ISOLATION AND CHARACTERISATION OF A LOCUS DISRUPTED IN A TRANSGENIC MOUSE MUTANT EXHIBITING SEX-LINKED CLEFT PALATE

Hugh Gerard Laverty

Thesis presented in partial fulfilment
of the degree of Doctor of Philosophy
to the University of Glasgow

Robertson Laboratory of Biotechnology
Division of Molecular and Cellular Biology
Institute of Biomedical and Life Sciences
University of Glasgow
Glasgow, G11 6NU

March, 1995.

"......feeling the drunkeness of all things being various...."

Snow, Louise MacNeice.

CONTENTS

LIST OF FIGURES AND TABLES

FIGURES:

TABLES

Acknowledgements:

Special thanks to Fiona for all her support and understanding over the previous months. I would like to thank my supervisor Joanna Wilson for her support, advice and project. Many more people deserve thanks for the past few years than can be mentioned, but here is a balanced scientific sample. Thanks to Stephen and Jo for putting up with me and everybody else in the lab (especially the cleaners who have taken many's a year off my life); to all my fellow departmental Friday night drunkards especially, Leslie, Sandy, Maurice, John, Jo (yes another one!) and Tam; my family and friends also deserve some thanks, especially my Mother, Michael, Martin and everybody at Queen's; the menagerie of friends and acquantainces bumped into and fallen over in the pubs and clubs of Glasgow. A special thanks to those who work in the dens of iniquity of The Cul de Sac, Whistler's Mother, The Exchequer and Rooftops. Other thanks go to James Joyce, Flann O'Brien, Franz Kafka, MBV, The Fall, RFTK and all my other books and records for giving me refuge and a constant reminder that there's more to life than this. And the deepest heart felt thanks to Arthur Guinness and Co.

> " When money's tight and hard to get
> and your horse has also ran,
> When all you have is a heap o'debt
> A pint of plain's your only man."
>
> *At Swim Two Birds*, Flann O'Brien.

Thanks and gratitude to the Wellcome Trust who funded this work and continue to support my scientific endeavours.

Abstract

A unique transgenic mouse line displayed the phenotype of sex-linked cleft palate. As no other mouse line expressing this transgene demonstrated a similar phenotype, it is argued that the disorder is due to the disruption of the cellular locus. At the very least this mutation results in redirecting expression of the transgene in a unique fashion, alternatively, a gene(s) at the locus is critical to normal secondary palate development. Therefore 32.8Kb of cellular DNA of the cellular locus disrupted in this mutant has been cloned and characterised.

This 32.8Kb of wild type DNA has been characterised at the level of conservation of sequence between species and expression. No potential gene encoding sequences were initially found. Subsequently, using the technique of exon-trapping several exons have been isolated from the wild type locus. When these exons were hybridised to Northern blots of embryonic and post-natal RNAs they revealed a 4.5Kb, a 0.8Kb and a 0.5Kb transcript, each expressed in the head and body during embryonic development. The 0.8Kb message was shown to be expressed post-natally in a tissue specific manner.

Five individual exons have been isolated and sequenced. They range in size from 81bp to 446bp. One exon (exon p50, 446bp in size) was found to be composed of repetitive sequence but it did not contain an open reading frame (ORF) that spanned the exon. The other four exons (exon p51, p52, p56 and p57) were found to contain an ORF that spanned the exon. Database analysis with the sequence of the exons revealed that two of the exons (exon p56 and exon p57) showed no homology to nucleotide sequences in databases, and another (exon p52) exhibited a high level of similarity to CaM kinase II at both the nucleotide and amino acid level.

Using RT-PCR, two of the exons (exon p52 and exon p57) were shown to be expressed at the RNA level during development in the wild type mouse. Exon p52 and exon p57 were also found to be spliced together. Cloning and sequencing of this exon p52-exon p57 RT-PCR product revealed that this RT-PCR product contained a single ORF that spanned the sequence of both exons spliced together. The amino acid sequence of the putative peptide

encoded by this ORF revealed 47% identity to the calmodulin-binding domain of CaM kinase II. Analysis of this RT-PCR product in RNA derived from transgenically positive spleen revealed that several products are present which are absent in the wild type. These novel products hybridise with exon p52 and exon p57 sequences as well as the transgene sequences, indicating that the splicing of the RT-PCR product is aberrant in the transgenic mutant.

A human form of heritable sex-linked cleft palate has been well documented. The phenotypes observed in the transgenic mutant and the cloning of the gene from the wild type locus may lead to the identification of the human gene involved in heritable sex-linked cleft palate, as well as provide a model for the analysis of the disorder.

ABBREVIATIONS

EBV	Epstein-Barr Virus
LMP-1	Latent membrane protein 1
LINE-1	Long Interspersed Nuclear Element
LRE	LINE retrotransposable element
kDa	kilo daltons
bp	base pairs
kb	kilo bases
aa	amino acids
μg	micrograms
μl	microlitres
secs	seconds
mins	minutes
hrs	hours
RT	Room temperature
FCS	Foetal calf serum
DMEM	Dulbeccos modified Eagles medium
ORF	open reading frame
EST	Expressed sequence tag
CaM kinase	Calcium/calmodulin-dependent protein kinase
PCR	polymerase chain reaction
RT-PCR	Reverse transcribed-PCR
CP+A	Cleft palate plus ankylossia

The nomenclature for restriction enzymes used in this thesis is wrong, rather than have *EcoRI* the enzyme abbreviations should read *Eco* RI. This applies to all the enzymes quoted in the text.

Single letter Code for Amino Acids:

Alanine	A
Arginine	R
Asparagine	N
Aspartic acid	D
Cysteine	C
Glutamic acid	E
Glutamine	Q
Glycine	G
Histidine	H
Isoleucine	I
Leucine	L
Lysine	K
Methionine	M
Phenylalanine	F
Proline	P
Serine	S
Threonine	T
Tryptophan	W
Tyrosine	Y
Valine	V

1 INTRODUCTION

1.1 Mouse developmental mutants

This thesis describes the isolation and characterisation of a locus whose disruption by transgene insertion resulted in a mouse developmental mutant that exhibited sex-linked cleft palate. Much of the progress in mammalian developmental biology throughout this century is based on studies of the developmental genetics of mutations in the laboratory mouse. Many of the 1300 or so genetic loci currently mapped in the mouse were identified originally on the basis of a mutant developmental phenotype in inbred mouse strains (Lyon and Searle, 1989). These classic mutants have been augmented by the use of highly efficient chemical mutagens and X-ray irradiation to generate new mutants (for review, see Rinchik 1991). For the major part of this century the descriptions of the temporal and spatial aspects of these mutant gene phenotypes helped reveal information about the respective wild type gene action, while the underlying molecular mechanisms of gene action were inaccessible. With the advances in recombinant DNA technology and improvements in genetic and physical mapping methodologies, the large collection of mutants generated is now being defined at the molecular level (Reith and Bernstein, 1991).

The advances in recombinant DNA technology and embryonic stem cell culture allow the generation of new mouse mutants for known loci (Capecchi, 1989; Rudnicki and Jaenisch, 1991). When this capability is coupled with the new methodologies for analysing the genes of organisms, the genomes of organisms can now be modified to address specific developmental problems. Over the past 10 years transgenic technology has succeeded in generating a range of exciting developmental mutants. The analysis of these mutants has given a great insight into the mechanisms of mammalian embryogenesis and its genetic control. With the increased capabilities of technology the goal of the present day developmental biologist remains unchanged from that of the pioneers', namely to identify the role of the genes responsible for controlling the developmental mechanisms active during embryogenesis.

Some genes that play critical roles in murine development have been identified because there was pre-existing knowledge about the basic biochemical defect, or they were the homologues of genes essential for normal development in other organisms, a process known as 'functional cloning'. Unfortunately, 'functional cloning' is of no use for the vast majority of genes active during development as no functional information exists. An alternative approach to identifying gene loci is the method of 'positional cloning' (Collins, 1992). It requires the physical location of the gene in the genome to be known and usually little information is required regarding the gene's function. Positional cloning is based on the methods of chromosome walking and the identification of expressed sequences. This can be extremely laborious and time consuming especially if the region of the chromosome to be walked is large. Sometimes other features of the genome can aid in the cloning of the gene, e.g. cytogenetic markers such as translocations and rearrangements. In some cases the nature of the mutant gene's sequence can aid in its isolation, e.g. the expanded trinucleotide repeat of fragile X locus (Verkerk *et al.*, 1991) can be detected by Southern analysis.

However for the majority of genes present in the mouse genome the position of the locus is unknown. In these cases the size of the region to which the gene is assigned (the critical region) is usually more than two megabases, (2×10^6 base pairs), requiring the isolation and examination of all the genes contained in that region. Positional cloning is very labour intensive and expensive and it will benefit greatly from the ability to assign gene loci accurately on the genome. This assignment requires the generation of a high resolution molecular genetic linkage map of the mouse genome (Copeland and Jenkins, 1991). Positional cloning will also benefit from the construction of overlapping yeast artifical chromosome contigs that cover the entire genome.

With the increased information available from a physical map of the genome and the new methodologies for analysing DNA, in time, functional and positional cloning will be superseded by the candidate gene approach (Ballabio, 1993). The candidate gene approach does not require the isolation of new genes, but relies on the availability of information from previously isolated genes. With the steady increase in transcript density for each mouse chromosome the candidate gene approach will gradually improve with time. A striking example of the success of the candidate gene approach is the cloning of the human glycerol kinase gene (GK). Two groups spent several years mapping and cloning the GK critical region before finally isolating the gene by positional cloning, whereas another group mapped a randomly isolated expressed sequence tag (EST), homologous to bacterial glycerol kinase, to the human Xp21 region (Ballabio, 1993). The second group found that the gene they had isolated directly using EST screening was the candidate gene for human glycerol kinase deficiency. It was the combined information on function and position which led to the gene's identification, bypassing the time consuming approach of positional cloning.

When a locus has no known function and no candidate gene is available the rate limiting step in gene identification is the need to sift through hundreds of thousands of base pairs to identify a transcript. A method to avoid the time consuming and costly sift of genetic sands is to tag the gene in some way. The tag acts as a genetic landmark that simplifies identification of the gene. The tag sometimes generates a phenotype by its presence and this phenotype may reveal information about the gene's normal function. Since the advent of transgenic technology there have been several interesting reports of mutants generated by transgene insertion. There has also been a drive to produce a more systematic method of generating insertional mutations using enhancer trap and reporter constructs (Rijkers *et al.*, 1994). These new systems, coupled with advances in embryonic stem cell culture allow a prescreening for patterns of expression before the analysis of phenotype requires extensive breeding. Here, however only insertional mutations as a result of retroviral infection or microinjection of transgenic DNA and some spontaneous mutants are discussed further.

1.2 Insertional mutagenesis

The process of generating transgenic mice requires the introduction of foreign DNA into a host's genome. There are two major strategies for achieving this, the first is retroviral infection and the second is microinjection of linear constructs. It is logical to assume that if the pattern of integration of the transgene is purely random then the transgene should integrate in host genes at a certain frequency. The presence of the transgene may disrupt or insertionally mutate a host gene resulting in a mutant phenotype. This method of generating mutants aids the molecular analysis of the disrupted gene as the problems associated with positional cloning are not encountered. The presence of transgene sequences in or near the host gene acts as a readily identifiable marker for the host gene enabling the cloning of the interrupted locus. This allows the structure and function of the mutated gene to be analysed using molecular genetic techniques. However, problems do occur depending on the method chosen to introduce the transgenic sequences into the host.

There are now well documented examples of the generation of insertional mutants in the mouse by either, microinjection of cloned DNA into the pronucleus of fertilised oocytes or by retroviral infection of pre-implantation embryos.

Retroviral insertional mutagenesis

Retroviruses are RNA viruses that replicate via a DNA intermediate called the provirus. Transgenic mice can be generated at high frequency by retroviral infection of pre-implantation embryos. A commonly used retrovirus in the generation of transgenic mice is the Moloney-Murine leukaemia virus (Mo-MLV). This retrovirus can cause viremia, as a result of its replication in the host cells which has been shown to induce leukaemia in mice. To prevent this, replication incompetent recombinant retroviruses are used to generate the transgenic mice. Concentrated recombinant retrovirus is introduced to 8- or 16-cell embryos by infection prior to introducing the embryos into the uteri of pseudopregnant mice. Approximately 80% of the preimplantation embryos exposed to the retrovirus give rise to transgenically positive mice. During replication a single provirus integrates into the hosts genome.

In contrast to pronuclei microinjection, retroviral integration results in a short duplication of cellular sequence on either side of the provirus. Therefore, any mutant phenotype associated with proviral integration is more likely to be a result of the integration of the provirus rather than larger rearrangements in genomic DNA. The presence of a single copy of the provirus and the lack of any major alterations in the cellular sequence of the transgenic mouse greatly aids in the interpretation of any phenotype that may result from proviral integration. This is a great aid when it comes to the isolation of genes affected by retroviral integration. However, a draw back of retroviral insertional mutagenesis is the preference shown by retroviruses to integrate near DNase-1-hypersensitive regions, suggesting a preference for transcriptionally active chromatin or chromatin in an open configuration. This may result in proviral integration in or near the subset of genes active during the earliest stages of embryogenesis. Although the majority of insertional mutants have been generated by microinjection, because of the frequent complexity of integration most of the best characterised mutants are those that have been generated by retroviral insertion (Rudnicki and Jaenisch, 1991). These transgenic mutants exhibit a wide range of interesting phenotypes that allow a study of many different molecular processes and the subsequent identification of genes involved.

