pseudogene are indicated, and numbered for the protein sequence (Vandekerckhove and Weber, 1979a). The LTR putative regions U3, R and U5 are indicated. The numbering of the sequence is only for the inserted element. (b) Comparision of the members of the 46 base-pair repeat in IE 36 is shown. Asterisks indicate the position at which differences occur.

CA. Twenty base pairs to the start of the U5 region, there is the polyadenylation signal sequence, AATAAA. In U5, 10 to 25 base pairs after the end of R, there is TTGT or some closely related sequence, which is thought to be important in termination of viral RNA synthesis. Some retroviral LTRs have been found to contain core enhancer sequences ($TGG^{T}/_{A}{}^{T}/_{A}{}^{T}/_{A}$), which are thought to be involved in activation of RNA transcription of nearby genes (Weihler *et al.*, 1983). The enhancer sequence appears to be part of the LTRs of all IAPs , and is usually associated with a potential Z-DNA forming sequence, which is located 3' to the core enhancer. A consensus glucocorticoid recognition sequence (TGTTCT), first detected in mouse mammary tumour virus and the LTRs of an IAP associated with the mouse renin gene (Burt *et al.*, 1984; Scheidereit and Beato, 1984), is also a feature of some retrovirus LTRs. This segment is usually located 5' to the core enhancer.

With reference to the sequence homology of IE 36 to rc-mos, and data obtained from other IAP LTRs (Christy *et al.*, 1985), the boundaries between the putative U3, R, and U5 domains within the IAP LTR have been indicated in Figure 5.4a, as well as several of the functionally important sequence motifs found in some, although not all, retroviral LTRs (Temin, 1981; Varmus, 1982) . The solo IAP LTR in IE 36 contains : (a) an imperfect version of the 4 base pair terminal inverted repeat, 5'TGTG/TAGA3', which is unusual since most retroviral-like LTRs have been found to be flanked by perfect inverted repeats, 5'TGTT/AACA3' (Ono and Ohishi, 1983). However, imperfect inverted repeats have been reported in IAP LTRs characterised by Christy *et al.* (1985) ; (b) a glucocorticoid recognition sequence (TGTTCT at position 52) ; (c) a core enhancer (TGGTAA at position 61) ; (d) a potential stretch of Z-DNA consisting of 11 base pairs in reasonably close proximity to the enhancer (at position 77) ; (e) a CAT box (CCAAT at position 155) ; (f) a potential TATA box with 5 out of 9 matches to the consensus (CGAGAATAA at position 182) ; (g) a potential polyadenylation / processing signal with 5 out of 6 matches to the consensus (TATAAA at position 435) ; and (h) a potential transcriptional termination signal with 3 out of 4 matches to the consensus (TGCT at position 465). It appears that the solo IAP LTR in IE 36 contains similarly arranged putative nucleotide sequences for promotion, initiation, polyadenylation, and termination of viral RNA transcription to those of other retroviral-like LTRs (Temin, 1981). Of course there is no evidence that these sequences in the LTR ever function as transcriptional promoter and polyadenylation signals in λmA36.

As already mentioned, it can be seen from Figure 5.4a that IE 36 contains an additional region of DNA, 165 nucleotides in length. This appears to have arisen by successive duplication of the 46 nucleotide region represented by nucleotides 396 to 441 in Figure 5.4a, generating five copies, one of which subsequently underwent an internal deletion (Figure 5.4b). Sequence differences due to single base changes between four of the five repeats indicate that these were not generated during cloning. The lack of even short homologous regions flanking the prototype of the repeat makes it unclear whether the initial duplication was a result of non-homologous recombination, or slippage during replication. It should be remarked that other forms of expansion and rearrangements have been observed in the LTRs of some other IAPs. The LTR regions of mouse IAP genes have been found to be very heterogeneous (Christy *et al.*, 1985). However, this particular heterogeneity is confined to the R region of the LTR, IAP genes with longer R regions having been shown to consist of several repeated oligonucleotide stretches which are absent in LTRs with shorter R regions. IE 36 also conforms to this pattern, with the repeated stretch of nucleotides (although different from the others described previously) falling into the putative R region.

Although solo LTRs have not previously been reported for IAPs, they have been described for other mouse retrovirus-like elements (Wirth *et al.*, 1983). The ratio of full-sized elements to their solo LTR counterparts is highly variable in different cases. For example, in the family of murine retrovirus-related sequences (MuRRS), the 5.7 kb long elements with their corresponding LTRs are repeated 50 to 100 times in the mouse haploid genome, whereas their identical solitary LTR-like counterparts are found to repeated 500 to 1000 times in the mouse haploid genome (Schmidt *et al.*, 1985). In contrast, in VL30 DNAs there are at least twenty times more full-sized elements than solo LTRs (Rotman *et al.*, 1984). However, the factors that determine this variability are not known.

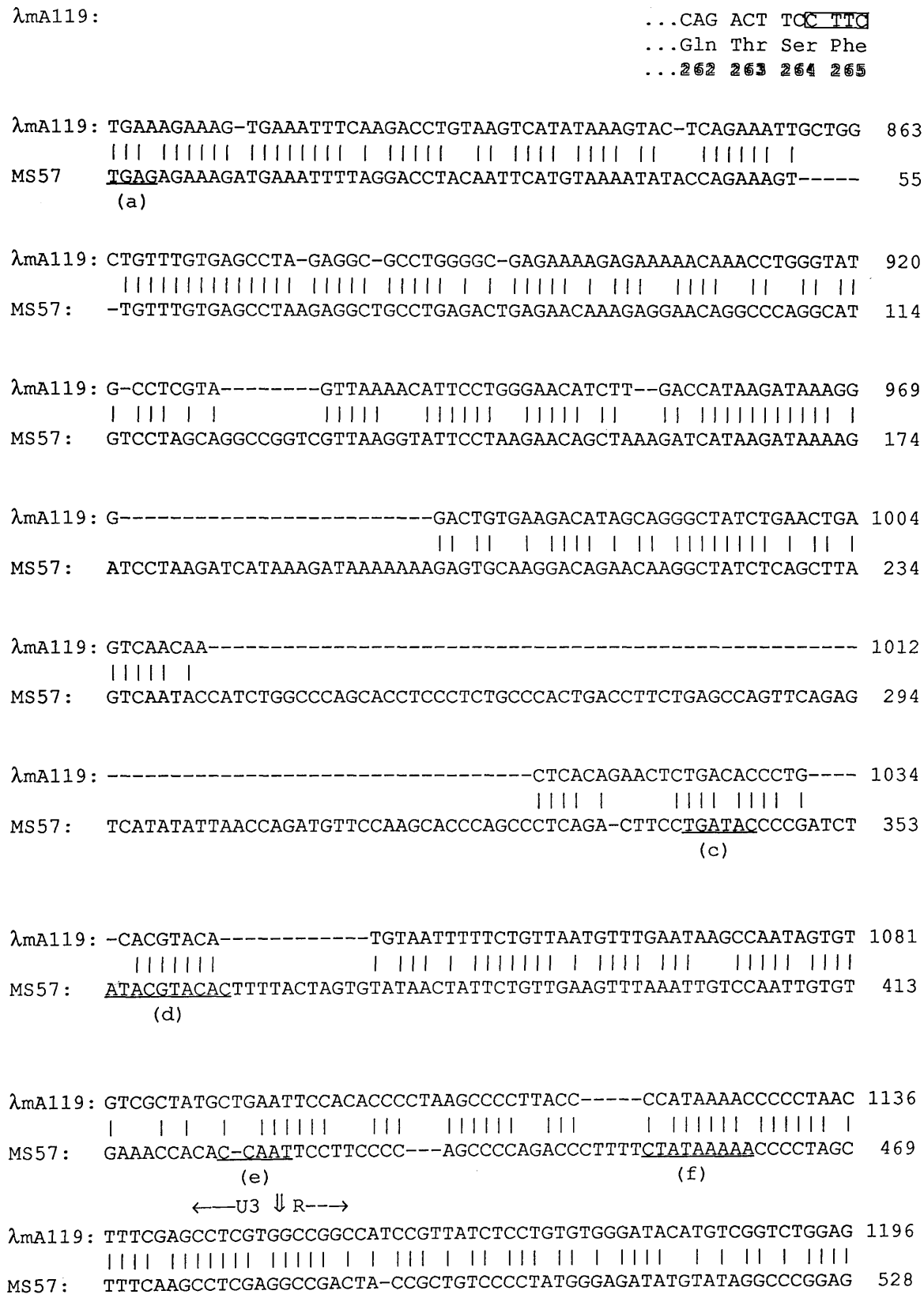
Hybridisation of a probe obtained from IE 36 to Southern blots of mouse liver DNA (Figure 4.11) revealed numerous hybridising bands, indicating that the IAP LTR is highly repeated in the mouse genome. It has been estimated (see section 4.3.4) that there are approximately 1,900 copies of IE 36 per mouse haploid genome, which is consistent with the 1,000 to 2,000 integrated copies of IAP genes per mouse haploid genome estimated by others (Kuff, *et al.*, 1983a).

5.2.2 Analysis of IE 119

Sequence analysis of λmA119 showed that IE 119 was flanked by a four base pair duplication, which is part of the actin cDNA coding for Ser^{264} and Phe^{265}. As before, this suggests that IE 119 was integrated at a staggered break into the actin processed-pseudogene. The structure of IE 119 (501 nucleotides in length) is related to a previously reported retroviral-like LTR element, MS57 (Propst and Vande Wonde, 1984), as illustrated by the sequence comparison shown in Figure 5.5. The methods used for distinguishing the putative U3, R, and U5 regions in LTRs (outlined in section 5.2.1) were adopted for the solo LTR in IE 119; and it was found to contain : (a) a 7 base pair imperfect inverted repeat, 5'TGAAAGA/TCTTACA3', flanking the LTR nucleotide sequence (at positions 805 and 1300) ; (c) a potential core enhancer with 4 out of 6 matches to the consensus (TGATAC at position 1025) ; (d) an 8 base pair stretch of Z-DNA which is in close proximity with the enhancer (at position 1035) ; (e) a potential CAT box with 3 out of 5 matches to the consensus (CTGAAT at position 1091) ; (f) a potential TATA box with 7 out of 9 matches to the consensus (CCATAAAAA at position 1120) ; (g) a polyadenylation/processing signal (AATAAA at position 1204) ; and (h) a potential transcriptional termination signal with 3 out of 4 matches to the consensus (TTCT at position 1247). It is clear that IE 119 is a solo LTR, despite the absence of the glucocorticoid recognition sequence, which is not always present in all retroviral-like LTRs (Scheidereit and Beato, 1984).

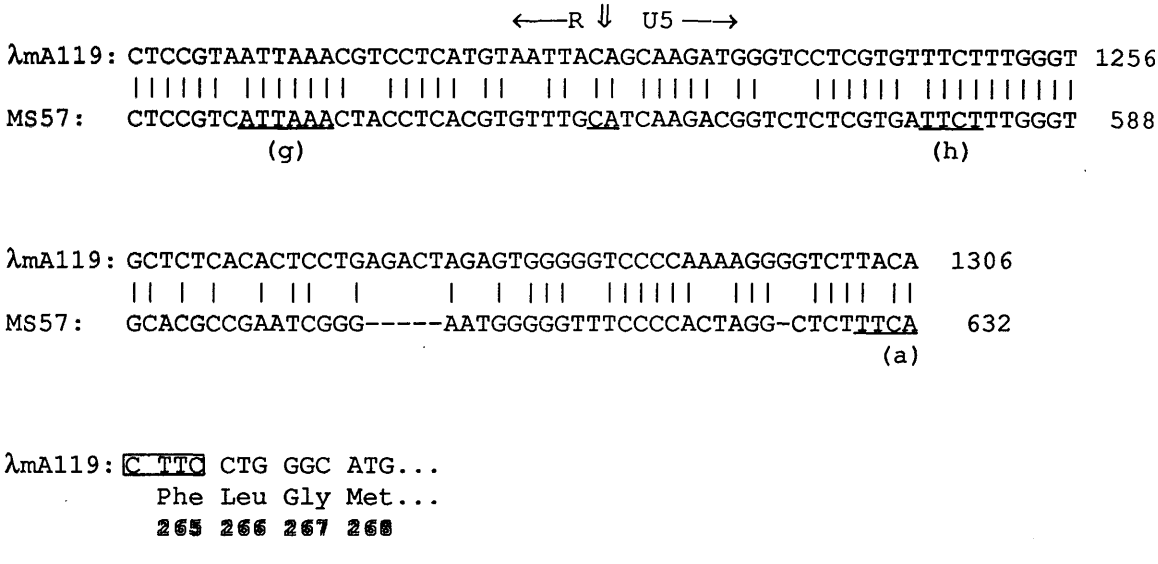
Figure 5.5 shows that considerable homology exists between IE 119 and MS57, although large gaps need to be introduced to allow sequence alignment. If one scores each gap (or deletion) and each base change as a single mutational event, the alignment indicates a sequence homology of 70% with the LTR element, MS57. The large gaps of DNA missing from the putative U3 region of the IE 119 (totalling approximately 120 nucleotides), may be due either to several insertions in the MS57 LTR or, alternatively, deletions in the IE 119 sequence. A full-length retroviral-like gene, GLN-3, has recently been shown to have an LTR related to MS57 (Itin and Keshet, 1986). This is only 430 base pairs in length (compared with 501 for IE 119) and contains none of the long sequences present in MS57 but absent from IE 119. Thus these latter most likely arose by insertion or expansion; and this suggests that in the GLN-3 LTR it is the U3 region that is particularly susceptible to alteration, rather than the R region, as was the case for the IAP LTRs. It is worth recalling that the variation

Figure 5.5 Comparision of IE 119 with related retroviral-like LTR



continued overleaf....

continued. . . .



Comparision of IE 119 with the related retroviral-like LTR, MS57 (Propst and Vande Woude, 1984) is shown. Vertical lines between sequences indicate identity, target site direct repeats are boxed, and sequences of possible functional importance are underlined. Gaps introduced to optimise alignment are indicated by hyphens. The amino-acid equivalents of the flanking nucleotides of the actin processed pseudogene are indicated, and numbered for the protein sequence (Vandekerckhove and Weber, 1979a). The numbering of the sequence is as in Figure 3.10. The LTR putative regions U3, R and U5 are indicated.

in size of the LTRs of different retroviruses generally reflects variations in the size of their U3 regions (Temin, 1981).

It has been estimated that there are approximately 2,300 copies of IE 119 per mouse haploid genome (see section 4.2.4), which conforms with the figures obtained for solo retroviral-like LTRs from MS57 and GLN-3 (Propst and Vande Wonde, 1984; Itin and Keshet, 1986).

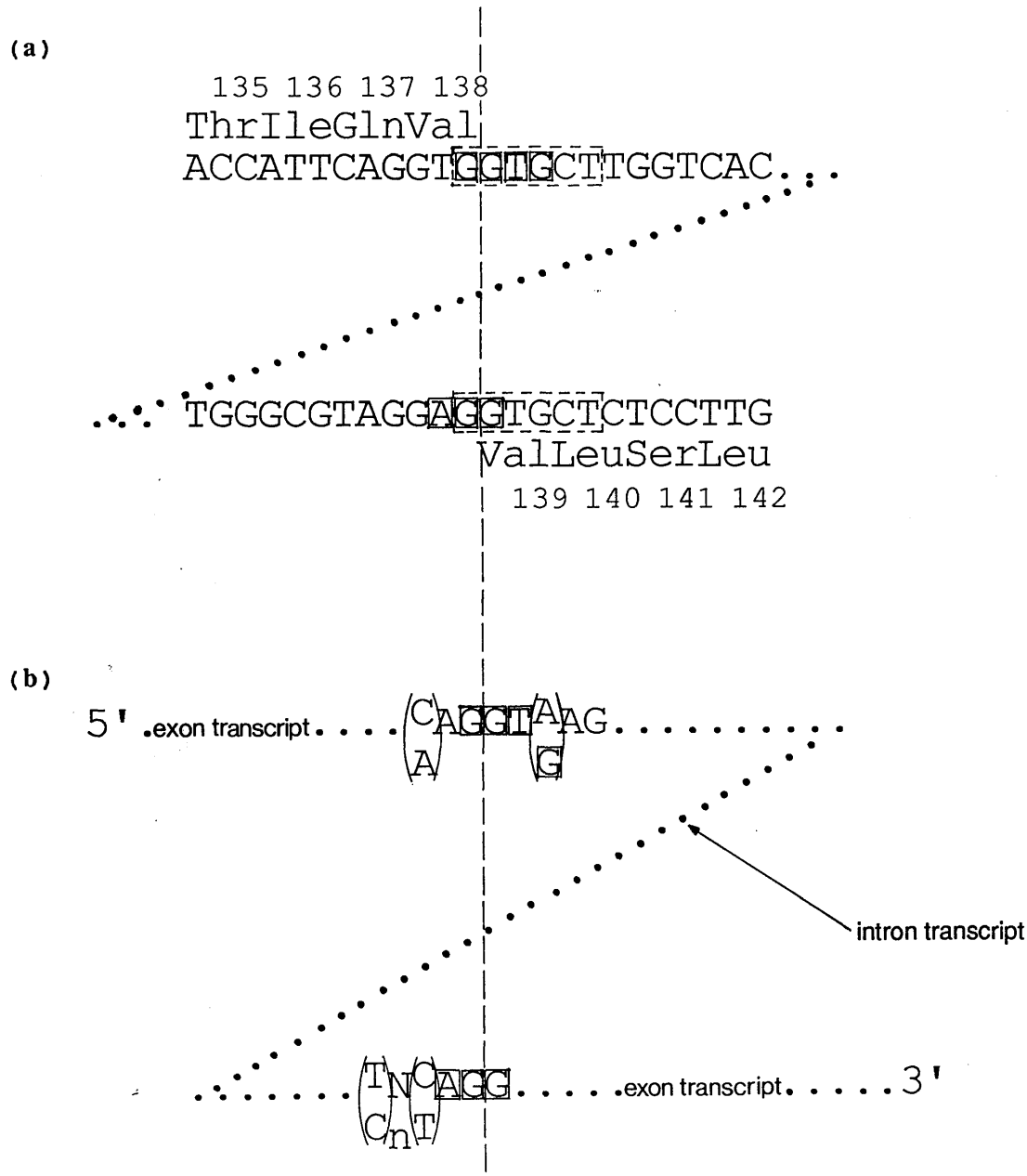
5.2.3 Analysis of IE 118

(i) Relationship to Introns

It is unlikely that the sequences interrupting the pseudogenes and designated as inserted elements (IE) represent introns for several reasons other than the overwhelming evidence that IE 36 and IE 119 are retroviral-like LTRs. None of the inserted elements is situated at a predicted intron position for γ-actin. Table 1.2 shows the position of introns in the actin genes of various organisms. Although the gene for the mammalian γ-actin actin has yet to be identified, a number of conserved intron positions in different actin isoforms of vertebrates, namely 41, 267 and 327, have been identified. From the sequencing data, it can be seen that the three inserted elements IE 118, IE 119 and IE 36, do not occupy any of these positions. The data also show that they are flanked by short direct repeats, which are not a feature of introns but are a common feature of the integrated state of transposable elements and retroviral-like elements. Figure 5.6a illustrates the flanking direct repeats of IE 118 at Val^{139} and Leu^{140}, the 6 base pair GGTGCT. Thus the proposed mode of insertion of IE 118 is via a staggered break in the mouse processed pseudogene in this region, resulting in a target site duplication.