The first described insertional mutation was the *Mov13* mouse (Jaenisch *et al.*, 1983). It had been generated by the infection of day 8 mouse embryos by Mo-MLV. Embryos homozygous for the insertion at the *Mov13* locus die at day 12 of development, this results from necrosis of erythropoietic and mesenchymal cells in the liver. Integration of the Mo-MLV was found to be in the first intron of the $\alpha1$(I) collagen gene, reducing its level of expression in most tissues (Kratochwil *et al.*, 1989). Evidence suggested that the reduction in expression observed in the *Mov13* line results from disruption of *cis* -acting regulatory elements within the first intron.

However, the introduction of of the human proα1 (I) collagen gene into the *Mov13* line resulted in only partial rescue of the homozygous phenotype. This indicated that either there is a subtle functional defect in the hybrid protein or that a second functional gene was disrupted by the Mo-MLV integration (Wu *et al.*, 1990).

Another interesting insertional mutant mouse is the *Mpv17* mouse line resulting from the infection of preimplantation embryos with an MPSVneo retrovirus (Weiher *et al.*, 1990). The vast majority of mutants characterised show embryonic failures, however, the *Mpv17* mutant develops normally and only exhibits a phenotype in the adult mouse. Adult mice homozygous for the *Mpv17* integration develop nephrotic syndrome and chronic renal failure, occurring between 2 and 9 months post-natally. Fragments of DNA detected flanking the provirus revealed a 1.7kb RNA transcript that was ubiquitously expressed during embryonic development and expressed at high levels in the kidney, brain and heart of the adult. This transcript was not detected in *Mpv17* homozygous for the insertion mice. The 1.7kb transcript encodes a 176 amino acid residue protein that contains two hydrophobic domains suggesting membrane association. This *Mpv17* mouse line has not only been used to identify a gene important in kidney function, but it may also prove to be a mouse model for human nephrotic syndrome (Weiher *et al.*, 1990). To isolate a gene of this nature from human patients with renal failure would have been very expensive and very labour intensive. Insertional mutagenesis offers the opportunity to generate mouse models of human diseases either by design or good fortune.

Spontaneous insertional mutations

Since the first description of a retroviral insertional mutation there have been several reports of mutations generated by spontaneous retroviral and retrotransposon insertion. Bowes *et al.*, (1993), have shown that the mutated gene responsible for the disease of retinal disintegration in the *rd* mouse encodes the β subunit of cGMP phosphodiesterase (β-PDE). It was observed that an endogenous xenotropic murine leukaemia virus (*Xmv*) was closely linked to the *rd* locus by genetic linkage analysis. Closer analysis revealed that the *Xmv-28* provirus had integrated within the first intron of the *rd* β-PDE gene in the opposite orientation with respect to the sense strand of the β-PDE gene. This integration event was associated exclusively with the *rd* phenotype. In *rd* mice the level of β-PDE transcripts was markedly decreased and the transcripts are of abnormal size. It appears that proviral integration results in the *rd* phenotype by disrupting transcription of the β-PDE gene.

There have been several reports of proposed retrotransposable element integrations resulting in mutant phenotypes in humans (Holmes *et al.*, 1994). The L1 elements are found in all mammalian genomes and belong to a class of retrotransposable elements that lack long terminal repeats. When the L1 element replicates it usually exhibits a variable length target site duplication flanking the element. Other features include a 3' poly A tail, along with 5' truncation of the majority of elements (only about 3% of the 50,000-100,000 elements present in the human genome are the full length of 6.1kb). Holmes *et al.*, (1994), reported the identification of an L1 insertion into the dystrophin gene resulting in Duchenne muscular dystrophy and the isolation of its precursor. They were able to identify the L1 element that was the precursor of the L1 element that inserted in the dystrophin gene based on the presence of a single-copy unique sequence component (USC) flanking the insertion and the precursor.

Insertional mutagenesis by microinjection of zygotes

Microinjection of cloned DNA into the pronucleus of fertilised mouse oocytes results in integration of the transgenic DNA in about 25% of animals born (Rudnicki and Jaenisch, 1991). Insertional mutants are generated at a frequency of about 8% and many of these are recessive lethal. This frequency may be a low estimate as many of the phenotypes may be very subtle and difficult to detect. The majority of transgenic mice generated by microinjection contain multiple copies of the transgene construct, often as tandem repeats at a single locus. At the site of integration, the host chromosome often has undergone deletions, inversions, duplications, translocations, and other major modifications. Most of the insertional mutants that have been characterised were generated by microinjection. This reflects the popularity of microinjection over other methods of introducing transgenic DNA, rather than any inherent advantage of microinjection for generating insertional mutants. Also, possibly the rearrangements generated a more dramatic phenotype (more mutational).

Although microinjection of pronuclei has generated several exciting insertional mutants, there are a number of serious limitations, the major limitation being the integrity of the integration site in the host DNA. As mentioned above when microinjected transgenes integrate they do so in a complex fashion, usually inducing profound changes in the chromosomal structure and organisation of cellular sequence at the integration site. These changes in cellular sequence, e.g., large inversions and translocations may affect more than one gene, making it difficult to identify the cause of the observed phenotype, e.g., the *legless* mouse insertional mutant has at least two integrated transgenes separated by 30kb of cellular sequence integrated on chromosome 12 (Singh *et al.*, 1991).

The *legless* mutant mentioned earlier is a transgenic insertional mutation found in the pHT1-1 line of transgenic mice (McNeish *et al.*, 1988). All the mice homozygous for this mutation have deficient hindlimbs, with structures distal to the femur absent. The forelimbs of legless mice are also malformed, but this malformation is highly variable.The mutant mice have aberrant brains, with the most anterior structures usually absent. Craniofacial clefts are observed in about 50% of the mutants and half of the mutant mice exhibited situs inversus, a mirror image arrangement of internal organs. Molecular analysis of the disrupted region indicated two tightly linked insertion sites separated by 30kb of cellular sequence. Evidence suggests that transgene insertion has disrupted more than one gene, and several cDNAs have been isolated from libraries (Singh *et al.*, 1991).

Several of the insertional events described in the literature resulted in the mutation of genes important in development. It should be stressed that the insertional mutations were serendipitous events, the transgenic mice were created to investigate the effect of a transgene's expression, not to create an insertional mutant. Also, many insertional events generating mutant phenotypes are not identified as they result in embryonic lethality and are not detected in the screen for the expected phenotypes. While some of the insertional mutations result in the identification of novel loci involved in developmental pathways, other insertional events produce phenotypes similar to well characterised mouse mutants. An example of this is the insertional mutation anterior digit-pattern deformity, *add*. Generated during the construction of transgenic mice harbouring the human α1-acid glycoprotein genes (Dente *et al.*, 1988), *add* mice exhibit a forelimb defect where the thumb becomes elongated and the second digit often contains an extra phalanx (Pohl *et al.*, 1990). Using mouse DNA from the site of insertion flanking the transgene, *add* has been mapped close to the centromere on chromosome 13. Evidence suggests that the *add* locus is tightly linked to a genetically mapped locus *extra-toes* (Xt). The phenotype of the double mutant *add* /Xt suggests that they are allelic.

Another insertional mutant that exhibits limb deformities is the novel mouse mutant *fused toes* (Ft), discovered in the course of generating transgenic mice with the activated human *Ha-ras* gene under the control of the human β-interferon promoter (Van der Hoeven *et al.*, 1994). A founder female exhibited the novel phenotype characterised by fused toes on the fore limbs and hemizygous animals displayed a thymic hyperplasia. Homozygosity of the transgene insertion leads to malformation of the developing brain, loss of the genetic control of left-right asymmetry and to death around day 10 of embryonic development. Analysis of limb development and immature thymocytes *in vitro* indicated that apoptosis (programmed cell death) is affected in the mutant line. As the phenotype may be a result of transgene insertion, it was possible to use the transgene sequences to map the site of insertion to chromosome 8. As with the *fused toes* mutant no gene has so far been identified at this locus. To date no gene has been isolated from around the site of insertion, but this insertional mutation may allow the identification of a novel gene involved in apoptosis during murine development. In case of both the *add* and Ft mutants it has not been formally proven that transgene expression is not involved in the phenotypes observed.

The rearrangements accompanying transgene insertion that occurs with microinjection of zygotes makes the identification of loci responsible for the mutant phenotype very difficult. Wilson *et al.*, (1993), have reported a transgenic mouse line that exhibits a mutant phenotype of sex-linked cleft palate, the transgene contains the Latent Membrane Protein-1 (LMP-1) from Epstein-Barr Virus (EBV). Analysis of this transgenic line revealed that the transgene was present on the X chromosome as a tandem repeat of two copies in a head to tail configuration. Unexpectedly the site of transgene integration appeared not to have undergone any major rearrangements, the transgene flanking fragments were isolated and found to be contiguous in wild type DNA (McFarlane and Wilson, in press). Therefore the affected locus is only disrupted by the presence of the transgene and not affected by major rearrangements. This thesis describes the work aimed at characterising the disrupted locus in this sex-linked cleft palate insertional mutant.

This sex-linked cleft palate transgenic mutant offered the chance to isolate a gene involved in secondary palate formation during embryogenesis. The advantages of this mouse model are:

(1) The clean insertion of the transgene will aid the genetic interpretation of the mutation.

(2) The phenotype presented in the mutant is "relatively simple".

(3) The palate is a well studied structure through development but little is known about the genetic controls.

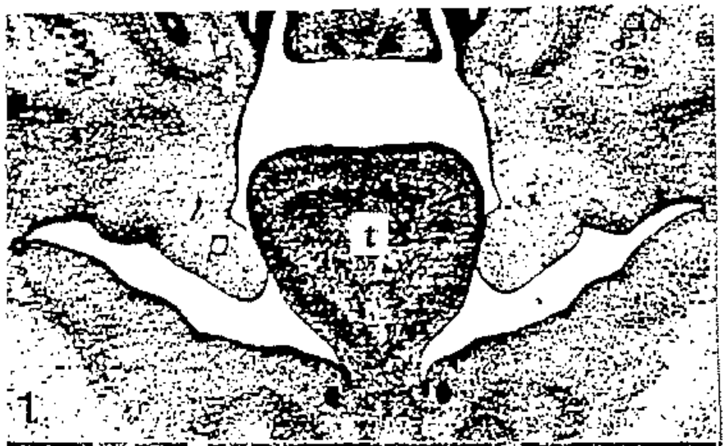
(4) This insertional mutant is a possible mouse model for the documented and mapped human heritable disorder of sex-linked cleft palate. The non sex-linked form of cleft palate is a very common disorder in humans.

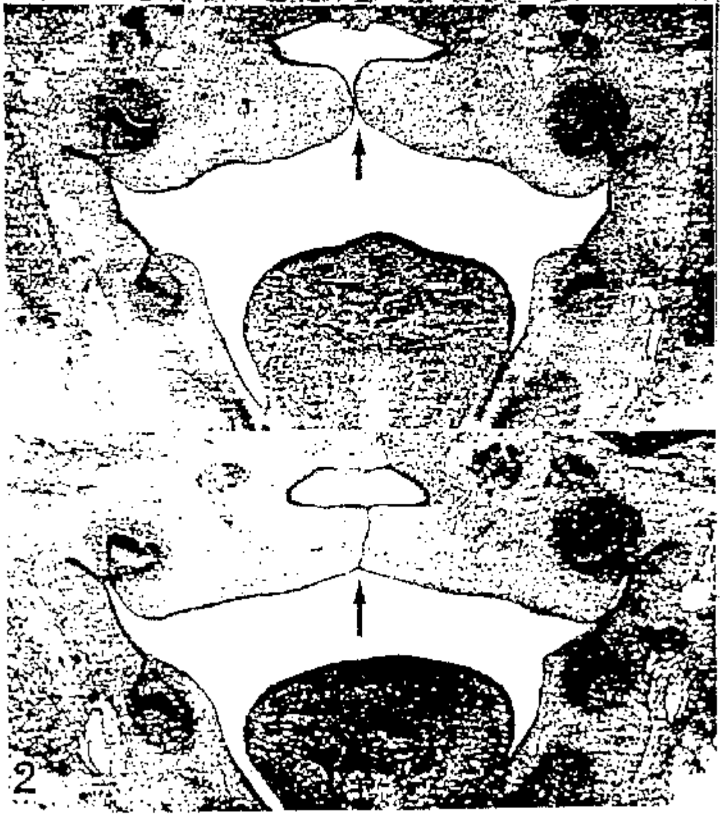
1.3 Morphogenesis of the secondary palate

The developing palate is often used as a model system for investigating fundamental questions common to the embryogenesis of many structures as it appears relatively late in embryogenesis and can be excised and cultured under chemically defined conditions (Ferguson, 1988). It also exhibits many important features such as morphogenetic movements, extracellular matrix synthesis, neurotransmitter synthesis, cell adhesion, epithelial-mesenchymal interactions and regional patterning (reviewed in Ferguson, 1987).

Mesenchymal cells from the neural crest migrate to the primitive oral cavity where in association with craniopharyngeal ectoderm they form the bilateral maxillary processes. The palatal shelves arise from the maxillary processes and initially grow vertically down the sides of the tongue. At a precise moment during development they elevate to a horizontal position above the tongue where by a combination of processes they fuse at about day 15 of embryonic development (Figure 1). The palatal shelves associate with each other through the medial edge epithelia, this rapidly degenerates by a combination of cell death and cell migration. This establishes mesenchymal continuity across the unified palate, the mesenchymal cells then consolidate and differentiate. Cleft palate may result from the perturbation of any number of critical processes during palate formation: defective palatal shelf growth, delayed or failed shelf elevation, defective shelf fusion, failure of cell death

Figure 1: Histological section through the anterior region of the palate in a (a) day 13.5 embryonic mouse head and traverse histological sections through the mid-palate region of (b) early day 14 and (c) late day 14 embryonic mouse head. Note the vertical palatal shelves (p) firmly wedged against the tongue (t) in preparation for shelf elevation in figure 1(a). The palatal shelves first approximate each other, figure 1(b), their medial edge epithelia (arrowed) then contact with each other and fuse to form the midline epithelial seam, figure 1(c). x21 (Ferguson, 1988).


t
1.