When the consensus sequence of the intron-exon splice sites were compared with the boundaries of IE 36 and IE 119, little similarity was evident. However, in the case of IE 118 surprising similarity of the flanking regions to the intron/exon consensus were observed (Figure 5.6b). The sequences of exon-intron boundaries (the splice site at the 5' end of the intron which is also known as the donor site) and intron-exon boundaries (the splice site at the 3' end of the intron which is also known as the acceptor site) are very well conserved, so that deviations are small, if any (Mount, 1982). The most invariant aspect of the consensus sequence is the GT doublet at the beginning of the intron transcript and AG at the end of the transcript (the so-called GT/AG

Figure 5.6 Flanking direct repeat of IE 118 and comparison with the intron splice site consensus sequence at position Val^{138}



The flanking regions of IE 118 are shown in (a), the predicted 6 base pair target site direct repeat being boxed with broken lines. Numbering of amino-acid is as in Figure 3.11. The intron splice site consensus sequence is shown in (b). The splice point of the intron and the exon boundary is indicated by the vertical hyphenated line, and this is extended to IE 118 at position Val^{138}. Homologous nucleotides between the consensus sequence and those of the flanking regions of IE 118 at the same positions are outlined by solid boxes.

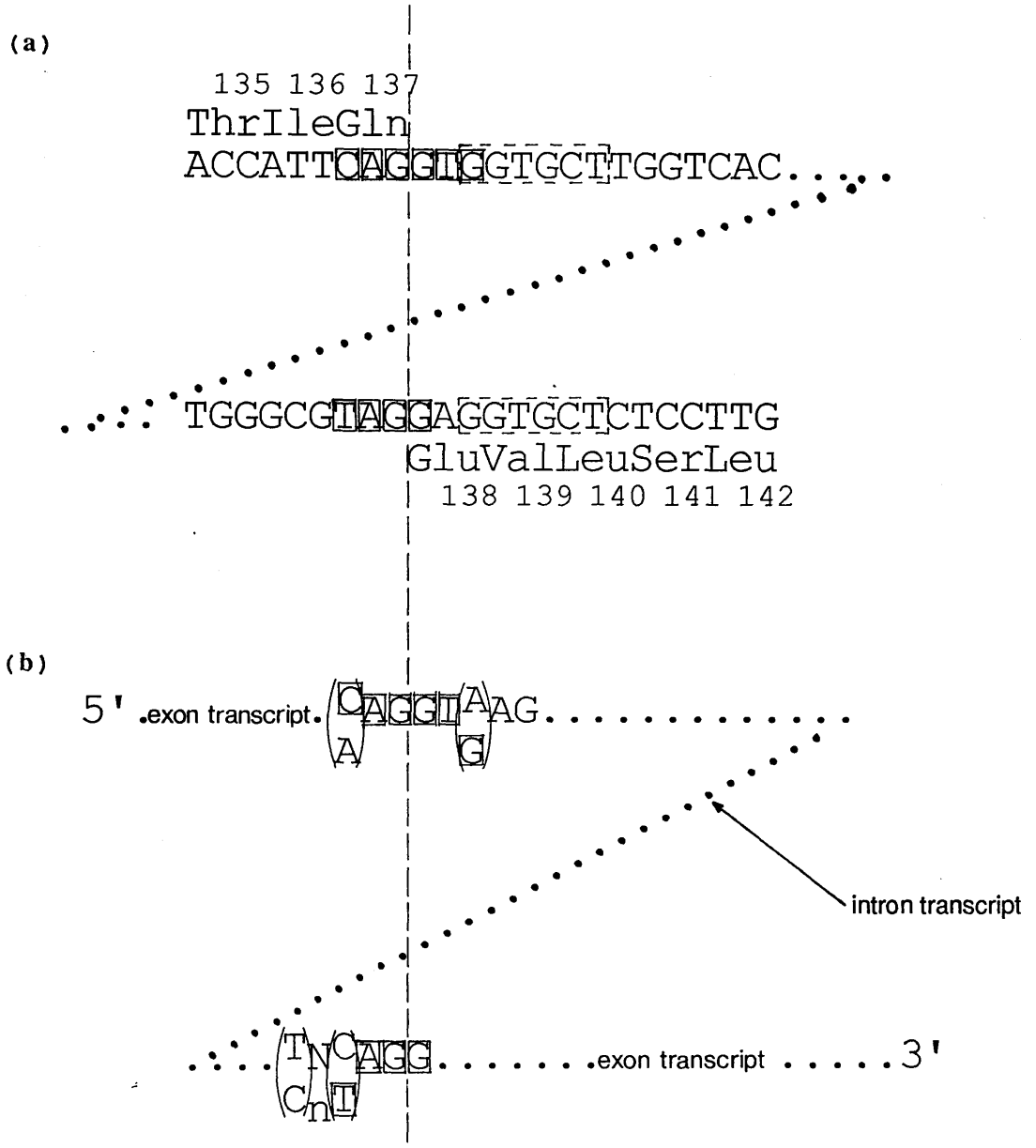
rule when applied to the genomic DNA). When the consensus sequence is aligned with the flanking regions of IE 118 close to the direct repeat, two regions of distinct sequence homology can be seen (Figure 5.6a). At position Val^{138} there are 4 out of 8 nucleotides which match the donor consensus, $C/_A AGGTA/_G AG$, (GTGGTGCT at position 435 to 442), and 4 nucleotides matching the last 5 of the acceptor consensus, $(T/_C)_n N C/_T AGG$, (AGGAGG at position 1308 to 1313). Furthermore, when the alignment is shifted 3 base pairs upstream to position Gln^{137} (Figure 5.7a), 6 out of 8 nucleotides match the donor consensus (CAGGTGGT at position 432 to 439), and the last 5 nucleotides of the acceptor consensus are completely matched (CGTAGG at position 1305 to 1310). (This creates Glu at position 138 instead of the previous Val. However, this is equally acceptable as Ala is the amino-acid actually at this position in the γ-actin sequence.) Nevertheless the similarity to intron/exon splice sites breaks down at the sequence immediately before the intron-exon junction. The consensus sequence here is always pyrimidine-rich and devoid of the dinucleotide AG, and this is not the case in either position of alignment of IE 118.

Although neither position Gln^{137} and Val^{138} corresponds to a known intron position in the actins of vertebrates (Table 1.2), there are actin genes which have their own unique intron positions. For example, the human smooth muscle actin (aorta) contains an intron at codon position 84, which is not found in others (Table 1.2; Ueyama *et al.*, 1984). For reasons discussed below, it is considered likely that IE 118 is the remenant of a retroposon rather than a residual intron in a processed pseudogene of the type apparently found in the preproinsulin I gene (Soares *et al.*, 1985). However, this similarity to an intron raises the possibility that mammalian retroposons could, by chance, give rise to new introns. Although the dominant view is that introns evolved early in evolution and are gradually being eliminated (Zakut *et al.*, 1982; Blake *et al.*, 1983; Nudel *et al.*, 1984; Carroll *et al.*, 1986), Rogers (1985) has argued for the opposite possibility.

(ii) Possible Identity of IE 118

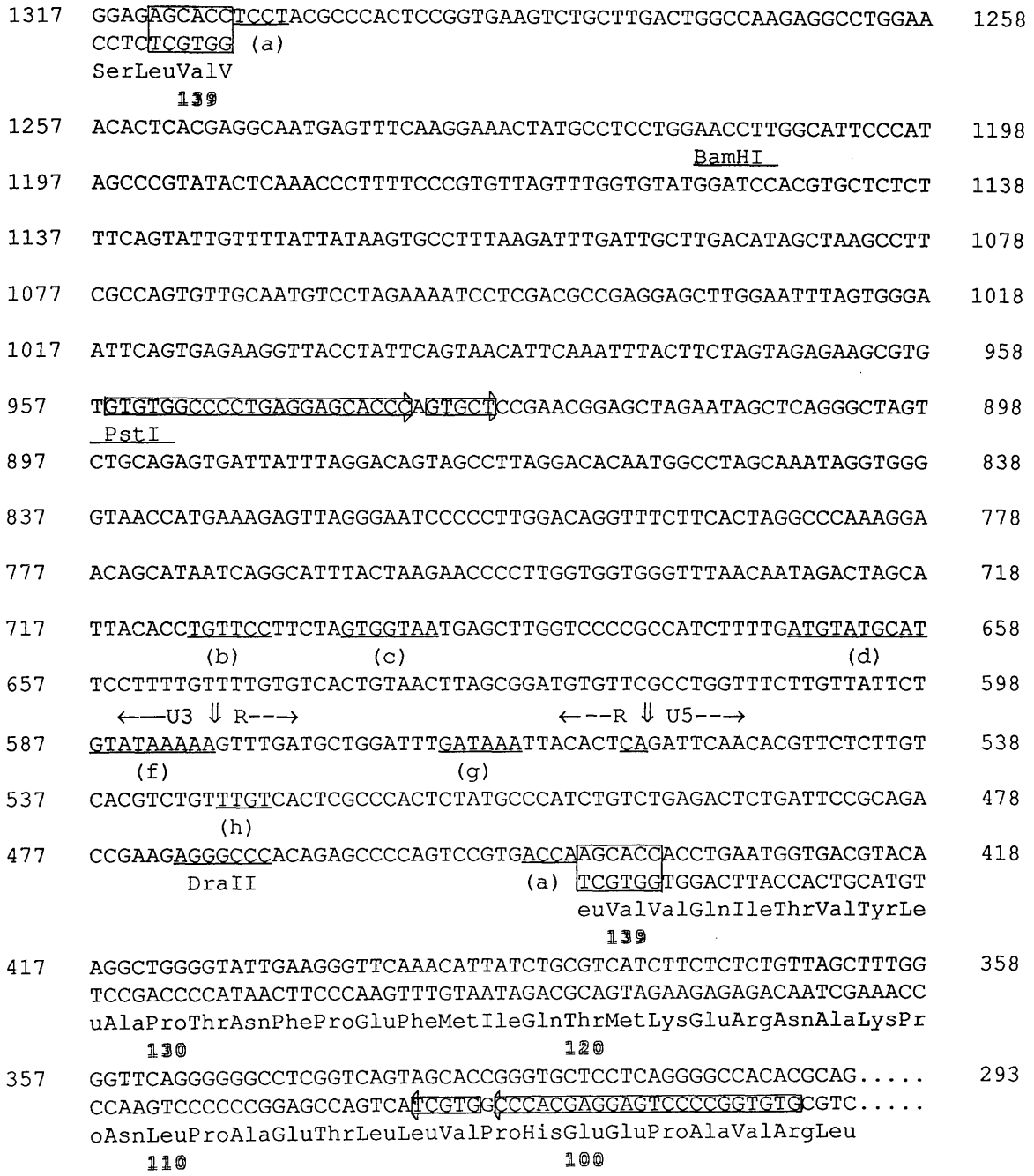
The sequence of IE 118 and part of the flanking actin pseudogene are shown in Figure 5.8. IE 118 is flanked by a 6 base pair duplication of actin sequence, indicating that it is inserted at a staggered break. Computer searches of GenBank (release 40) and EMBL (release 9) nucleotide sequence databanks have not revealed any sequence having extensive homology to IE 118 (see

Figure 5.7 Flanking direct repeat of IE 118 and comparison with the intron splice site consensus sequence at position Gln^{137}



The flanking regions of IE 118 are shown in (a), the predicted 6 base pair target site direct repeat being boxed by broken lines. Numbering of amino-acid is as in Figure 3.11. The intron splice site consensus sequence is shown in (b). The splice point of the intron and the exon boundary is indicated by the vertical hyphenated line, and this is extended to IE 118 at position Gln^{137}. Homologous nucleotides between the consensus sequence and those of the flanking regions of IE 118 at the same positions are outlined by solid boxes.

Figure 5.8 Nucleotide sequence of IE 118 and flanking regions



The nucleotide sequence of IE 118 and flanking regions is shown. Target site direct repeats are boxed and sequences of possible functional importance are underlined. The amino-acid equivalents of the opposite strand of the flanking nucleotides of the actin processed pseudogene are indicated, and numbered for the protein sequence (Vandekerckhove and Weber, 1979a). The numbering of the sequence is as in Figure 3.8. The LTR putative regions U3, R and U5 are indicated.

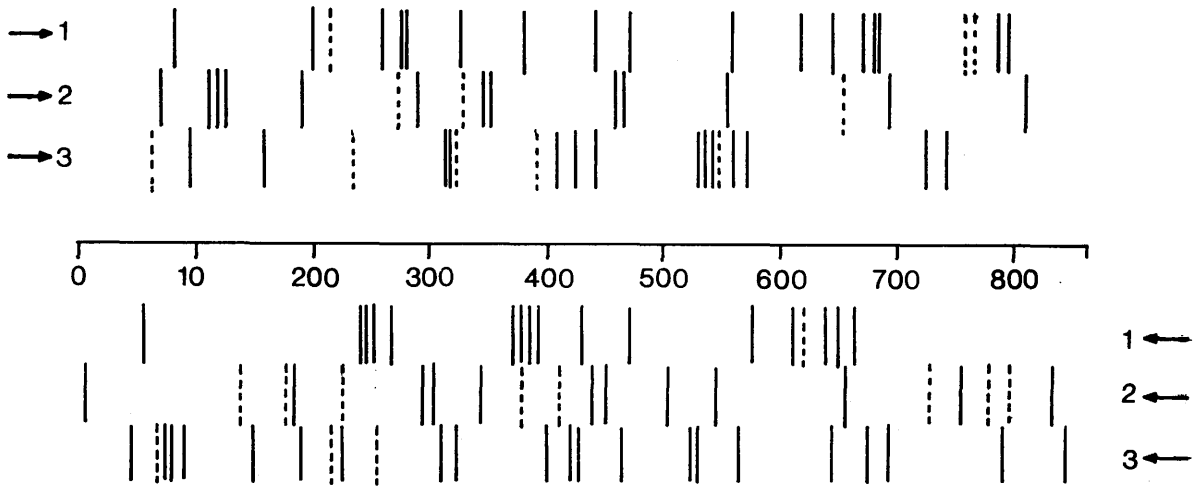
section 4.1.2). Thus it is not a known retroposon or part thereof.

A map of all the potential open reading frames on both strands of the DNA of IE 118 has been constructed (Figure 5.9). The distances between the AUG start codons and the stop codons are such that the longest possible open reading frame in IE 118 is 130 base pairs in length. This open reading frame could only code for a protein a little over 40 amino-acids long, which seems too small to be of any significance. Thus it is unlikely that IE 118 is a protein coding sequence.

Because IE 36 and IE 119 had been shown to be solo LTRs of murine retroviral-like elements the possibility was addressed that IE 118 is also a solo LTR of an endogenous mouse retrovirus or retroviral-like sequence not previously recognised. It can be seen from Figure 5.8 that it is possible to identify several of the functionally important sequence motifs of retroviral LTRs (Temin, 1981; Varmus, 1982), including the putative U3, R, and U5 domains within the LTR (compare also Figure 5.4a and Figure 5.5). Their status in IE 118 is as follows : (a) the terminal 4 base pair inverted repeat (5'TCCT/ACCA3') is imperfect, although the first two nucleotide pairs of the termini TC/CA (at position 1307 and 447) may be related to the inverted repeat TG/CA, which is found in most prokaryotic and eukaryotic transposable elements; (b) the glucocorticoid recognition sequence matches 5 out of 6 to the consensus (TGGTAA at position 736) ; (c) the core enhancer matches the consensus (TGGTAA at position 736) ; (d) a potential stretch of Z-DNA sequence consisting of 10 base pairs is found in reasonably close proximity to enhancer (at position 667) ; (g) the polyadenylation/ processing signal matches 5 out of 6 to the consensus followed after 8 base pairs by a potential CA site of poly A addition (GATAAA at position 562) ; (h) a transcription termination signal is present at a suitable region (TTGT at position 528). The lack of a perfect terminal inverted repeat need not argue against IE 118 being an LTR, as it has already been seen that imperfect inverted repeats are present in IE 36 (see section 5.2.1), and other IAP LTRs (Christy *et al.*, 1985). It may also be noted that the polyadenylation signal in IE 36 and the CAT box in IE 119 have suffered mutation. Thus it is tentatively concluded that IE 118 is most likely to be a new retroviral-like LTR.

Another feature of IE 118 deserves comment. This is a 28 nucleotide region (arrows in Figure 5.8) in which 27 nucleotides are identical to those in a 28 nucleotide region on the opposite strand of the actin target sequence (this includes 22 consecutive nucleotides of perfect identity). Such a large stretch of homology seems unlikely to have arisen by chance, and raises the possibility

Figure 5.9 Potential open reading frames in IE 118



The potential open reading frames in IE 118 in both the 'coding' and 'non-coding' strands of DNA are shown. The stop codons are indicated by vertical lines, and the Met start codon are indicated by broken vertical lines.

that it may be the result of a gene conversion event. The presence of a region of 10 base pairs of potential Z-DNA (nucleotides 952 to 962) overlapping this repeat may be significant in this respect, as such sequences have been found associated with other examples of gene conversion (Slightom *et al.*, 1980; Flanagan *et al.*, 1984).

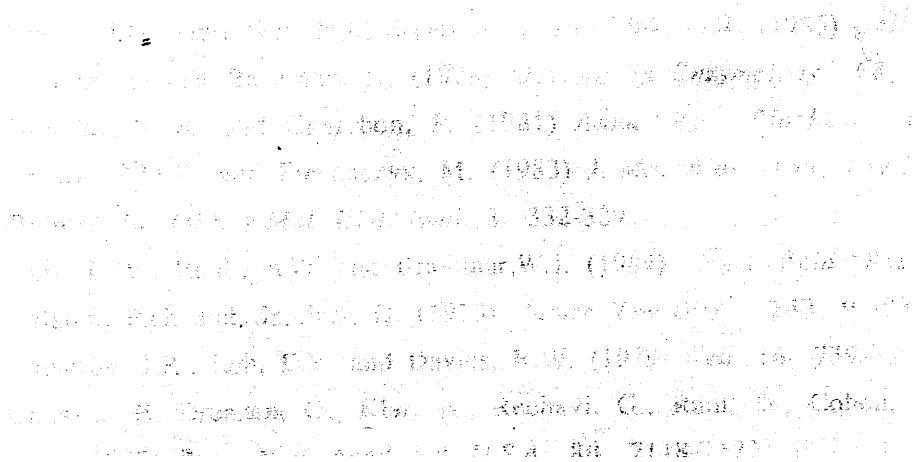
If IE 118 is related to a family of mouse retroviral-like elements it would be expected to be repeated in the mouse genome. Figure 4.3 shows genomic Southern blots of mouse DNA indicating that this is the case (Figure 4.3). Two ^{32}P-labelled probes from IE 118 (see section 4.1.3; Figure 5.8) were used for this purpose: probe 118**a** (a PstI-DraII fragment), containing all the putative functional motifs, and probe 118**b** (a BamHI-PstI fragment), which lacked these but contained the inverted repeat sequence. When these probes were used to screen a mouse genomic library, 118**b** hybridised to the majority of plaques to which 118**a** hybridised, although there were significantly more plaques to which only 118**b** hybridised than those to which 118**a** hybridised. The hybridisation of the mouse genomic library with probe 118**a** gave a generally weaker signal than with the others. A possible trivial explanation of this is that the time for which the first filter (hybridised to 118**a**) was left on the agar plate was in fact considerably less than for the other replica filters. The plaque hybridisation allowed the estimation that there are approximately 1,000 sequences related to 118**a** per haploid mouse genome, and approximately 2,000 sequences related to 118**b**. These figures are consistent with those of retroviral-like LTRs. It is unclear why there appear to be more copies of sequences related to 118**b** than 118**a** in the mouse genome, but it is possible that IE 118 contains extra sequences than the putative retroviral-like element from which it is derived, and that these are repeated in the mouse genome.