2

or cell migration along the medial epithelial edge or failure of subsequent mesenchymal differentiation resulting in postfusion rupture (Ferguson, 1988).

Trying to unravel the molecular processes that govern palate development is therefore very difficult, considering the number and complexity of the mechanisms involved. Each of these molecular mechanisms will be regulated by any number of genes and so cleft palate may result from several genetic dysfunctions. The susceptibility of the palate to clefting is also reflected in the ability to induce cleft palate with the use of chemicals such as retinoic acid and alcohol. The high incidence of cleft palate observed with craniofacial malformations suggests that the development of the palate is very dependent on the orofacial environment. These failures can result from defects in the palate itself, e.g. failed shelf elevation or steric interference from craniofacial morphology. It was observed that during the period of shelf elevation, there is almost no growth in head width, but constant growth in head height. This means that the base of the growing palatal shelves are moving away from the tongue, resulting in the path of least resistance for shelf elevation being above the tongue (Diewart, 1978). Therefore, changes in craniofacial morphology may well impair shelf elevation preventing subsequent palatal shelf fusion.

Shelf elevation is a rapid event occurring in minutes or hours *in vivo*. It is governed by an intrinsic force generated within the palatal shelves, the major component of which is a regionally specific build up of glycosaminoglycans, especially hyaluronic acid. Hyaluronic acid is highly electrostatically charged and capable of binding up to 10 times its weight in water. In the early stages of palatal shelf growth the accumulation of hyaluronic acid is responsible for the majority of the increase in shelf size by extracellular matrix swelling rather than mesenchymal cell division. The shelf elevating force is partly delivered by stout bundles of type I collagen which run down the centre of the shelf from the base. The epithelial covering of the shelves exhibits differential traction that helps direct the swelling force. The mesenchymal cells at the centre of the shelves are themselves contractile and able to secrete neurotransmitters that regulate the degradation of glycosaminoglycans and influence the contractility of the mesenchymal cells themselves (Ferguson, 1988). When the palatal shelves make contact they adhere to each other by means of a sticky cell surface glycoprotein coat, to form an epithelial seam. This epithelial interaction is specific, the epithelium of the palatal shelves will not fuse with any other epithelium.

Mouse medial edge epithelial cells rapidly form desmosomes and accumulate desmoplakin on their cell membranes just prior to contact (Ferguson, 1988). Antibodies generated to the medial edge epithelium have identified several new cell surface molecules, none of the antibodies to known cell adhesion molecules produce a similar pattern of staining of medial edge epithelia. These new antibodies may identify novel proteins active during palatal shelf fusion. Certain lectins have been identified at the medial edge epithelium suggesting the presence of surface molecules rich in carbohydrate. In the chicken, which displays physiological cleft palate, the medial edge epithelia exhibit none of the sticky surface coat specialisations, do not form desmosomes and do not adhere to each other.

No sooner has the epithelial seam formed than it begins to thin. This is achieved by an increase in palatal height and a migration of epithelial cells onto the oral and nasal surfaces of the palate. Some of the epithelial seam cells undergo apoptosis while others are observed to migrate into the mesenchyme. Single palatal shelves have been shown to undergo differentiation, so this process of medial edge epithelium disruption is not dependent on shelf contact.

A series of experiments carried out by Ferguson and Honig, (1984), revealed that the signal for medial edge epithelium is specified by the underlying mesenchyme. Epithelial-mesenchymal recombination experiments using combinations of mandibular, limb and palatal tissues within and between mouse, chicken and alligator embryos were attempted. It was found that for the chicken and alligator the mesenchyme signals nasal, medial and oral epithelial differentiation in a species-specific fashion. The mouse exhibited similar results except that the mouse epithelium receives its signal to differentiate at the onset of palatogenesis and will continue to do so even if placed on a 'neutral' mesenchyme. The mouse mesenchyme can signal differentiation to epithelium of other species.

There are four possible ways in which the mesenchyme could signal epithelial differentiation; (a) extracellular matrix molecules, (b) soluble factors, (c) direct cell-cell contact and (d) a combination of a-c. Direct cell-cell contact does not occur between mesenchymal cells and epithelial cells as shown by electron microscopy, it therefore seems unlikely that direct cell-cell contact is involved in the signalling. An immunocytochemical survey of the developing palate revealed the presence of collagen types I-X, laminin, fibronectin, tenascin, heparan sulphate proteoglycan, chrondroitin 0, 4, 6, sulphates in patterns of expression that suggested these molecules were not involved in the signalling pathway. Collagen IX exhibited an interesting pattern of expression. It is absent from the mesenchymal cells of the developing palatal shelves at day 12 to day 13 of development, but just prior to palatal shelf elevation type IX collagen appears on the surface of the epithelial cells at the medial edge. This is the precise time of mesenchymal signalling. However it is not clear whether the appearance of type IX collagen is the mesenchymal-epithelial signal or whether it is actually a response to the signal.

The mesenchymal-epithelial signal may be a soluble growth factor, such as epidermal growth factor, transforming growth factor alpha (TGFα), transforming growth factor beta (TGFβ), acidic and basic fibroblast growth factors (FGF) and platelet-derived growth factor (PDGF), all of which are expressed during palate development. These growth factors have a plethora of conflicting effects on cell behaviour and on the extracellular matrix e.g. the synthesis of extracellular matrix molecules such as type IX collagen is stimulated by TGFα and TGFβ, while it is inhibited by FGF, all of which are known to be present in the developing palate. TGFα and PDGF stimulate cell division in mesenchymal cells, whereas TGFβ inhibits palatal mesenchymal cell growth and kills some cells. It should be noted that the same growth factor can stimulate cell division at one concentration and inhibit it at another; populations of cells respond differently to the same concentration of a growth factor depending upon their density and the nature of their substratum (Massague, 1987). No account has been taken of the presence of the receptors for the growth factors. These may exist in any pattern of distribution across the palatal shelves. So we have another layer of complexity with gradients of growth factors interacting with gradients of their respective receptors. The expression of soluble growth factors during palate development results in a complex pattern of interactions between the effects of the various growth factors, some of which will be synergistic and others antagonistic.

The different developmental mechanisms that may result in cleft palate are understood at the morphological level. The genes controlling these events are less easily defined mainly due to the numerous processes and pathways involved or implicated. The following molecules alone are all implicated: extracellular matrix molecules particularly hyaluronic acid, collagen molecules e.g. type IX, tenascin, soluble factors e.g. EGF, PDGF, TGFα, TGFβ, FGF, cytokeratins and cell surface molecules.

1.4 Genes whose disruption result in cleft palate

Recently it has been reported that TGFβ 3 plays a vital role in palatogenesis. A transgenic knockout mouse for TGFβ 3 has been created and the only phenotype observed was that of cleft palate (Mark Ferguson, personal communication). TGFβ 3 is expressed in many tissues during embryonic development, however it was expressed at very high levels at the tips of the palatal shelves prior to palatal shelf fusion. This TGFβ 3 knockout mouse has revealed a single gene whose disruption specifically results in cleft palate. How this gene's diruption impinges on palate development is not understood as yet, but provides a key signal from which relevant molecular interactions may be investigated.

Many of the cases of cleft palate observed in both mouse and human do not involve mutations of genes specifically involved in palatogenesis. Rather the cleft palate observed is a secondary effect of other mutated developmental pathways.

Recently several genes responsible for patterning of the head have been identified. Chisaka and Capecchi, (1991), produced direct evidence that Hox genes play an important role in the patterning of the head. They generated a germline mutation in the mouse *Hox-1.5* gene using homologous recombination to disrupt the endogenous gene in embryonic stem cells. Lufkin *et al.*, (1991), also generated germline mutations in a Hox gene, the *Hox-1.6* gene. Both mutations resulted in neonatal lethality of mice homozygous for the disrupted allele. Both of these mutations resulted in cleft secondary palate though they are not directly involved in palatogenesis. Mice homozygous for the *Hox-1.5* mutation exhibit a wide range of throat abnormalities, defects of the heart and arteries as well as craniofacial abnormalities.

The *Hox-1.6* mutant links *Hox* genes and the branchial region of the head. The abnormalities observed are localised to cells derived from rhombomeres 4-7 and their associated second and third branchial arches. The abnormal neural crest derivatives are confined to the neurogenic structures (Lufkin *et al.*, 1991). Many of the affected structures derive from interactive tissues and so it is not possible to differentiate between primary and secondary effects of gene disruption. However what is clear is the altered regional specification of the neural epithelium, leading to altered patterning in the rhombomeres and the subsequent alteration of patterning of the branchial arches. This results in the absence of certain cranial nerves and ganglia and multiple malformations of the inner ear and bones of the skull (Lufkin *et al.*, 1991).

The *Hox-1.5* homozygous mutant phenotypes are observed in a complex set of tissues largely confined to the head and thorax, and not all tissues that express the *Hox-1.5* gene are affected (Chisaka and Cappechi, 1991). The abnormalities in this case appear confined to a subset of the mesenchymal components, while the neurogenic derivatives appear completely normal. Cranial neural crest generates most of the mesenchyme for bone and connective tissue of the branchial arch. It is believed that crest migration takes place in two waves: the first phase comprising the mesenchymal components, and the second phase representing the neurogenic crest. Comparison of the results from the two mutant mice indicates that patterning of the subpopulations of the neural crest occurs before migration from the neural epithelium. In both cases the rhombomeres themselves appear morphologically normal suggesting that *Hox* genes have a role in specifying segmental identity rather than generating the individual segments. The disruption of *Hox-1.5* and *Hox-1.6* reveal how the developing palate is not only very susceptible to perturbations in genes required for palatal development but also to the disruption of genes not normally associated with the developing palate. Disruption of a particular gene may alter the patterning or a regulatory cascade during development so that phenotypes are observed in tissues not normally associated with that gene's expression domain. An alternative explanation is that cleft palate is a secondary event resulting from the abnormalities in craniofacial development in these animals.

These *Hox* genes have been shown to be fundamental to pattern formation in the developing cranial structures. They are by no means sufficient to explain the molecular mechanisms responsible for head development. A recent report has exposed the phenotypic effects of mutations in a structural protein on the developing palate. Mouse cartilage matrix deficiency (*cmd*) is an autosomal recessive mutation characterised by cleft palate, short limbs, tail and snout (Watanabe *et al.*, 1994). Biochemical and immunological studies have demonstrated a major reduction in the amount of aggrecan in the cartilage matrix of *cmd* mice, although normal levels of link protein and type II collagen were detected. Aggrecan, a proteoglycan, is one of the major structural components of cartilage whose function is to absorb shock and resist compression in the joints. It is extensively modified by substitution with chondroitin sulphate and keratin sulphate chains, and by other oligosaccharides. Watanabe *et al.*, (1994) reported the mapping of the aggrecan gene to the *cmd* locus and the identification of a mutation within the coding region of aggrecan.

The aggrecan gene contains 18 exons and was mapped to chromosome 7. A 7 base pair deletion was found in exon 5 of a genomic clone for the N-terminal portion of the *cmd* associated aggrecan gene. The deletion results in a truncated polypeptide which may affect the binding of hyaluronic acid, this coupled with possible interference in oligosaccharide binding may result in an unstable protein (Watanabe *et al.*, 1994). The cause of cleft palate in this mutant remains to be elucidated, though it is quite possible that the aggrecan mutation does not affect palate development directly. The presence of a shortened snout and a normal sized tongue would suggest a greatly reduced oronasal cavity during development. This smaller oronasal cavity may effect palatal shelf elevation resulting in cleft palate. Another possibility is that the binding of hyaluronic acid, which has been shown to be important in shelf elevation, is affected and this impinges on palatal shelf elevation, prevents shelf fusion and results in cleft palate.

An important feature of palate development is the epithelium-mesenchymal signal that induces palatal differentiation. It is believed that mutation of the genes that encode these epithelial-mesenchymal signals is involved in disruption of palate development. Recently, homeobox genes of the *Msx* class have been shown to be expressed in the epithelial or mesenchymal components of disparate tissues undergoing morphogenesis. In vertebrates, the expression of *Msx-1* and *Msx-2* is observed in many embryonic tissues in which epithelial-mesenchymal interactions occur during morphogenesis including the limb, the developing mandible and teeth (Hill *et al.*, 1989, MacKenzie *et al.*, 1991 and MacKenzie *et al.*, 1992).

In the vertebrate limb, the apical ectodermal ridge (AER) regulates the mesenchymal expression of *Msx-1* and *Msx-2*, as demonstrated by transplantation experiments (Robert *et al.*, 1991) and heterospecific grafting experiments (Brown *et al.*, 1993). *Msx-2* gene expression and skeletal differentiation of the first pharyngeal arch ectomesenchyme exhibit a strict requirement for an epithelial signal, suggesting that *Msx* expression is required for the formation of craniofacial bone. In the developing tooth bud, the dental epithelium is required to induce the underlying mesenchyme to differentiate into the dental follicle and dental papilla and also to activate *Msx-1* expression in the mesenchyme (Jowett *et al.*, 1993). It appears that

Msx-1 and *Msx-2* are primary mediators of inductive signals transmitted between epithelial and mesenchymal layers in these tissues.