5.2.4 Conclusion

The results presented here provide further evidence for the mobility of endogenous murine retroviral-like elements, and are consistent with the view that the majority of processed pseudogenes are functionless. However, it does seem rather surprising that 3 distinct γ-actin processed pseudogenes should be the targets of such retroviral-like elements. Reports of human (Zaberarovsky *et al.*, 1984; Shimada *et al.*, 1984) and rat (Lemischka and Sharp, 1982; Scarpulla, 1985) processed pseudogenes with incorporated SINE or LINE members have been documented. However, the copy number of the latter

mobile elements is thought to be several orders of magnitude greater than that of mouse endogenous retroviral-like elements. From analysis of the present literature it appears that only about 1 in 10 of the human or rat processed pseudogenes so far examined have been targets for such elements. Although examples of mouse processed pseudogenes which are interrupted by inserted elements are lacking, the number of mouse processed pseudogenes on which this assessment is based number no greater than a dozen. There is no reason to suppose that actin pseudogenes are favoured as targets, the points of insertion shown by the three inserted elements described here are all different and bear no sequence similarity. The tendency of other retroposons to insert into one another has been mentioned (Rogers, 1985), but this generally involves sequences close to regions which are adenylate-rich.

The high incidence of insertions into the mouse γ-actin processed pseudogenes may reflect only the small sample of processed pseudogenes so far investigated. However, an alternative explanation is that the number of different classes of endogenous retroviral-like elements in the mouse genome is, in fact, much greater than currently estimated. Although close examination of other related genomic sequences is still required, IE 118 may in fact be a portent of such possible murine retroviral genomic ubiquity. The most pressing task emerging from the work described in this thesis will be to look for a retroviral-like element, the LTRs of which could be related to IE 118.



REFERENCES

Allet, B., Katagiri, K.J. and Gesteland, R.F. (1973) *J. Mol. Biol.* **78**, 589-600.

Alonso, S., Minty, A., Bourlet, Y. and Buckingham, M. (1984) *J. Mol. Evol.* **23**, 11-22.

Anagnou, N.P., O'Brien, S.J., Shimada, T., Nash, W.G., Chen, M. and Nienhuis, A.W. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5170-5174.

Battey, J., Max, E., McBride, O., Swan, D. and Leder, P. (1982) *Proc. Natl. Acad. Sci. U.S.A* **79**, 5956-5960.

Bayev, A.A., Krayev, A.S., Lyubomirskaya, N.V., Ilyin, Y.V., Skryabin, K.G., and Georgiev, G.P. (1980) *Nucleic Acids Res.* **8**, 3263-3273.

Bell, G.I., Pictet, R. and Rutter, W.J. (1980) *Nucleic Acids Res.* **8**, 4091-4109.

Benchimol, S., Jenkins, J.R., Crowford, L.V., Leppard, K., Lamb, P., Williamson, N.M., Pim, D.C. and Harlow, E. (1984) *Cancer Cells* **2**, 383-391.

Benham, F.J., Hodgkinson, S. and Davis, K.E. (1984) *EMBO J.* **3**, 2635-2640.

Bennett, K.L. and Hastie, N.D. (1984) *EMBO. J.* **3**, 467-472.

Bergsma, D.J., Chang, K.S. and Schwartz, R.J. (1985) *Mol. Cell. Biol.* **5**, 1151-1162.

Bernstein, L.B., Mount, S.M. and Weiner, A.M. (1983) *Cell* **32**, 461-472.

Bingham, P. and Judd, B. (1981) *Cell* **25**, 705-711.

Bird, A.P. (1980) *Nucleic Acids Res.* **8**, 1499-1501.

Birnboim, H.C., and Doly, J., (1979), *Nucleic Acids Res.* **7**, 1513-1523.

Blake, C. (1983) *Nature* **306**, 535-537.

Blattner, F.R., Blechl, A.E., Denniston-Thomson, K., Faber, H.E., Richards, J.E., Slightom, J.L., Tucker, P.W. and Smithies, O. (1978) *Science* **202**, 1279-1284.

Boeke, J.D., Garfinkel, D.J., Styles, C.A. and Fink, G.R. (1985) *Cell* **40**, 491-500.

Bolivar, F. and Backman, K. (1979) *Methods in Enzymology* **68**, 245-267.

Breathnach, R. and Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349-383.

Brown, S.M.D. and Piechaczyk, M. (1983) *J. Mol. Biol.* **165**, 249-256.

Bulmer, M. (1986) *Mol. Biol. Evol.* **3**, 332-329.

Burt, D.W., Reith, A.D. and Brammar,W.J. (1984) *Nucl. Acids Res.* **12**, 8579-8593.

Calarco, P.G. and Szollosi, D. (1973) *Nature New Biol.* **243**, 91-93.

Cameron, J.R., Loh, E.Y. and Davies, R.W. (1979) *Cell* **16**, 739-751.

Canaani, E., Dreazen, O., Klar, A., Rechavi, G., Ram, D., Cohen, J.B. and Givol, D. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 7118-7122.

Carbonare, B.D. and Gehring, W.J. (1985) *Mol. Gen. Genet.* **199**, 1-6.

Carmon, Y., Czosnek, H., Nudel, U., Shani, M. and Yaffe, D. (1982) *Nucleic Acids Res.* **10**, 3085-3097.

Carroll, S.L., Bergsma, D.J. and Schwartz, R.J. (1986) *J. Biol. Chem.* **261**, 8965-8976.

Chang, K.S., Rothblum, K.N. and Schwartz, R.J. (1985) *Nucleic Acids Res.* **13**, 1223-1237.

Chang, K.S., Zimmer, W.E., Jr., Bergsma, D.J., Dodgson, J.B. and Schwartz, R.J., (1984) *Mol. Cell Biol.* **4**, 2498-2508.

Chang, L.Y.E. and Slightom, J.L., (1984) *J. Mol. Biol.* **180**, 767-784.

Chen, M.J., Shimada, T., Moulton, A.D., Harrison, M. and Nienhuis, A.W. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 7435-7439.

Cheng, S.M. and Schildkraut, C.L. (1980) *Nucleic Acids Res.* **8**, 4075-4090.

Christy, R.J., Brown, A.R., Gourlie, B.B. and Huang, R.C.C. (1985) *Nucleic Acids Res.* **13**, 289-302.

Clare, J. and Farabaugh, P. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2829-2833.

Cleveland, D.W., Lopata, M.A., MacDonald, R.J., Cowan, N.J., Rutter, W.J. and Kirschner, M.W. (1980) *Cell* **20**, 95-105.

Cohen, J.C. and Varmus, H.E. (1979) *Nature* **278**, 418-423.

Collins, J. and Elinza, M. (1975) *J. Mol. Chem.* **250**, 5915-5920.

Cooper, E.D. and Crain, W.R., Jr. (1982) *Nucleic Acids Res.* **10**, 4081-4091.

Coulondre, C., Miller, J.H., Farabough, P.J. and Gilbert, W. (1978) *Nature* **224**, 775-780.

Czelusniak, J., Goodman, M., Hewett-Emmett, D., Weiss, M.L., Venta, P.J. and Tashian, R.E. (1982) *Nature* **298**, 297-300.

Denison, R.A., Van Arsdell, S.W., Bernstein, L.B., and Weiner, A.W. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 810-814.

Denison, R.A. and Weiner, A.M. (1982) *Mol. Cell. Biol.* **2**, 815-828.

Devereux, J.R., Haeberli, P. and Smithies, O. (1984) *Nucleic Acid Res.* **12**, 387-395.

DiGiovanni, L., Haynes, S.R., Misra, R. and Jelinek, W.R. (1983) *Proc Natl. Acad. Sci. U.S.A.* **80**, 6533-6537.

DiNocera, P.P., Digan, M.E. and Dawid, I.B. (1983) *J. Mol. Biol.* **168**, 715-727.

Doring, H.P. and Starlinger, P. (1984) *Cell* **39**, 253-259.

Doring, H.P., Tillmann, E. and Starlinger, P. (1984) *Nature* **307**, 127-130.

Dudov, K.P. and Perry, R.P. (1984) *Cell* **37**, 457-468.

Duncan, C.H., Jagadeeswaran, P., Wang, R.R.C. and Weissman, S. (1981) *Gene* **13**, 185-196.

Dunsmuir, P., Brorien, W.J., Simon, M.A. and Rubin, G.M. (1980) *Cell* **21**, 576-579.

Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Barralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) *Cell* **21**, 653-668.

Eldridge, J., Zehner, Z. and Paterson, B.M. (1985) *Gene* **36**, 55-63.

Engel, J., Gunning, P. and Kedes, L. (1982) *Mol. Cell. Biol.* **2**, 674-684.

Fanning, T.G. (1982) *Nucleic Acids Res.* **10**, 5003-5013.

Fanning, T.G. (1983) *Nucleic Acids Res.* **11**, 5073-5091.

Farabaugh, P.J. and Fink, R.R. (1980) *Biochemistry* **19**, 5842-5850.

Fedoroff, N.V. (1983) In *Mobile Genetic Elements* (ed. J.A. Shapiro), Academic Press, New York, pp. 1-65.

Fedoroff, N.V., Mauvais, J. and Chaleff, D. (1983) *J. Mol. Appl. Genet.* **2**, 11-29.

Feinberg, A.P. and Vogelstein, B. (1984) *Anal. Biochem.* **137**, 266-267.

Fincham, J.R.S. and Sastry, G.R.K. (1974) *Annu. Rev. Genet.* **8**, 15-50.

Finnegan, D.J. (1981) In *Chromosomes Today* (M.D. Bennett, M. Bobrow, and Hewitt, eds.), Allen and Unwin, London, pp. 84-91.

Finnegan, D.J., Rubin, G.M., Young, M.W. and Hogness, D.S. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 1053-1063.

Flanagan, J.G., Lefranc, M-P. and Rabbitts, T.H. (1984) *Cell* **36**, 681-688.

Flavell, A.J. and Ish-Horowicz, D. (1981) *Nature* **292**, 591-595.

Foran, D.R., Johnston, P.J. and Moore, G.P. (1985) *J. Mol. Evol.* **22**, 108-116.

Fornwald, J.A., Kuncio, G., Peng, I. and Ordahl, C.P. (1982) *Nucleic Acids Res.* **10**, 3861-3875.

Freytag, S.O., Bock, H-G.O., Beaudet, A.L. and O'Brien, W.E. (1984) *J. Biol. Chem.* **259**, 3160-3166.

Fyrberg, E.A., Bond, B.J., Hershey, N.D., Mixter, K.S. and Davidson, N. (1981) *Cell* **24**, 107-116.

Fyrberg, E.A., Klindle, K.L., Davidson, N. and Sodja, A. (1980) *Cell* **19**, 365-378.

Gabbiani, G., Schmid, E., Winter, S., Chaponnier, C., de Chastonay, C. Vandekerckhove, J., Weber, K. and Franke, W.W. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 298-302.

Gafner, J. and Philippsen, P. (1980) *Nature* **286**, 414-418.

Gallwitz, D. and Sures, I. (1980) *Proc. Natl. Acad. Sci. U.S.A.* 77, 2546-2550.

Gebhard, W., Meitinger, T., Hochtl, J. and Zachau, H.G. (1982) *J. Mol. Biol.* **157**, 453-471.

Gehring, W.J. and Paro, R. (1980) *Cell* **19**, 897-904.

Goldberg, M., Paro, R. and Gehring, W.J. (1982) *EMBO J.* **1**, 93-98.

Goodman, M., Koop, B.F., Czelusniak, J., Weiss, M.J. and Slightom, J.L. (1984) *J. Mol. Biol.* **186**, 803-823.

Goosens, M., Dozy, A., Embury, S., Zacharides, Z., Hadjimas, M.,

Stamatoyannoboulos, G. and Kan, Y.W. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 518-521.

Grimaldi, G.J., Skowronski, J. and Singer, M.F. (1984) *EMBO J.* **3**, 1753-1759.

Grindley, N.D.F. (1985) *Cell* **32**, 3-5.

Grindley, N.D.F. and Reed, R.R. (1985) *Annu. Rev. Biochem.* **54**, 863-896.

Gundelfinger, E.D., Krause, E., Melli, M. and Dobberstein, B. (1983) *Nucleic Acids Res.* **11**, 7363-7374.

Gunning, P., Ponte, P., Okayama, H., Engel, J., Blau, H. and Kedes, L. (1983) *Mol. Cell. Biol.* **3**, 787-795.

Hall, B.G. (1983) In *Evolution of Genes and Proteins* (ed. M. Nei and R.K. Koehn), Sinauer, pp. 234-257.

Hamada, H., Petrino, M.G. and Kakunaga, T. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 5901-5905.

Hanauer, A. and Mandel, J.L. (1984) *EMBO J.* **3**, 2627-2633.

Hanukoglu, I., Tanese, N. and Fuchs, E. (1983) *J. Mol. Biol.* **163**, 673-678.

Hardies, S.C., Edgell, M.H. and Hutchison III, C.A., (1984) *J. Biol. Chem.* **259**, 3748-3756.

Harris S., Barrie, P.A., Weiss, M.L. and Jeffreys, A.J. (1984) *J. Mol. Biol.* **180**, 785-801.

Hauber, J., Nelbock-Hochstetter, P. and Feldman, H. (1985) *Nucleic Acids Res.* **13**, 2745-2758.

Hawley, R.G., Shulman, M.J., Murialdo, H., Gibson, D.M. and Hozumi, N. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 7425-7429.

Hayashi, K. (1981) *Nucleic Acids Res.* **9**, 3379-3388.

Haynes, S.R., Toomey, T.P., Leinwand, L. and Jelinek, W.R. (1981) *J. Mol. Cell Biol.* **1**, 573-583.

Hayward, W.S., Neel, B.G. and Astrin, S.M. (1981) *Nature* **290**, 475-480.

Heffron, F. (1983) In *Mobile Genetic Elements* (ed. J.A. Shapiro), Academic Press, New York, pp. 223.

Higgs D.R., Old, J.M., Pressley, L., Clegg, J.B. and Weatherall, D. (1980) *Nature* **284**, 632-635.

Hirashima, M., Yodoi, J. and Ishizaka, K. (1980) *J. Immunol.* **125**, 1442-1448.

Hollis, G.F., Hieter, P.A., McBride, O.W., Swan, D. and Leder, P. (1982) *Nature* **296**, 321-325.

Holmes, D.S. and Quigley, M. (1981) *Anal. Biochem.* **114**, 193-197.

Houck, C.M., Rinehart, F.P. and Schmid, C.W. (1979) *J. Mol. Biol.* **132**, 289-306.

Hu, M.C.T., Sharp, S.B. and Davidson, N. (1986) *Mol. Cell. Biol.* **6**, 15-25.

Hughes, S.H., Payvor, F., Spector, D., Schimke, R.T., Robinson, H.L., Payne, G.S., Bishop, J.M. and Varmus, H.E. (1979) *Cell* **18**, 347-359.

Ikenega, H. and Saigo, K. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 4143-4147.

Ilyin, Y.V., Chmeliauskaite, V.G., Ananiev, E.V. and Georgiev, G.P. (1980a)

Chromosoma **81**, 27-53.

Ilyin, Y.V., Chmeliauskaite, V.G., Georgiev, G.P. (1980b) *Nucleic Acids Res.* **8**, 3439-3457.

Ilyin, Y.V., Tchurikov, N.A., Ananiev, E.V., Ryskov, A.P., Yenikolopov, G.N., Limborska, S.A., Maleeva, N.E., Gvozder, V.A. and Georgiev, G.P. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 959-969.

Ishizaka, K. (1984) *Annu. Rev. Immunol.* **2**, 159-182.

Itin, A. and Keshet, E. (1986) *Journal of Virology* **59**, 301-307.

Jacq, C., Miller, J.R. and Brownlee, G.G. (1977) *Cell* **12**, 109-120.

Jaenisch, R., Harbers, K., Schnieke, A., Lohler, J., Chumakov, I., Jahner, D., Grotkopp, D. and Hoffman, E. (1983) *Cell* **32**, 209-216.

Jagadeeswaran, P., Forget, B.G. and Weissman, S.M. (1981) *Cell* **26**, 141-142.

Jeffreys, A.J. and Harris, S. (1984) *Bioessays* **1**, 253-258.

Jeffreys, A.J., Harris, S., Barrie, P.A., Wood, D., Blanchetot, A. and Adams, S.M. (1983) In *Evolution from Molecules to Men* (ed. D.S. Bendall), Cambridge University Press, pp. 175-195.

Jenkins, N.A., Copeland, N.E., Taylor, B.A. and Lee, B.K. (1981) *Nature* **293**, 370-374.

Johnson, M.S., McClure, M.A., Feng, D.F., Gray, J. and Doolittle, R.F. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7648-7652.

Kalb, V.F., Glasser, S., King, D. and Lingrel, J.B. (1983) *Nucleic Acids Res.* **11**, 2177-2184.

Karin, M. and Richards, R.I. (1982) *Nature* **299**, 797-802.

Karn, J., Brenner, S., Barnett, L. and Cesareni, G. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 5172-5176.

Katzir, N.G., Rechavi, G., Cohen, J.B., Unger, T., Simoni, F., Segal, S., Cohen, D. and Givol, D. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1054-1058.

Klein, A. and Meynhas, O. (1984) *Nucleic Acids Res.* **12**, 3763-3776.

Kohler, G. and Shulman, M.J. (1980) *Eur. J. Immunol* **10**, 467-476.

Kost, T.A., Theodorakis, N. and Hughes, S.H. (1983) *Nucl. Acids Res.* **11**, 8287-8301.

Krayev, A.S., Kramerov, D.A., Skryabin, K.G., Ryskov, A.P., Bayev, A.A. and Georgiev, G.P. (1980) *Nucleic Acids Res.* **8**, 1201-1215.

Kuff, E.L., Feenstra, A., Lueders, K.K., Rechavi, G., Givol, D. and Canaani, E. (1983b) *Nature* **302**, 547-548.

Kuff, E.L., Feenstra, A., Lueders, K., Smith, L., Hawley, R., Hozumi, N. and Shulman, M. (1983a) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1992-1996.

Kuff, E.L., Smith, L.A. and Lueders, K.K. (1981) *Mol. Cell. Biol.* **1**, 216-227.

Kuff, E.L., Wivel, N.A. and Lueders, K.K. (1972) *Proc. Natl. Acad. Sci. U.S.A.* **69**, 218-222.

Lauer, J., Shen, C.K.J. and Maniatis, T. (1980) *Cell* **20**, 119-130.

Leader, D.P.,Gall, I. and Campbell, P.C. (1986b) *Bioscience Reports* **6**, 741-747.

Leader, D.P., Gall, I., Campbell, P. and Frischauf, A.M. (1986a) *DNA* **5**, 235-238.

Leader, D.P., Gall, I. and Lehrach, H. (1985) *Gene* **36**, 369-374.

Leder, A., Swan, D., Ruddle, F., D'Eustachio, P. and Leder, P. (1981) *Nature* **293**, 196-200.

Lee, M.G.S., Lewis, S.A., Wilde, C.D. and Cowan, N.J. (1983) *Cell* **33**, 477-487.

Leibhaber, S.A., Gossens, M. and Kan, Y.W. (1981) *Nature* **290**, 26-29.

Lemischka, I. and Sharp, P.A. (1982) *Nature* **300**, 330-335.

Lerman, M.I., Thayer, R.E. and Singer, M.F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3966-3970.

Leube, R. and Gallwitz, D. (1986) *Nucleic Acids Res.* **14**, 6339.