Satokata and Maas (1994), have generated homozygous *Msx-1*$^{-}$ mice. They report that these mice die in the immediate postnatal period and exhibit marked abnormalities in craniofacial development, including cleft secondary palate, a failure of tooth and aveolar bone development in the mandible and maxilla, and abnormalities of the skull, malleus, nasal bones and conchae (Satokata and Maas, 1994). The abnormal dental development could be predicted from the pattern of gene expression (MacKenzie *et al.*, 1991, and MacKenzie *et al.*, 1992). But the disruption of palate development could not be predicted as *Msx-1* is normally absent from the developing palate. Also *Msx-1* has a widespread pattern of expression throughout the developing embryo yet it is predominantly the craniofacial structures that are affected. This suggests that *Msx-1* has a unique function in the developing head. The phenotypes observed in the *Msx-1*$^{-}$ mouse suggests that during limb development *Msx-1* expression is either not required or its absence can be compensated by the expression of other genes e.g. *Msx-2* . The abnormal phenotypes observed offer the possibility that *Msx-1* is a candidate gene for mutation in craniofacial abnormalities observed in humans especially cleft palate.

The palatal shelves do elevate in the *Msx-1*$^{-}$/*Msx-1*$^{-}$ mice but are developmentally delayed. This resembles the phenotype seen in many cases of human cleft palate. Although *Msx-1* is expressed at day 12 of embryonic development anteriorly in the developing palatal shelves, it is not expressed at the time of shelf elevation (day 13.5). The palatal shelves are adjacent to the alveolar maxillae and also adjacent to the dental follicle mesenchyme. In the *Msx-1*$^{-}$/*Msx-1*$^{-}$ mice these alveolar maxillae are deficient. Satokata and Maas (1994), propose that the absence of *Msx-1* expression in the dental follicle mesenchyme is related to impaired palatal development. This would suggest a link between the occurrence of cleft palate and abnormal tooth development. In fact 35% of individuals with isolated cleft palate exhibit abnormalities of their incisors and molars. Alternatively, the cleft palate in the *Msx-1* mice may arise because of a deficiency of the neural crest derived mesenchymal cells, as a result of problems with either neural crest cell proliferation or migration.

The abnormalities associated with mandibular growth and development could result in the pathological trapping of the palatal shelves between an abnormally large highly situated tongue on a small mandible. The *Msx-1*$^{-}$ mice have a small mandible and so it is possible that the cleft palate observed is secondary to this abnormality. However, the palatal shelves observed in the *Msx-1*$^{-}$ mice are raised but do not fuse. This suggests that palatal shelf elevation occurs but that the palatal shelves are deficient in some respect that prevents palatal fusion, this may be a result of delayed shelf elevation. But it is still possible that many human cases of cleft palate result from the disruption of a gene involved in mandibular growth and development.

As the developing palate is a very complicated developmental system, the genes described to date whose mutation results in cleft palate may impinge on palate development in many ways. TGFβ3 is the only candidate gene whose pattern of expression suggested that it may have a direct critical role in palate development. The TGFβ3 knockout mouse supports this, in that mice homozygous for the knockout mutation only exhibit cleft palate (Ferguson, personal communication). This mouse mutant may be a model for the human condition of cleft palate, though most cases of human isolated cleft palate probably occur as a result of multifactorial inheritance. However, several families have been documented that inherit cleft palate in a X-linked Mendelian fashion. Wilson *et al.*, (1993), have reported the generation of a transgenic mouse mutant that exhibits sex-linked cleft palate and may be a mouse model for the human heritable condition.

1.5 A transgenic mouse line displays sex-linked cleft palate

A series of transgenic mice were generated by microinjection of transgenic constructs designed to express the Latent Membrane Protein (LMP) from Epstein-

Barr Virus (EBV) (Wilson *et al.*, 1990). One of the transgenic mouse lines generated exhibited the unique phenotype of sex-linked cleft palate, none of the other LMP expressing mouse lines exhibited this phenotype. The transgenic line was designated PyLMP.5. All transgenically positive male offspring from PyLMP.5, died within 24 hours of birth. The survival of transgenically positive females was strain dependent. When positive females were bred into the mouse strain C57Bl/6, transgenically positive females died within the first two weeks post-natally. When the positive females were bred into the outbred strain CD1, a few females survived to maturity. The ill health of the adult females resulted in small, infrequent litters. In the fourth generation from the founder female the PyLMP.5 line became extinct. The founder female (F2 of C57Bl/6 x DBA2) may have been sufficiently mosaic or strain hybrid to avoid the severe mutant phenotype. 2/3 of the founders' offspring were transgenically negative but the transgene was transmitted to subsequent generations at the expected Mendelian ratio of 50% supporting the idea of a mosaic founder.

Of the transgenically positive progeny, new born males were found to have cleft secondary palates, while females surviving to maturity did not. Both transgenically positive male and female mice were 50% smaller than transgenically negative counterparts and had altered cranial morphology, with a shortened (retognathic) lower jaw, indicative of abnormal development of the mandible and maxillae (Wilson *et al.*, 1993).

Histological examination of the head of a neonatal male presenting cleft palate revealed that the only major difference between affected and unaffected animals was the presence of the cleft palate. In affected animals the palatal shelves had elevated and were horizontal, but much smaller than normal. They extended only a third of the width of the palate. The oral and medial edge epithelia of the affected palatal shelves were heavily keratinised, while in the anterior palatal region the mesenchyme contained condensations of pre-osseous tissue and blood vessels. The presence of these features is associated with the differentiation normally observed in the palatal mesenchyme of the mid-palatal suture and indicates that the palatal shelves have undergone a normal pattern of differentiation in the mutant. There were no abnormalities appparent in the snout, premaxillary region, or teeth (both incisors and molars).

Additional phenotypes were observed in this transgenic line. As females matured they began to display kinks in their tails and finally they assumed a hunched posture, due to spinal kinks. Sectioning of these kinks revealed that the bone was surrounded by an inflammatory cell infiltrate. The diagnosis was of cellulitis. None of the other LMP expressing transgenic lines displayed this bone deformation. The PyLMP.5 mouse did share other phenotypes with other LMP expressing lines such as retarded growth, delayed eye opening and sparse fur growth, but these were noted as being more severe in this transgenic line (Wilson *et al.*, 1990 and Wilson *et al.*, 1993).

1.6 Cloning of the *DXRiB 1* locus

The cellular sequences flanking the transgene were isolated and cloned. Wilson *et al.*, (1993), showed that the integrant existed as two copies of the transgene in a head to tail tandem array on the X chromosome. The cellular sequences flanking the transgene were cloned from a recombinant phage λ library constructed using the mutant genomic DNA. The cloned isolates consisted of 1.8kb and 1.2kb of cellular sequence bounded by *EcoR I* sites (used in the cloning), flanking the 5' and 3' ends of the transgene. When these cloned fragments were used as probes on Southern blots of genomic DNA from transgenically positive and transgenically negative mice, both were found to be unique sequence. The preintegration site was present as a single copy in transgenically negative males (a 3.0kb *EcoR I* fragment) and two copies in transgenically negative female. In transgenically positive males this single copy was disrupted (10.6kb *EcoR I* fragment), while one copy was disrupted in transgenically positive females and the other copy remained intact, thus demonstrating that the transgene had integrated into the X chromosome as predicted by the phenotype.

Both the isolated cellular fragments of DNA from the transgenic line hybridised to the 3.0kb *EcoR I* fragment, indicating that they are contiguous in the preintegration site. This also suggested that the transgene insertion was not accompanied by any gross rearrangements of cellular DNA. In this instance the absence of any gross rearrangements should aid the identification of the mutated gene(s). The phenotypes observed may result from the transgene influencing a cellular gene in a number of ways. Transgene integration may have physically disrupted a cellular locus. It may have altered the pattern of expression of the cellular locus by either perturbing the cellular locus' regulatory sequences or by the introduction of transgenic regulatory sequences. There is also the possibility that the splicing of the cellular locus is perturbed resulting in hybrid transcripts and possibly transient hybrid proteins.

The cloned flanking sequences were highly conserved across all mouse sub species tested and specific in rat, hamster and marmoset indicating high homology to these sequences. The cloned cellular flanking sequences were used as probes on Northern blots of post-natally derived RNAs. This analysis would identify if the cellular sequences were part of a transcription unit post-natally. No transcripts were identified in this analysis, so it would appear that none of the sequences is part of a transcript expressed after birth (Wilson *et al.*, 1993). This Northern analysis was repeated on a Northern blot of embryonic RNAs, again no messages were detected (unpublished). This suggests that the cellular sequences are not part of a transcript at any of the time points analysed and this was further substantiated by the complete sequencing of the sequences, no open reading frames (ORF) were found.

However, when a Northern blot of RNAs from transgenically positive tissues was probed with LMP sequences two hybrid LMP/cellular transcripts were detected. The 2.5kb and 0.8kb LMP specific transcripts were detected on a Northern blot of total RNAs from a day 10 post-natal transgenically positive female mouse, when probed with LMP sequences. These messages were expressed in the skin and tongue (they were also detected in a different PyLMP transgenic mouse line that had an autosomal site of transgene insertion).

In addition to these two LMP specific transcripts two larger transcripts were detected, unique to line PyLMP.5. Approximately 4.5kb and 10kb in size these transcripts were detected in all tissues analysed except the liver, therefore these transcripts have a pattern of expression completely different to that of the transgene sequences. Considering that the transgene is present at the site of insertion as a double copy totalling 7.6kb, the larger of the two transcripts must contain cellular sequence. This suggests that there are cellular sequences present at or near the site of transgene insertion capable of being expressed at the RNA level. No hybrid proteins were detected in skin samples (using anti-LMP antisera) in which the hybrid transcripts were detected (Wilson *et al.*, 1990). The presence of these hybrid transcripts that must contain cellular sequence and the lack of similar phenotypes in other LMP-1 expressing transgenic lines suggested that transgene integration had disrupted a cellular locus and thereby resulted in the sex-linked cleft palate phenotype.

However an understanding of the phenotypes that result from transgene integration at the *DXRiB 1* locus is complicated by transgene expression. Firstly, the transgene itself encodes a protein, LMP-1, that has been shown to have pleiotropic effects. LMP-1 has been shown to inhibit differentiation of human epithelium (Dawson *et al.*, 1990) and to cause morphological transformation of immortalised human keratinocytes (Fahraeus *et al.*, 1990). In addition to these phenotypic effects LMP-1 has been shown to induce the expression of *Bcl-2* (Henderson *et al.*, 1991). Bcl-2 is a proto-oncogene and its expression has been shown to promote the survival of B cell precursors and to block programmed cell death in B cells. The oncogenic activity of Bcl-2 is thought to result from its ability to promote cell survival rather than cell proliferation (Martin *et al.*, 1993). LMP-1 has been shown to induce changes in the profile of keratins expressed in the skin of mice and induce hyperplasia (Wilson *et al.*, 1990). The evidence to date is that LMP-1 induces a range of changes in the cell that affects its state of differentiation and proliferation, one of the cell types affected is epithelial cells.

In palate development the epithelium plays a crucial role in inducing and responding to changes in the mesenchymal cells it is in contact with. Due to the speed of palatal development the epithelium responds rapidly to the inducing signals. If anything changes the differentiation state of the cells in the developing palate then it may be expected that palate development itself may be effected. In the transgenic line the insertion of a LMP-1 expressing transgene resulted in the phenotype of sex-linked cleft palate. The fact that LMP-1 has been shown to induce changes in the differentiation state of several cell types raises the possibility that it is the LMP-1 expression and not a disrupted gene that results in cleft palate. Several other transgenic lines were created that expressed the LMP-1 transgene and none of these exhibited the mutant phenotype (Wilson *et al.*, 1990). However the transgenic line that exhibited cleft palate expressed LMP-1 at the highest level of all the transgenic lines but only in the skin. So the phenotype observed in the mutant line may be a result of redirected transgene expression due to the cellular locus.

What is not clear is if it is the LMP protein or disruption of a cellular protein that results in the observed phenotype. The minimalist interpretation is that the cellular locus causes the expression of LMP-1 in the developing palate or skull with the result that the phenotype is unique to this line. While the maximal interpretation is that LMP-1 expression is not involved and the presence of the transgene at this locus disrupts a cellular protein and this disruption alone results in the phenotype. There is also the possibility that the cleft palate phenotype results from a position between the minimalist and maximal interpretations, i.e. a combination of cellular disruption and transgene expression.

If the sex-linked phenotype does result from transgene expression it is still valuable to clone this locus. The transgene specific and hybrid transcripts exhibited different patterns of expression and the hybrid transcripts were not detected in any of the other LMP-1 expressing lines. This indicates that there are sequences present at the cellular locus directing the expression of these hybrid transcripts and possibly the LMP transcripts in development. Also the larger of the two hybrid transcripts was too large to be transgene sequence alone, it therefore must contain some cellular sequences. It is also quite possible that the phenotype of sex-linked cleft palate results from a combination of effects from the transgene site of insertion and transgene expression. However, all the observations from the mutant mouse indicate that there are important cellular sequences present at this locus. Both minimal and maximal interpretations indicate that there is a developmentally expressed gene at this locus, therefore it is relevant to clone this locus as all interpretations point to the presence of a gene. The cloning of this locus will allow some understanding of the sequences present, their functions and their involvement in the sex-linked cleft palate phenotype.

Unfortunately the transgenic line that exhibited sex-linked cleft palate became extinct and so it was not possible to investigate the role of LMP-1 expression in the sex-linked cleft palate phenotype. This problem can only be addressed by the regeneration of the transgenic mouse for the site of insertion with a different transgene. If these transgenic mice exhibit sex-linked cleft palate then it is the disrupted product of the cellular locus that resulted in the mutant phenotype and not LMP-1 expression. The generation of these transgenic mice is considered in the discussion.

As the 3kb of transgene flanking sequence consisted of unique sequence it could be used to map the site of insertion to known markers on the X chromosome. The cloned cellular DNA was given the locus map nomenclature *DXRib1* (DNA segment, X chromosome, Robertson Institute of Biotechnology, 1) and mapped by interspecific backcross analysis using progeny from the matings of [(C57B1/6J x *Mus spretus*)F1 X C57B1/6J] (Copeland and Jenkins, 1991). The mapping results indicated that *DXRib1* mapped to the proximal region of the X chromosome, closely linked to but separate from *Araf* and *DXPas3.*

As mentioned earlier cleft palate is a common developmental abnormality (both heritable and spontaneous) in the human population. The incidence of cleft palate is approximately 1/1500 live births, with little variation between races (Stanier *et al.*, 1993). Three families have been studied which display heritable sex-linked, non-syndromic cleft palate and it is quite possible that this disrupted *DXRib1* locus is the homologue of the locus affected in the human form of sex-linked cleft palate. However, the closest marker to *DXRib1* is *Araf* and *Araf* maps to Xp11 on the human X chromosome. The locus thought responsible for sex-linked cleft palate (CP+A) has recently been mapped to the other arm Xq21.1-Xq21.31 of the human X chromosome. The gene order of the mouse and human X chromosome are well conserved with breaks in blocks of synteny. Therefore, *DXRiB 1* may yet be the homologue of CP+A at a break in synteny, or alternatively be an entirely different locus. Ankyloglossia (tongue-tied) is often associated with the human X-linked cleft palate and is found in the three kindreds studied (Moore *et al.*, 1987; Rollnick and Kaye, 1986; Gorski *et al.*, 1992). The similarity in phenotypes between these families suggests that the same gene might be involved in each case with the possibility that failure of programmed cell death in the tongue is the primary site of the gene defect, with cleft palate as a secondary effect. The tongue-tied phenotype was not observed in the mouse mutant nor has it been observed in any mice which may be a due to of species difference. It is interesting to note that Stanier *et al.*, (1993), propose that the X-linked cleft palate is a secondary effect of gene dysfunction. The presence of the craniofacial abnormalities and spinal kinks in the transgenic line suggests that the cleft palate observed with this line, is possibly a secondary effect. Even if the gene disrupted in the transgenic line is not involved directly in palate development, it may still be a developmentally important gene with a vital role in craniofacial morphogenesis.

Given the observations suggesting the presence of important gene sequences at this locus and the strong possibility that transgene insertion had disrupted a cellular locus, the 3kb of unique DNA that flanked the transgene was used to screen a wild type mouse genomic DNA to isolate a larger region of genomic DNA from around the site of transgene insertion. This wild type genomic DNA was then characterised.

2 RESULTS:

2.1 Isolation of wild type mouse genomic DNA from around the site of transgene insertion

The first step in isolating the disrupted gene was to isolate a large fragment from the wild type locus. A recombinant Lambda phage library containing C57 Black/6, female mouse spleen DNA (Stratagene) was screened using the 3kb of cloned cellular sequences that flanked the transgene. Seven positive phage were identified hybridising to the flanking sequence probe. After difficulties in the first rounds of screening to obtain pure isolates of authentic recombinants, the mouse genomic library was rescreened using the transgene flanking sequence as the probe. The plaque purification of identified phage was continued. After twenty rounds of plaque purification two phage isolates from the first library screen were found to exhibit an expected restriction pattern upon digestion with *EcoR I*. Further recombinants were identified in a second screen that contained the transgene flanking regions of cellular DNA. Three of the identified phage from the two rounds of screening were further studied in detail to yield the restriction map of the wild type locus (Figure 3).

These were designated Phage 1, Phage 4 and Phage 7. Digestion of Phage DNA with *Not I* liberated the genomic DNA present in each recombinant phage DNA clone. When restriction digested with *Not I*, Phage 1 was found to contain 18.2kb of cellular DNA, Phage 4 contained 20.8kb of cellular DNA and Phage 7 contained 17.9kb of DNA. These cellular DNA fragments overlap to give 32.8kb of contiguous DNA from the wild type mouse X chromosome that surrounds the site of transgene insertion (Figure 3).

Each of the individual DNA fragments from the isolated phage was restriction mapped using enzymes with sites present in the multiple cloning site of the phage DNA (Figure 2). This information was used to produce a restriction map

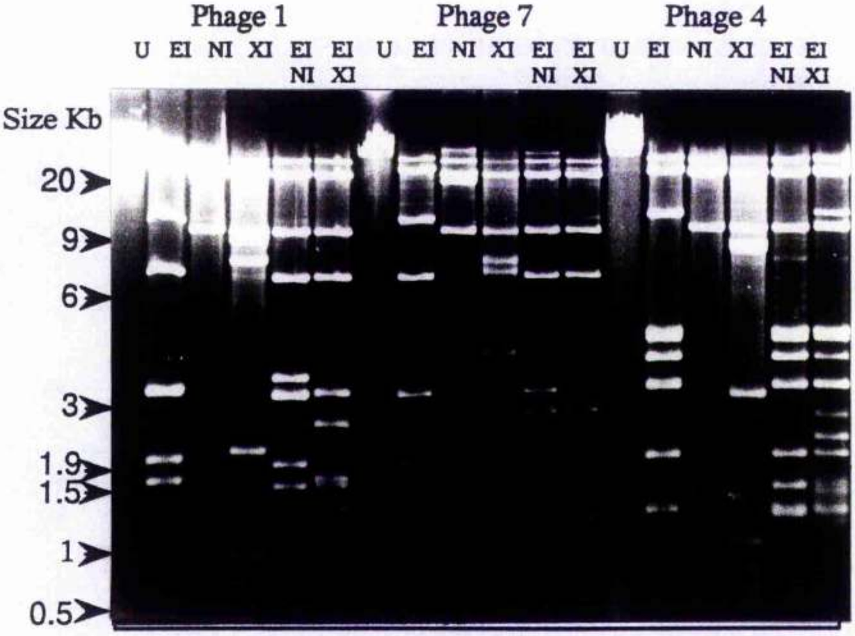

Phage 1
Phage 7
Phage 4
U EI NI XI EI NI EI XI U EI NI XI EI NI EI XI U EI NI XI EI NI EI XI
Size Kb
20
9
6
3
1.9
1.5
1
0.5

Figure 2:Restriction Digest analysis of isolated recombinant Phage DNA.

Phage DNA from the three isolated Phage was restriction digested with the enzymes indicated; U, undigested; EI, *EcoR I*; NI, *Not I*; XI, *Xba I*; EI NI, *EcoR I + Not I*; EI XI, *EcoR I + Not I*. The individual phage isolates are indicated above the enzymes used.

Note; the *EcoR I* digests of each phage isolate are represented in the Physical map of the locus Figure 3.

In the phage *EcoR I, Not I, Xba I* and double digests the 20kb and 9kb Phage arms are apparent. In the Phage 1 *EcoR I + Not I* double digest a 6kb, 3.2kb, 3kb, 1.9kb, 1.5kb and 0.7kb fragments are observed.

In the Phage 7 *EcoR I + Not I* double digest a 6kb, 3kb, 2.7kb, 1.9kb, 1.5kb, 1.3kb and 1.2kb fragments are observed. In the Phage 4 *EcoR I + Not I* double digest 4.1kb, 3.9kb, 3.5kb, 3.0kb, 1.9kb, 1.4kb, 1.3kb and 1.25kb fragments are observed.

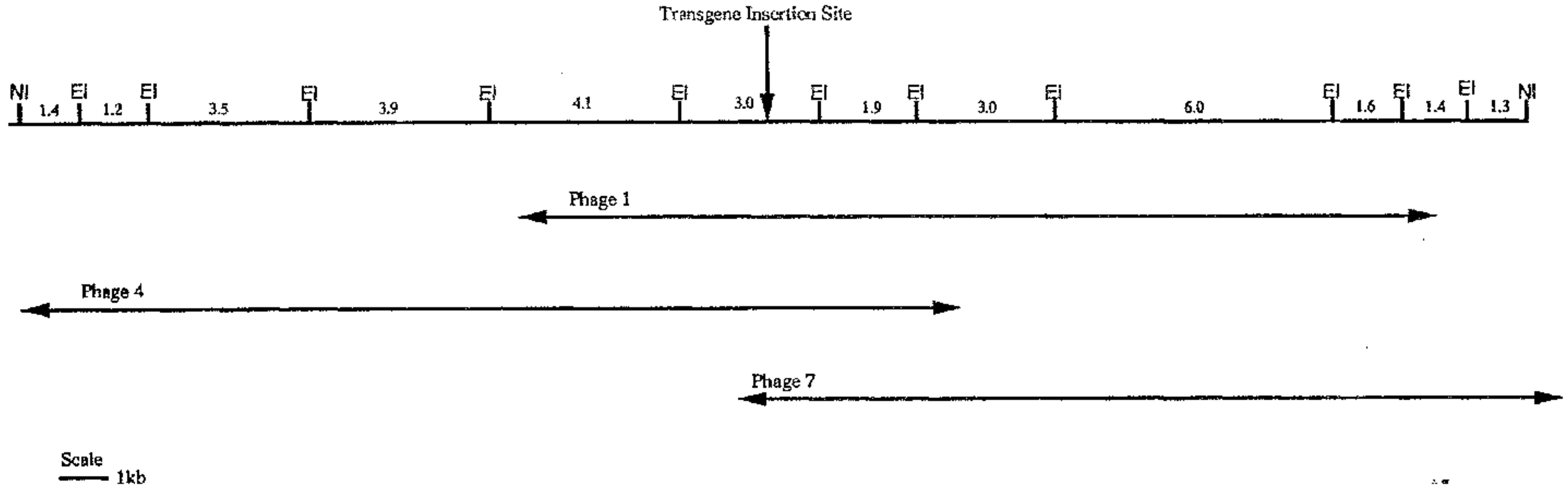


Figure 3: Restriction Map of DNA cloned from the wild type *DXRiB 1* locus. Restriction Map of the cloned DNA from the *DXRiB 1* locus. Sites shown are *EcoR I* (EI) and *Not I* (NI) sites. The positions of the recombinant phage that form the 32.8kb are shown below the restriction map. Site of transgene insertion is indicated by the plain arrow. The size of each fragment is given above the map.

spanning the entire 32.8kb of mouse genomic DNA isolated. The fragments generated by digestion with *EcoR I*, ranging in size from 0.7-6kb, were subcloned into the plasmid pBluescript (Materials and Methods).

Based on previous studies where genes have been isolated from large stretches of genomic DNA, the approach of using the subclones as probes on blots was taken. Noted in Goodfellow *et al.*, (1991), species conserved regions are generally associated with a gene or important motif. For these subclones, it was important to separate unique and repetitive sequences and evaluate whether they are expressed or not. Southern blots of restriction digested mouse genomic DNA were probed with the individual *EcoR I* fragments in order to identify regions that contained unique or repetitive sequences. These fragments were also used as probes on Northern blots of embryonic mouse total RNAs to evaluate whether a fragment of DNA contained transcribed sequence or not. The conservation of sequence across species was investigated by probing Southern blots of DNAs derived from a number of species, Zoo blots, with selected individual subcloned *EcoR I* fragments. Those DNA fragments that were conserved across species and expressed were sequenced.

2.2.1 Identification of unique sequence at the wild type locus

Unique sequence was identified at this locus as it is not possible to interpret the expression pattern from a DNA fragment containing repetitive sequence. Individual *EcoR I* fragments derived from the isolated phage were used as probes on Southern blots of mouse genomic DNA restriction digested with *EcoR I*. This analysis would reveal whether the DNA present was unique sequence or composed of repetitive sequence. Examples of the results analysis are shown in Figure 4 and summarised diagramatically in Figure 7. A large proportion of the subfragments from the 32.8kb of isolated DNA contained repetitive sequence. Several fragments of the isolated DNA consisted solely of unique sequence, including the original 3kb *EcoR I* fragment found in all phage isolates and used to screen the phage λ library. Some of the fragments containing repeated sequences are relatively large (e.g. the 6kb *EcoR I*

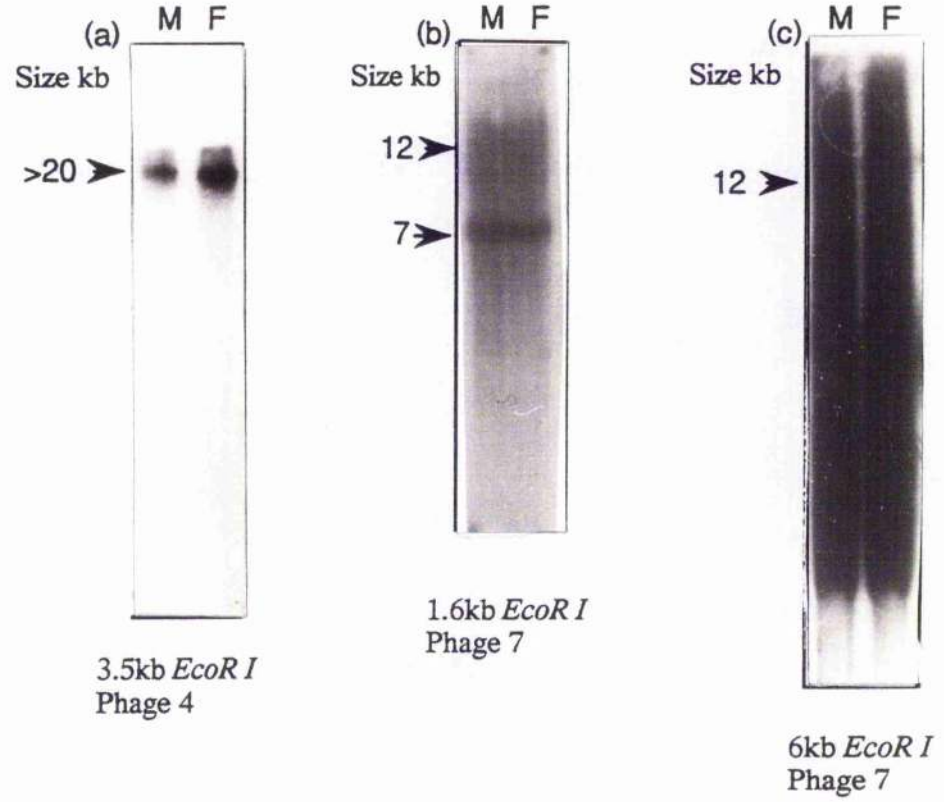


Figure 4: Southern blot analysis of mouse genomic DNA probed with *EcoR I* fragments from the isolated genomic clones.