Levis, R., Dunsmuir, P. and Rubin, G.M. (1980) *Cell* **21**, 581-588.

Li, W.H., Gojobori, T. and Nei, M. (1981) *Nature* **292**, 237-239.

Liebermann, D., Hoffman-Lieberman, B., Weinthal, J., Childs, G., Maxson, R., Mauron, A., Cohen, S.N. and Kedes, L. (1983) *Nature* **306**, 342-347.

Limbach, K.J. and Wu, R. (1985) *Nucleic Acids Res.* **13**, 617-630.

Lloyd, (1983) In *The Cytoskeleton in Plant Growth and Development,* Academic Press, London, pp. 3-29.

Loeb, D.D., Padgett, R.W., Hardies, S.C., Ron Shehee, W., Comer, M.B., Edgell, M.H. and Hutchison III, C.A. (1986) *Mol. Cell. Biol.* **6**, 168-182.

Loening, U.E. (1967) *Biochem J.* **102**, 251-257.

Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. and Tizard, R. (1979) *Cell* **18**, 545-558.

Low, B. (1968) *Proc. Natl. Acad. Sci. U.S.A.* **60**, 160-167.

Lu, R. and Elzinga, M. (1977) *Biochemistry* **16**, 5801-5806.

Lueders, K.K. and Kuff, E.L. (1977) *Cell* **21**, 963-972.

Lueders, K.K. and Kuff, E.L. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 3571-3575.

Lueders, K.K. and Kuff, E.L. (1981) *Nucleic Acids Res.* **9**, 5917-5929.

Lueders, K.K., Segal, S. and Kuff, E.L. (1977) *Cell* **11**, 83-94.

Lund, E. and Dahlberg, J.E. (1984) *J. Biol. Chem.* **259**, 2013-2021.

McClintock, B. (1951) *Cold Spring Harbor Symp. Quant. Biol.* **16**, 13-47.

McDonnell, M.W., Simon, M.N. and Studier, F.W. (1977) *J. Mol. Biol.* **110**, 119-146.

McGrath, J.P., Capon, D.J., Smith, D.H., Chen, E.Y., Seeburg, P.H., Goeddel, D.V. and Levinson, A.D. (1983) *Nature* **304**, 501-506.

McKeown, M. and Firtel, R.A. (1981) *J. Mol. Biol.* **151**, 593-606.

MacLeod, A.R. and Talbot, K. (1983) *J. Mol. Biol.* **167**, 523-537.

Manser, T. and Gesteland, R.F. (1981) *Cell* **29**, 257-264.

Manuelidis, L. (1982) *Nucleic Acids Res.* **10**, 3211-3219.

Martin, S.L., Vincent, K.A. and Wilson, A.L. (1983) *J. Mol. Biol.* **164**, 513-528.

Martin, S.L., Voliva, C.F., Burton, F.H., Edgell, M.H. and Hutchinson III, C.A., (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 2308-2312.

Masters, J.N., Yang, J.K., Cellini, A. and Attardi, G. (1983) *J. Mol. Biol.* **167**, 23-36.

Maxam, A.M. and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560-564.

Maxam, A.M. and Gilbert, W. (1980) *Meths. in Enzymolology* **65**, 499-560.

Mayer, Y., Czosnek, H., Zeelon, P.E., Yaffe, D. and Nudel, U. (1984) *Nucleic Acids Res.* **12**, 1087-1100.

Mellor, J., Malim, M.H., Gull, K., Tuite, M.F., McCready, S., Dibbayawan, T., Kingsman, S.M. and Kingsman, A.J. (1985) *Nature* **313**, 243-246.

Meselson, M. and Yuan, R. (1968) *Nature* **217**, 1110-1114.

Messing, J. (1983) *Method In Enzymology* **101**, 20-78.

Messing, J., Crea, R. and Seebury, P.H. (1981) *Nucleic Acids Res.* **9**, 309-321.

Miller, J.R., Cartwright, E.M., Brownlee, G.G., Fedoroff, N.V. and Brown, D. (1978) *Cell* **13**, 717-725.

Miller, J.R. and Melton, D.A. (1981) *Cell* **24**, 829-835.

Minty, A.J., Alonso, S., Caravatti, M., and Buckingham, M.E. (1982) *Cell* **30**, 185-192.

Minty, A.J., Alonso, S., Guenet, J.L. and Buckingham, M.E. (1983) *J. Mol. Biol.* **167**, 77-101.

Minty, A.J., Caravatti, M., Robert, B., Cohen, A., Daubas, P., Weydert, A., Gros, F. and Buckingham, M.E. (1981) *J. Biol. Chem.* **256**, 1008-1014.

Miyata, T. and Hayashida, H. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 5739-5743.

Miyata, T. and Yasunaga, T. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 450-453.

Miyoshi, J., Kagimoto, M., Soeda, E. and Sakaki, Y. (1984) *Nucleic Acids Res.* **12**, 1821-1828.

Monstein, H-J., Hammarstrom, K., Westin, G., Zabielski, J., Philipson, L., and Pettersson, U. (1983) *J. Mol. Biol.* **167**, 245-257.

Moore, K.W., Jardieu, P., Mietz, J.A., Trounstine, M.C., Kuff, E.L., Ishizaka, K. and Martens, C.L. (1986) *J. Immunol.* **136**, 4283-4290.

Moos, M. and Gallwitz, D. (1982) *Nucleic Acids Res.* **10**, 7843-7849.

Moos, M. and Gallwitz, D. (1983) *EMBO J.* **2**, 757-761.

Mount, S.M. (1982) *Nucleic Acids Res.* **10**, 459-472.

Mount, S.M. and Rubin, G.M. (1985) *Mol. Cell. Biol.* **5**, 1630-1638.

Ng, R. and Abelson, J. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 3912-3916.

Ng, S.Y., Gunning, P., Eddy, R., Ponte, P., Leavitt, J., Shows, T. and Kedes, L. (1985) *Mol. Cell. Biol.* **5**, 2720-2732.

Nishioka, Y., Leder, A. and Leder, P. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2806-2809

Nojima, H. and Kornberg, R.D. (1983) *J. Biol. Chem.* **258**, 8151-8155.

Nudel, U., Mayer, Y., Zakut, R., Shani, M., Czosnek, H., Aloni, B., Melloul, D. and Yaffe, D. (1984) *Expl. Biol. Med.* (Karger, Basel) **9**, 219-227.

Nudel, U., Zakut, R., Katcoff, D., Carmon, Y., Czosnek, H., Shani, M. and Yaffe, D. (1982) In: *Muscle Development*, (Pearson, M.L. and Epstein, E.F., eds.), Cold Spring Harbor Laboratory, N.Y., pp. 177-188.

Nudel, U., Zakut, R., Shani, M., Neuman, S., Levy, Z. and Yaffe, D. (1983) *Nucleic Acids Res.* **11**, 1759-1771.

Ohno, S. (1970) *Evolution by Gene Duplication.* Allen and Unwin, London.

Ohshima, Y., Okada, N., Yani, T., Itoh, Y. and Itoh, M. (1981) *Nucleic Acids Res.* **9**, 5145-5158.

Ono, M.M., Cole, D., White, A.T. and Huang, R.C. (1980) *Cell* **21**, 465-473.

Ono, M. and Ohishi, H. (1983) *Nucleic Acids Res.* **11**, 7169-7179

Ordahl, C.P. and Cooper, T.A. (1983) *Nature* **303**, 348-349.

Orkin, S.H., Old, J., Lazarus, H., Altay, C., Gurgey, A., Weatherall, D.J. and Nathans, D.G. (1979) *Cell* **17**, 33-42.

Panganiban, A.T. (1985) *Cell* **42**, 5-6.

Panganiban, A.T. and Temin, H.M. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 7885-7889.

Paterson, B.M., Segal, S., Lueders, K.K. and Kuff, E.L. (1978) *J. Immunol.* **27**, 118-126.

Payne, G.S., Courtneidge, S.A., Crittenden, L.B., Fadly, H.M., Bishop, J.M. and Varmus, H.E. (1981) *Cell* **23**, 311-322.

Peled-Yalif, E., Cohen-Binder, I. and Meynhas, O. (1984) *Gene* **29**, 157-166.

Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980) *Cell* **20**, 555-566.

Peter, B. and Leader, D.P. unpublished.

Pohlman, R.F., Fedoroff, N.V. and Messing, J. (1984) *Cell* **37**, 635-643.

Ponte, P., Gunning, P., Blau, H. and Kedes, L. (1983) *Mol. Cell. Biol.* **3**, 1783-1791.

Ponte, P., Ng, S.Y., Engel, J., Gunning, P. and Kedes, L. (1984) *Nucleic Acids Res.* **12**, 1687-1696.

Potter, S.S. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1012-1016.

Potter, S.S., Brorien, W.J., Dunsmuir, P. and Rubin, G.M. (1979) *Cell* **17**, 415-427.

Propst, F. and Vande Woude, G.F. (1984) *Nucleic Acids Res.* **12**, 8381-8392.

Proudfoot, N.J., Gill, A. and Maniatis, T. (1982) *Cell* **31**, 553-563.

Proudfoot, N.J. and Maniatis, T. (1980) *Cell* **21**, 537-544.

Razin, A. and Riggs, A.D. (1980) *Science* **210**, 604-610.

Rechavi, G., Grivol, D. and Canaani, E. (1982) *Nature* **300**, 607-608.

Rigby, P., Dieckmann, M., Rhodes, C. and Berg, P. (1977) *J. Mol. Biol.* **113**, 237-251.

Robert, B., Daubas, P., Akimenko, M-A., Cohen, A., Garner, I., Guenet, J-L. and Buckingham, M.E. (1984) *Cell* **39**, 129-140.

Roeder, G.S. and Fink, G.R. (1980) *Cell* **21**, 239-249.

Roeder, G.S. and Fink, G.R. (1983) In *Mobile Genetic Elements* (J.A. Shapiro, ed.), Academic Press, New York, pp. 299-328.