3.3μg of mouse genomic DNA from male (M) and female (F) was digested with *EcoR V*, electrophoresed on an 0.7% agarose gel, Southern blotted and probed with radio-labelled *EcoR I* fragments from the isolated genomic clones spanning the *DXRiB 1* locus. The *EcoR I* fragments used to probe the Southern blots are shown with each blot shown. A summary of the Southern analysis for the entire 32.8kb of cloned DNA is presented in Figure 7.

fragment from Phage 1 and Phage 7) and probably include unique sequences within them. Therefore these fragments were further digested with other enzymes to generate smaller fragments and these were then used for the analysis.

In the case of the 6kb *EcoR I* fragment, it was digested with *BamH I* and *Hind III* yielding fragments of 2.2kb, 1.5kb, 1.2kb and 1.0kb in size which were subcloned into pBluescript (Materials and Methods). Southern analysis revealed that all of these smaller fragments still contained repetitive sequence, however the 1.2kb fragment appeared to hybridise to a unique band at about 6kb in size (data not shown).

Another fragment that contained highly repetitive sequence was a 3.0kb *EcoR I* fragment from Phage 1 and Phage 7. (It was possible to separate this 3.0kb fragment from the 3.0kb *EcoR I* fragment flanking the transgene in Phage 7, since the latter was incomplete in the phage being present at one end of the clone as a 2.7kb *EcoR I* + *Not I* fragment). The 3.0kb *EcoR I* fragment from Phage 7 was digested with *BamH I* and *Hind III*. This produced three fragments a 1.3kb *BamH I* + *Hind III* fragment, a 1.0kb *BamH I* + *Hind III* fragment and a 0.7kb *Hind III* fragment. Each of these fragments were also shown to contain highly repetitive sequences, however a predominant band was observable on the Southern blot probed with the 1.3kb fragment (data not shown).

Phage 4 *EcoR I* fragments 4.1kb and 3.9kb were composed of highly repetitive sequences. The subfragments generated by digestion with *Apa I* and *Sst I* of these fragments were used as probes on Southern blots. All of these were also found to be highly repetitive. As such, none of the subfragments that were derived from the large repetitive *EcoR I* fragments was found to contain unique sequence and these were not further restricted in an attempt to isolate unique sequences.

2.2.2 Reverse Southern analysis

An alternative approach to identify fragments harbouring repetitive sequences, Reverse Southern analysis, was used. In this method cloned DNA, (phage or plasmid), is probed with total, labelled, genomic DNA (Materials and Methods). Mouse genomic DNA is composed of repetitive sequences that are present at very high copy numbers, hence, when total genomic DNA is hybridised to a Southern blot it reveals those fragments that contain the most repetitive sequence. A Southern blot of all three phage digested with *EcoR I* was probed with total genomic DNA that had been radiolabelled by random priming. Figure 5 shows that the 6kb and the repetitive 3kb *EcoR I* fragments of Phage 1 and Phage 7 are highly repetitive. The 4.1kb *EcoR I* fragment of Phage 4 was shown to be highly repetitive too, confirming the earlier results (above). This technique also revealed that at least two fragments must be 3.0kb in size, the original unique fragment found flanking the transgene and a repetitive fragment, migrating together. The Reverse Southern analysis did not reveal the 3.9kb *EcoR I* fragment as being repetitive. This means that the 3.9kb *EcoR I* fragment, although containing highly repetitive sequences, contains sequences that are not widely dispersed throughout the mouse genome.

The Reverse Southern analysis was repeated on restriction digests of Phage 4, 4.1kb *EcoR I* fragment, the 6kb *EcoR I* (data not shown) and the 3kb *EcoR I* fragments shared between Phage 1 and Phage 7. All restriction fragments on these blots were found to be repetitive by Reverse Southern analysis. Figure 6 shows the results of Reverse Southern analysis of the Phage 1 and Phage 7, repetitive 3kb *EcoR I* fragment digests. Several different digests are shown in Figure 6. Three fragments were produced upon digestion with *BamH I* and *Hind III*, a 1.4kb, 1kb and a 0.7kb fragment. Reverse Southern analysis identified the 0.7kb fragment as the most highly repetitive. Reverse Southern analysis also identified the other two *BamH I* + *Hind III* subfragments, though the intensity of signal indicated that they were not as repetitive as the 0.7kb fragment.

2.3 Expressional analysis of isolated DNA

In order to identify transcribed sequences the *EcoR I* fragments isolated from the 32.8kb of DNA were used as probes on Northern blots of wild type and the

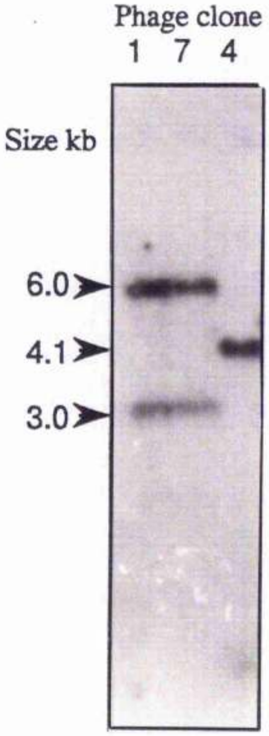


Figure 5: Reverse Southern analysis of *EcoR I* digested genomic clones.

1μg of genomic clones were digested with *EcoR I*, electrophoresed on an 0.7% agarose gel, Southern blotted and hybridised with radio-labelled total genomic DNA. The blot was washed at high stringency. Each lane represents *EcoR I* digests of the phage indicated.

The 3kb and 6kb *EcoR I* fragments from Phage 1 and Phage 7, along with the 4.1kb *EcoR I* fragment of Phage 4 hybridise to total genomic sequences.

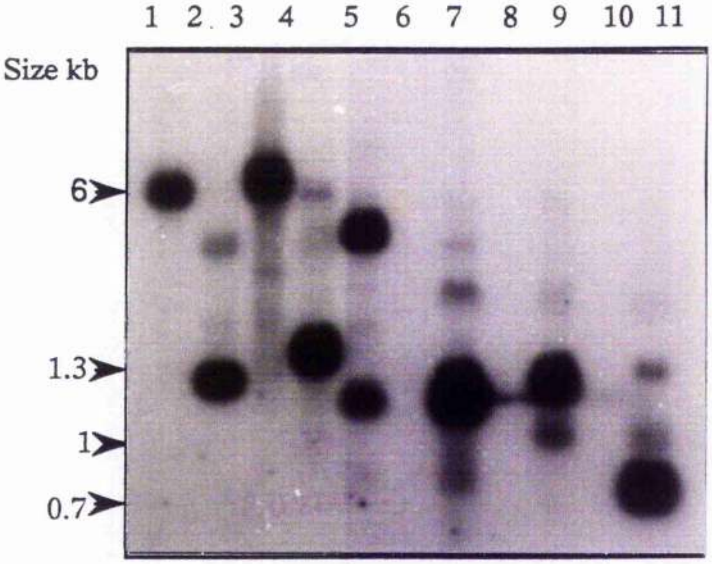
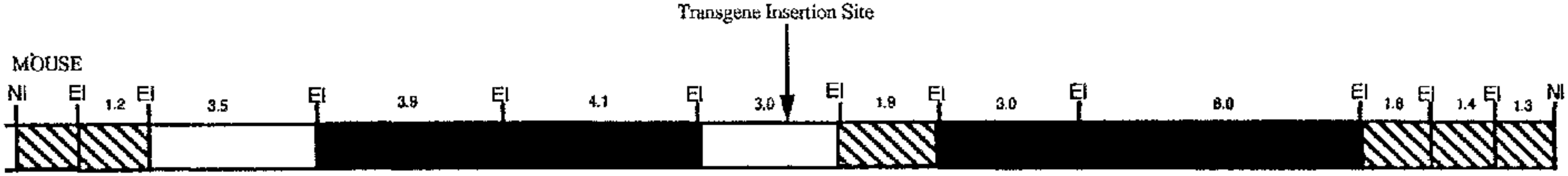


Figure 6: Reverse Southern analysis of the repetitive 3kb *EcoR I* fragment from Phage 1 and Phage 7.

1μg of the repetitive 3kb *EcoR I* fragment was digested with each of the enzymes indicated, lane 1, *AccI*; lane 2, *BamH I*; lane 3, *EcoR V*; lane 4, *Hind III*; lane 5, *Sst I*; lane 7, *Acc I* + *Sst I*; lane 9, *BamH I* + *EcoR V*; lane 11, *BamH I* + *Hind III*. Samples were electrophoresed on an 0.7% agarose gel, Southern blotted and hybridised with radio-labelled total genomic DNA. The blot shown was washed at high stringency.

All fragments hybridise to genomic sequences, the 0.7kb *BamH I-Hind III* fragment hybridises most intensely in lane 11.

SOUTHERN ANALYSIS OF CLONED DNA



Key:

■ Highly Repetitive Sequence

□ Unique Sequence

▨ Repetitive Sequence but band observable

Figure 7: Summary of Southern Analysis for the 32.8kb of DNA isolated from around the transgene site of insertion. The key details the type of DNA tested for each fragment isolated from the locus. The site of transgene insertion is indicated by the arrow. The size of each *EcoR I* fragment is given above it.

transgenic mutant RNA. A transcript that is involved in the phenotype at this locus should be disrupted in its pattern of expression between transgenically positive RNA and transgenically negative RNA. The hybrid transcripts detected by transgene probing were most highly expressed in the brain. It was therefore decided to investigate the expression of these sequences in RNA prepared from transgenically positive and transgenically negative mouse brains.

All of the *EcoR I* subclones were used as probes on Northern blots of RNA from transgenically positive and transgenically negative day 10 post-natal brains. Figure 8 shows the result of one of these hybridisations. The blot shown is that probed with the repetitive 3kb *EcoR I* fragment from Phage 1 and Phage 7. A very large message is observed, significantly larger than 10kb, in both transgenically positive and transgenically negative samples. At this resolution it is not possible to tell if there is a change in size of the transcripts between transgenically positive and negative samples. However, the hybrid transcripts previously identified in the transgenically positive RNA were not seen in this analysis.

Several other fragments hybridised to large transcript(s). The 6kb and 1.4kb *EcoR I* fragments from Phage 1 and Phage 7, along with the *EcoR I* + *Not I* 0.7kb fragment from Phage 1 and the *EcoR I* + *Not I* 1.3kb fragment from Phage 7, all exhibited this pattern of expression. This was also true of the 4.1kb and the 3.9kb fragments from Phage 4. However all of these were shown to harbour repetitive sequences. No expression could be detected from any of the unique DNA fragments.

The phenotype caused by the insertional mutation of cellular sequences results from a perturbation in craniofacial development, thus the cellular sequences should be expressed during embryonic development. The transgenic mutant mouse line that exhibited sex linked cleft-palate unfortunately became extinct. As a result no transgenically positive embryonic RNA was available for analysis. However, wild type RNA was prepared from whole mouse embryos at different stages of development and used to analyse the patterns of expression exhibited from the locus.

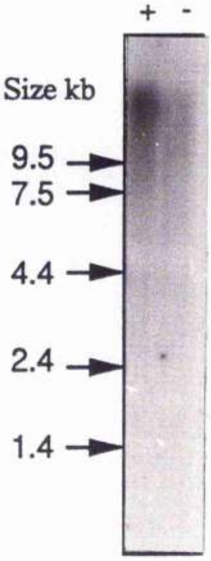


Figure 8: Expression analysis of isolated genomic DNA from the *DXRiB 1* locus.

10µg samples of total RNA isolated from the brains of line PyLMP.5 transgenically positive (+) and transgenically negative (-) day 13 post-natal mice, were electrophoresed on a 1% agarose formaldehyde gel. This gel was Northern blotted and probed with radio-labelled *EcoR I* fragments from the isolated genomic cloned DNA.

The blot shown was probed with the repetitive 3kb *EcoR I* fragment from Phage 7. It was found to hybridise to RNA transcripts greater than 10kb in size in both transgenically positive and transgenically negative brain RNAs.

The results of the analysis are summarised in Figure 11.

All the *EcoR I* fragments derived from the isolated 32.8kb of RNA were used as probes on Northern blots of mouse embryonic total RNA. RNAs were derived from the body and head of day 11.5 to day 16.5 embryonic mice (the total gestation period for the mouse is 20 days). A representative result of this expression analysis is presented in Figure 9. The repetitive 3kb *EcoR I* fragment from Phage 1 and Phage 7 was found to hybridise at high levels to non-discrete messages, that appear to increase in the level of expression during development. It was found that from Phage 1 and Phage 7 the 6kb *EcoR I* fragment was also expressed at very high levels. A fragment from Phage 7, a 1.3kb *EcoR I* fragment, which was shown to contain some repetitve sequence but was not completely repetitive, was also found to exhibit a similar pattern observed with the repetitive fragments. The transcripts observed were not discrete for all three fragments. Rather they were very large >10kb and probably as a result of RNA degradation, are very smeared. The fact that the 6kb fragment and the 3kb *EcoR I* fragments contain highly repetitive sequences the messages detected may be derived from numerous different unrelated genes which also harbour the repeats. The Northern blots were reprobed with the smaller subcloned fragments from the 6kb and the 3kb *EcoR I* fragments. The patterns of expression were identical to those observed with the original entire fragments (data not shown).

The 4.1kb *EcoR I* and the 3.9kb *EcoR I* fragments of Phage 4 were found to hybridise to embryonic RNAs. The transcripts observed were very large, >10kb, and apparently developmentally regulated. These Phage 4 fragments reproduced the patterns of expression observed with the Phage 7 fragments. Again, the DNA fragments from Phage 4 that exhibit expression during development are those identified as being the most highly repetitive. Expression was only detected with fragments that were shown to contain repetitive sequences, none of the fragments found to contain solely unique sequences demonstrated expression. The presence of repetitive sequences makes the interpretation of the analysis of expression very

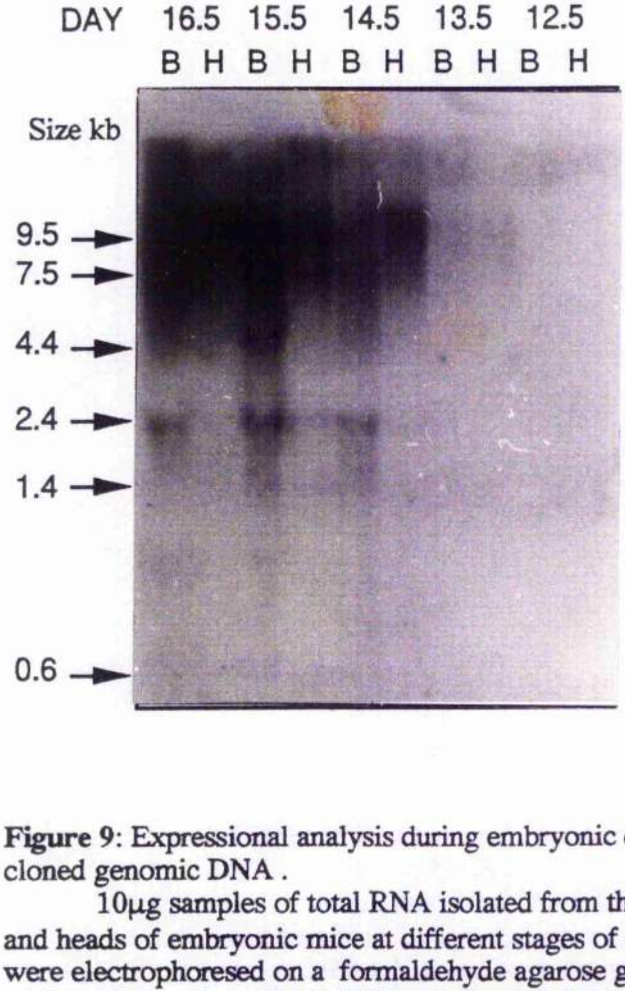


Figure 9: Expressional analysis during embryonic development of cloned genomic DNA .

10µg samples of total RNA isolated from the whole body and heads of embryonic mice at different stages of development were electrophoresed on a formaldehyde agarose gel. The day of embryonic development is indicated above the samples, 'B' indicates whole body RNA and 'H' indicates whole head RNA. This gel was Northern blotted and probed with the radio-labelled repetitive 3kb *EcoR I* fragment from the isolated Phage 1 and Phage 7 cloned genomic DNA.

difficult. Those fragments hybridising to the post-natal RNAs were also found to hybridise to the embryonic RNAs

2.3.1 Reverse Northern analysis

Since the patterns of expression observed by Northern analysis of embryonic RNAs were difficult to interpret Reverse Northern analysis was attempted. Consonant with the technique of Reverse Southern analysis, Reverse Northerns involve the probing of Southern blots, though instead of genomic DNA as the probe, labelled RNA is used (Materials and Methods). Only highly transcribed sequences would be detected using this method. The labelling of the RNA is actually first strand cDNA synthesis, incorporating radiolabelled nucleotides, primed by either oligo [dT] or random primers. Here, first strand cDNA synthesis of day 15.5 embryonic RNA was primed using oligo [dT]. This produced a cDNA probe from the Poly $[A]^+$ fraction which was used against a Southern blot of Phage DNAs digested with *EcoR I*.

The results of the Reverse Northern analysis are presented in Figure 10. The cDNA probe hybridises to the 6kb *EcoR I* fragment and the repetitive 3kb *EcoR I* fragment from Phage 1 and Phage 7. It also hybridised to the 4.1kb *EcoR I* fragment from Phage 4. On longer exposures of the blot the 1.4kb *EcoR I* fragment that was shown to contain some repetitive sequences from Phage 4 is found to hybridise to transcripts present in the labelled cDNA. The 1.9kb *EcoR I* fragment from Phage 1 and 7 is also seen to hybridise with labelled cDNA, but the 1.9kb *EcoR I* fragment from Phage 4 does not appear to hybridise to the cDNA. From restriction mapping it is believed that each Phage isolate shared the 1.9kb *EcoR I* fragment and hybridisation of *EcoR I* fragments from one phage to another confirmed this observation. This suggests that the lack of signal is an artifact due to unequal loading between samples. The Reverse Northern analysis was repeated using RNA radiolabelled by random priming. The results obtained using this

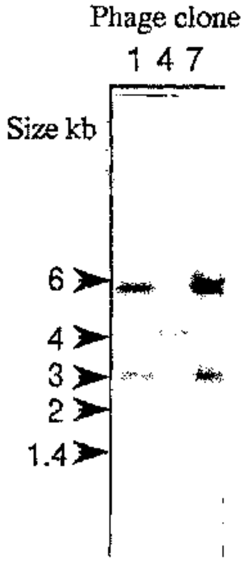


Figure 10: Reverse Northern analysis of *EcoR I* digested genomic clones.

1µg of isolated Phage genomic clones were digested with *EcoR I*, electrophoresed on an 0.7% agarose gel, Southern blotted and hybridised with reverse transcribed radio-labelled embryonic day 15.5 total RNA primed with oligo [dT]. Wash stringency of the blot shown was high; the phage clone is indicated for each track.

The 3kb and 6kb *EcoR I* fragments from Phage 1 and Phage 7, along with the 4.1kb *EcoR I* fragment of Phage 4 are detected. A 1.9kb *EcoR I* fragment from Phage 1 and 7 and a 1.4kb *EcoR I* fragment from Phage 4 are detectable on longer exposures of this blot.

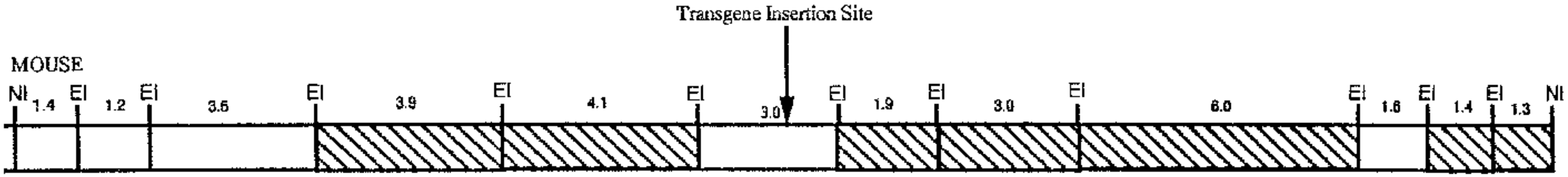
RNA probe were similar to those using the oligo [dT] primed RNA probe (data not shown).

There are differences between the results of the Northern analysis and the Reverse Northern analysis. Several of the fragments identified by Northern analysis as being expressed, e.g. the 0.7kb *EcoR I* + *Not I* fragment of Phage 1, are not identified by the Reverse Northern analysis. This is a reflection of the difference in sensitivity of the two methods. No hybridisation was observed to the unique fragments. The results of the Reverse Southern analysis and the results of the Reverse Northern analysis confirm that the most highly repetitive sequences are those that hybridise to expressed sequences. These expressed sequences may not be derived from this locus and are hybridising to expressed repetitive sequences from other genes. As only the repetitive sequences hybridise to transcripts, these analyses do not reveal whether a transcript is expressed from this locus or not. All of the information obtained from the previous Southern and Northern analyses are presented in the summary Figures 7 (for Southerns) and Figure 11 (for Northerns).

2.3.2 Analysis of differences in expression between transgenically positive and transgenically negative tissues

It was impossible to analyse the patterns of expression of the locus during mutant embryonic development due to lack of transgenically positive embryonic RNA. However post-natal tissues were available from transgenically positive mice. The large transcript(s) hybridising to the repetitive 3kb *EcoR I* fragment from Phage 7 was examined in post-natal RNAs from transgenically positive and transgenically negative tissues (Figure 12). There are no bands observable in most of these samples. A band is observed in the lung samples and there is a clear difference between the transgenically (+) and transgenically (-) lung samples. In transgenically negative lung RNA a band is observable at 4kb, this signal is not present in the transgenically positive RNA sample. A difference observed between both the lung samples and the other RNA samples was

ANALYSIS OF EXPRESSION OF CLONED DNA



Scale

—— 1kb

Key:

▨ Expressed Sequences

□ Non-Expressed Sequences

Figure 11: Northern Analysis of DNA cloned from around the site of transgene insertion. The figure summarises the results of expression analysis using embryonic and post-natal RNAs. It also includes information from the Reverse Northern analysis.The key indicates the difference between expressed and non-expressed sequences.

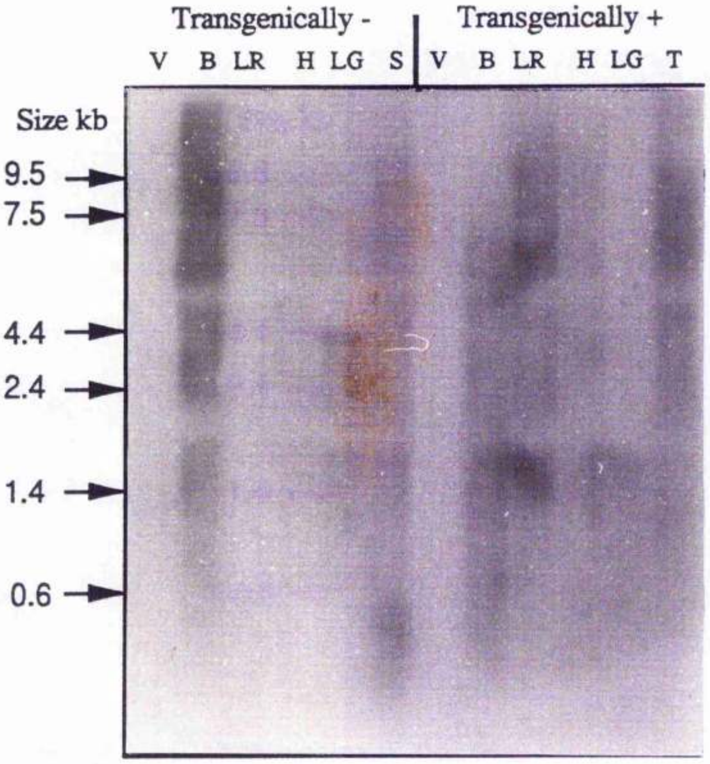


Figure 12: Analysis of expression between post-natal transgenically positive and transgenically negative tissues.

10µg samples of total RNA isolated from the tissues of age matched transgenically positive (+) and transgenically negative (-) tissues. Samples are V, viscera; B, brain; LR, liver; H, heart; LG, lung; S, spleen and T, thymus. Samples were electrophoresed on a 1% formaldehyde agarose gel. This gel was Northern blotted and probed with the radio-labelled repetitive 3kb *EcoR I* fragment from the isolated Phage 1 and Phage 7 cloned genomic DNA.

Note: a band is present in the transgenically negative lung sample running at about 4kb that is not observed in the transgenically positive lung sample.

a lack of background smears. This may be due to the lung samples being more intact than the other samples, although visualisation of the samples by Ethidium Bromide staining revealed that all samples appeared intact. The Ethidium Bromide staining of the gel indicated loading was even, apart from the brain RNA samples. The presence of repetitive sequence makes interpretation of the blot very difficult.

2.4 Analysis of sequence conservation across species

A feature of DNA encoding important sequence, ie. sequence that forms part of a functional gene product, is that it usually exhibits a degree of conservation between DNA from different species. The more necessary the function of a protein domain for protein activity then a higher degree of conservation of the sequence that encodes it is usually observed. The degree of sequence conservation of the isolated unique and expressed DNA fragments was investigated by hybridising these to Southern blots of DNAs derived from different species, referred to as Zoo blots. The Zoo blots used in this analysis consisted of DNAs derived from mouse, rat, hamster and human restriction usually digested with *EcoR I* but also *EcoR V*. The hybridisation conditions were made less stringent (by lowering the temperature and increasing the salt concentration) to allow sequences with some divergence from each other to hybridise. The washing conditions used, subsequent to the hybridisation, were very much less stringent than normal. The blots were then exposed and washed again, either decreasing the salt concentration of the wash or increasing the temperature of the wash and on occasion both. These rounds of washes with increasing stringency, resulted in blots that first displayed a pattern of non specific hybridisation. This pattern was replaced gradually by a more specific pattern of hybridisation. Hence the more conserved the sequence the better the ability to remain hybridised with the increasing stringency of the washes.