Rogers, J. (1983) *Nature* **301**, 460.

Rogers, J.H. (1985) *Int. Rev. Cytol.* **93**, 187-279.

Rotman, G., Itin, A. and Keshet, E. (1984) *Nucleic Acids Res.* **12**, 2273-2282.

Rubin, G.M., Kidwell, M.G. and Bingham, P.M. (1982) *Cell* **29**, 987-994.

Russell, G.J., Walker, P.M.B., Elton, R.A. and Subak-Sharpe, J.H. (1976) *J. Mol. Biol.* **108**, 1-23.

Sanger, F. (1981) *Science* **214**, 1205-1210.

Sanger, F. and Coulson, A.R. (1978) *FEBS Letters* **87**, 107-110.

Scarpulla, R.C. (1984) *Mol. Cell. Biol.* **4**, 2279-2288.

Scarpulla, R.C. (1985) *Nucleic Acids Res.* **13**, 763-775.

Scarpulla, R.C. and Wu, R. (1983) *Cell* **32**, 473-482.

Scheidereit, C. and Beato, M. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 3029-3033.

Scheller, R.H., McAllister, L.B., Crain, W.R., Jr., Durica, D.S., Posakony, J.W., Thomas, T.L., Britten, R.J. and Davidson, E.H. (1981) *Mol. Cell. Biol.***1**, 609-628.

Scherer, G., Tshudi, C., Perera, J., Deluis, H. and Pirrotta, V. (1982) *J. Mol. Biol.* **157**, 435-452.

Schliwa, M. (1981) *Cell* **25**, 587-590.

Schmid, C.W. and Jelinek, W.R. (1982) *Science* **216**, 1065-1070.

Schmidt, M., Wirth, T., Kroger, B. and Horak, I. (1985) *Nucleic Acids Res.* **13**,

Schon, E.A., Wernke, S.M. and Lingrel, J.B. (1982) *J. Biol. Chem.* **257**, 6825-6835.

Schwartzberg, P.J., Colicelli, J. and Goff, S.P. (1984) *Cell* **37**, 1943-1052.

Setlow, P. (1976) In *Handbook of Biochemistry and Molecular Biology : Nucleic Acids,* vol. 2 : 312, (ed. G.D. Fasman) CRC Press, Cleveland, OH.

Sharp, P.A. (1983) *Nature* **301**, 471-472.

Shen, S., Slightom, J.L. and Smithies, O. (1980) *Cell* **26**, 191-203.

Shen-Ong, G.L. and Cole, M.D. (1982) *J. Virol.* **42**, 411-421.

Shimada, T., Chen, M.J. and Nienhuis, A.W. (1984) *Gene* **31**, 1-8.

Singer, M.F. (1982) *Cell* **28**, 433-434.

Singer, M.F., Thayer, R.E., Grimaldi, G., Lerman, M.I. and Fanning, T.G. (1983) *Nucleic Acids Res.* **11**, 5730-5745.

Slightom, J.L., Blechl, A.E. and Smithies, O. (1980) *Cell* **21**, 627-638.

Soares, M.B., Schon. E., Henderson, A., Karathanasis, S.K., Cate, R., Zeitlin, S., Chirgwin, J. and Efstradiatis, A. (1985) *Mol. Cell. Biol.* **5**, 2090-2103.

Soriano, P., Meunier-Rotival, M. and Bernardi, G. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1816-1820.

Soriano, P., Szabo, P. and Bernardi, G. (1982) *EMBO J.* **1**, 579-583.

Southern, E.M. (1975) *J. Mol. Biol.* **98**, 503-517.

Spritz, R.A., DeRiel, J.K., Forget, B.G. and Weissman, S.M. (1980) *Cell* **21**, 639-646.

Staden, R. (1978) *Nucleic Acids Res.* **5**, 1013-1015.

Stein, J. P., Munjaal, R.P., Lagace, L., Chai, E., O'Malley, B.W. and Means, A.R. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 6485-6489.

Stossel, T.P. (1984) *J. Cell. Biol.* **99**, 15s-21s.

Strobel, E., Dunsmuir, P. and Rubin, G.M. (1979) *Cell* **17**, 429-439.

Suemura, M., Yodoi, J., Hirashima, M. and Ishizaka, K. (1980) *J. Immunol.* **125**, 148-154.

Sutton, W.D., Gerlach, W.L., Schwartz, D. and Peacock, W.J, (1984) *Science* **223**, 1265-1268.

Temin, H.M. (1981) *Cell* **27**, 1-3.

Ueda, S., Nakai, S., Nishida, Y., Hisajima, H. and Honjo, T. (1982) *EMBO J.* **1**, 1539-1544.

Ueyama, H., Hamada, H., Battula, N. and Kakunaga, T. (1984) *Mol. Cell. Biol.* **4**, 1073-1078.

Ullu, E. and Tschudi, C. (1984) *Nature* **312**, 171-172.

Ullu, E. and Weiner, A.M. (1984) *EMBO. J.* **3**, 3303.

Van Arsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, T. and Gesteland, R.F. (1981) *Cell* **26**, 11-17.

Van Arsdell, S.W. and Weiner, A.M. (1984) *Nucleic Acids Res.* **12**, 1463-1471.

Vandekerckhove, J., de Couet, H.-G. and Weber, K. (1983) In: *Actin Structure and Function in Muscle and Non-muscle Cells.* Academic Press, Sydney, Australia, pp. 241-248.

Vandekerckhove, J., Franke, W. and Weber, K. (1981a) *J. Mol. Biol.* **152**, 413-426.

Vandekerckhove, J., Leavitt, J., Kalunaga, T. and Weber, K. (1981b) *Cell* **22**, 893-899.

Vandekerckhove, J. and Weber, K. (1978a) *Proc. Natl. Acad. Sci. U.S.A.* **75**, 1106-1110.

Vandekerckhove, J. and Weber, K. (1978b) *J. Mol. Biol.* **126**, 783-802.

Vandekerckhove, J. and Weber, K. (1978c) *Eur. J. Biochem.* **90**, 451-462.

Vandekerckhove, J. and Weber, K. (1979a) *Differentiation* **14**, 123-133.

Vandekerckhove, J. and Weber, K. (1979b) *FEBS Lett.* **102**, 219-222

Vandekerckhove, J. and Weber, K. (1984) *J. Mol. Biol.* **179**, 391-413.

Vanin, E.F. (1984) *Biochem. Biophys. Acta.* **782**, 231-241.

Vanin, E.F., Goldberg, G.I., Tucker, P.W. and Smithies, O. (1980) *Nature* **286**, 222-226.

Varmus, H.E. (1982) *Science* **216**, 812-820.

Varmus, H.E. (1983) In *Mobile Genetic Elements* (ed. Shapiro, J.A.) Academic

Press, New York, pp. 411-503.

Voliva, C.F., Jahn, C.L., Comer, M.B., Edgell, M.H. and Hutchison III, C.A. (1984) *Nucleic Acids Res.* **11**, 8847-8859.

Walter, P. and Blobel, G. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7112-7116.

Warmington, J.R., Waring, R.B., Newlon, C.S., Indge, K.J. and Oliver, S.G. *Nucleic Acids Res.* (1985) **13**, 6679-6692.

Watanabe-Nagasu, N., Itoh, Y., Tani, T., Okano, K., Koga, N., Okada, N. and Ohshima, Y. (1983). *Nucleic Acids Res.* **11**, 1791.

Weihler, H., Konig, M. and Gruss, P. (1983) *Science* **219**, 626-631.

Weiss, R., Teich, N., Varmus, H. and Coffin, J. (1982) In *Molecular Biology of RNA Tumor Viruses.* Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Weissbach, A. (1977) *Annu. Rev. Biochem.* **46**, 25-47.

Weydert, A., Daubas, p., Caravatti, M., Minty, A., Bugaisky, G., Cohen, A., Robert, B. and Buckingham, M. (1983) *J. Biol. Chem.* **258**, 13867-13881.

Wiedemann, L.M. and Perry, R.P. (1984) *Mol. Cell. Biol.* **4**, 2518-2528.

Wilde, C.D., Crowther, C.E., Cripe, T.P., Lee, M. G-S. and Cowan, N.J. (1982a) *Nature* **297**, 83-84.

Wilde, C.D., Crowther, C.E. and Cowan, N.J. (1982b) *Science* **217**, 549-552.

Wilson, S.H. and Kuff, E.L. (1972) *Proc. Natl. Acad. Sci. U.S.A.* **69**, 1531-1536.

Wilson, R. and Storb, U. (1983) *Nucleic Acids Res.* **11**, 1803-1816.

Wirth,T., Gloggler, T., Baumrdker, T., Schmidt, M. and Horak, I. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3327-3330.

Witney, F.R. and Furano, A.V. (1984) *J. Biol. Chem.* **259**, 10481-10492

Yaffe, D., Nudel, U., Mayer, Y. and Neuman, S. (1985) *Nucleic Acids Res.* **13**, 3732-3737.

Yanisch-Perron, C., Vieira, J. and Messing, J. (1985) *Gene* **33**, 103-119.

Ymer, S., Tucker, W.Q.J., Sanderson, C.J., Hapel, A.J., Campbell, H.D. and Young, I.G. (1985) *Nature* **317**, 255-257.

Yotsuganagi, Y. and Szollosi, D. (1981) *J. Natl. Cancer Inst.* **67**, 677-685.

Young, R.A. and Davis, R.W. (1980) *Nature* **22**, 778-782.

Zabarovsky, E.R., Chumakov, I.M., Prassolov, V.S. and Kisselev, L.I. (1984) *Gene* **30**, 107-111.

Zakut, R., Shani, M., Givol, D., Neuman, S., Yaffe, D. and Nudel, U. (1982) *Nature* **298**, 857-859.

Zakut-Houri, R., Oren, M., Bienz, B., Lavie, V., Hazum, S. and Givol, D. (1983) *Nature* **306**, 594-597.