Some of the results of the Zoo blot analysis can be seen in Figures 13 to 18. The fragments of isolated DNA studied in depth were those that exhibited patterns of expression as detected by Northern blot analysis. Figures 13(a) and 13(b) show the

Figure 13 (a): Evolutionary conservation of the *Apa I +Sst I* DNA subfragments of the 4.1kb *EcoR I* DNA fragment from Phage 4.

4μg of genomic DNAs (listed below) were digested with *EcoR I,* electrophoresed, Southern blotted and hybridised with (i) the 0.7kb *Apa I+ Sst I* subfragment, (ii) the 0.8kb *Apa I+Sst I* subfragment, (iii) the 1.3kb *Apa I+Sst I* and (iv) the 1.4kb *Apa I+Sst I* subfragment. Wash stringency of the blots shown are 1xSSC, 0.1% SDS, 55°C.

DNAs were derived from the tissues of; lane 1, mouse female; lane 2, mouse, male; lane 3, rat; lane 4, hamster and lane 5, human.

Note that the fragments are highly repetitive in the mouse but bands are apparent in some of the other species DNA samples indicating the presence of unique sequence.

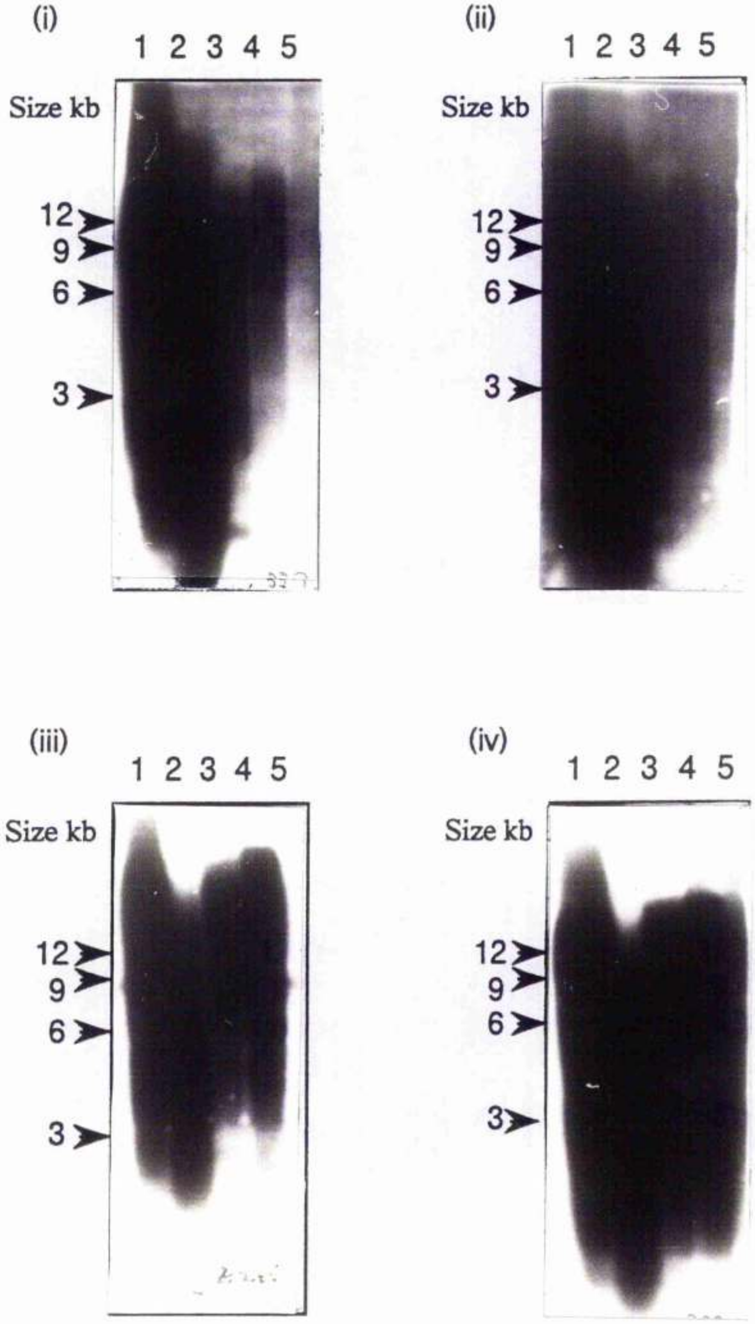


Figure 13(a): Evolutionary conservation of *Apa I*+ *Sst I* subfragments of the 4.1kb *EcoR I* fragment from Phage 4 washed at low stringency.

Figure 13 (b): Evolutionary conservation of the *Apa I+Sst I* DNA subfragments of the 4.1kb *EcoR I* DNA fragment from Phage 4.

4μg of genomic DNAs (listed below) were digested with *EcoR I*, electrophoresed, Southern blotted and hybridised with (i) the 0.7kb *Apa I+ Sst I* subfragment, (ii) the 0.8kb *Apa I+Sst I* subfragment, (iii) the 1.3kb *Apa I+Sst I* and (iv) the 1.4kb *Apa I+Sst I* subfragment. Wash stringency of the blots shown are 1xSSC, 0.1% SDS, 65°C.

DNAs were derived from the tissues of; lane 1, mouse female; lane 2, mouse, male; lane 3, rat; lane 4, hamster and lane 5, human.

Note that compared to the previous low stringency wash (figure 13(a)) unique sequence is apparent in the human DNA samples, indicating the conservation of sequence across species at medium stringency.

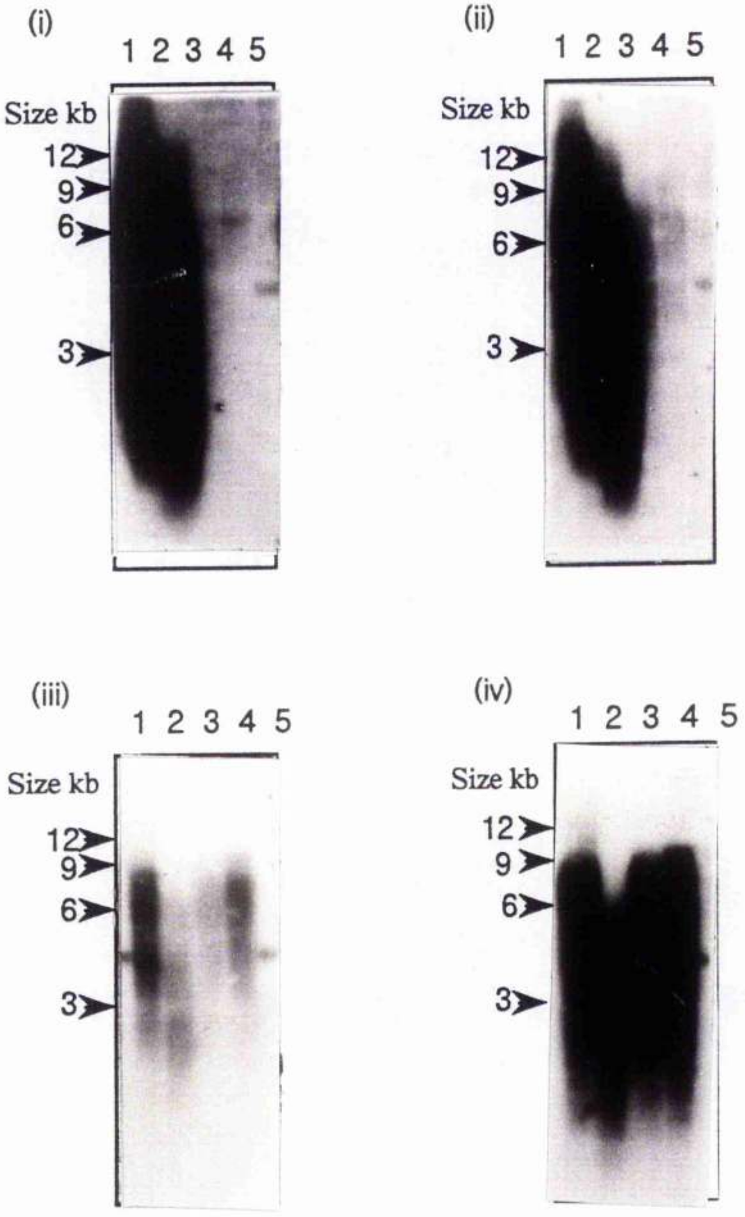


Figure 13(b): Evolutionary conservation of *Apa I+ Sst I* subfragments of the 4.1kb *EcoR I* fragment isolated from Phage 4. Blots were washed at medium stringency.

Zooblot analysis for the Phage 4, 4.1kb *EcoR I* fragment: *Apa I+Sst I* subfragments. At low stringency (55°C, 1xSSC) the highly repetitive nature of the *Apa I+Sst I* fragments derived from Phage 4, 4.1kb *EcoR I* fragment is obvious (Figure 13(a)). For each subfragment the sequence is highly repetitive in mouse, rat and hamster DNA. The human DNA does not appear to hybridise the mouse sequences as well as DNA from other species and does not give as repetitive a signal. The Phage 4, 4.1kb *EcoR I* fragment, 1.3kb *Apa I+ Sst I* subfragment hybridises to a discrete 3kb band in human DNA at low stringency (55°C and 1xSSC).

When increasing the stringency by decreasing the salt concentration to 0.1xSSC but maintaining the temperature of the wash at 55°C more information can be obtained from the blots. With the 0.7kb, the 0.6kb and the 1.4kb *Apa I + Sst I* subfragments the pattern of hybridisation observed with mouse DNA is highly repetitive, (Figure 13(b)). Discrete bands are however observable in the tracks of other species DNA. A band is observed in human DNA with all three subfragments and a hamster signal observed with the 0.7kb and the 0.6kb *Apa I + Sst I* subfragments. The strong repetitive signal of the female mouse track obscures the rat DNA track and it is difficult to determine whether a signal is present or not.

The 1.4kb *Apa I + Sst I* subfragment from the Phage 4, 4.1kb*EcoR I* fragment remains repetitive across all species except human where a faint band is observable at about 3kb (as with the 1.3kb subfragment). The signal observed with the 1.3kb *Apa I + Sst I* subfragment is much less repetitive at this stringency of 0.1xSSC. A band is clearly apparent in male mouse DNA, this band is not obvious in the female mouse DNA sample. This is due to the female DNA being slightly degraded. Rat DNA does not contain any bands at all and there may be a band present in hamster DNA. However there clearly is a band present in the human DNA sample.

When the temperature was increased to 68°C and the salt concentration increased to 1xSSC the pattern of hybridisation for all the subfragments of the Phage 4, 4.1kb *EcoR I* fragment remained the same (data not shown). The stringency was increased to high (68°C and 0.1xSSC), and the blots exposed for three days. The 1.4kb *Apa I* + *Sst I* subfragment was found to be still repetitive across mouse, rat and hamster, with no bands detected. The band in human DNA was still observable. These results indicate that the 1.4kb *Apa I* + *Sst I* subfragment is highly repetitive in mouse and this repetitive sequence is conserved in the genomes of rat and hamster. This repetitive sequence is not present in human DNA, but some unique sequence clothed in repetitive sequence in the mouse genome is conserved into human DNA (data not shown).

The 0.6kb and the 0.7kb *Apa I* + *Sst I* subfragments exhibited a similar pattern of hybridisation when the blots were washed at high stringency. The sequences are highly repetitive in the mouse where no bands are observed at all, just two black smears are present in the mouse DNA lanes. With both subfragments a band is observed in the human DNA lane, it migrates at 6kb in size. A faint band of about 8kb was observed in the hamster DNA sample with the 0.6Kb subfragment. These results reveal that the 0.7kb and 0.6kb *Apa I* + *Sst I* subfragments are highly repetitive in mouse and these repetitive sequences are not conserved in other species. As was observed with the 1.4kb *Apa I* + *Sst I* subfragment, there are sequences present in 0.6kb and 0.7kb *Apa I* + *Sst I* subfragments which are conserved in human.

The hybridisation pattern of the Zoo blot probed with the 1.3kb *Apa I* + *Sst I* subfragment from 4.1kb *EcoR I* fragment of Phage 4, after washing at high stringency, reveals that this fragment does not contain as repetitive a sequence as the other subfragments from the Phage 4, 4.1kb *EcoR I* fragment. A well defined band is present at 4.1kb in the male mouse DNA lane. This result confirms that the subfragment is derived from a 4.1kb *EcoR I* fragment. Although the repetitive nature of this subfragment is somewhat less than that of other subfragments, no obvious bands are present in the DNA samples of the other species, except the human sample where a band just larger than the 4.1kb fragment is evident. The ability of sequences present in the Phage 4, 4.1kb *EcoR I* fragment to hybridise to a discrete fragment present in human DNA, but not to clearly defined sequences in other species may suggest that the conserved sequence may be of importance in the human but not in rat or hamster. This is unlikely due to the higher conservation between rodents, than conservation between rodents and humans. Alternatively the size of the band in rat and hamster was either too large or too small to give good signal. The possibility of contamination of the human DNA sample has not been rigorously eliminated.

Using human DNA from different sources and digesting with *EcoR V* no homologous bands were evident. Figure 14 shows that when the Phage 4, 4.1kb *EcoR I*, 1.3kb *Apa I+ Sst I* subfragment was used to probe this Zoo blot a 20kb fragment was identified in all mouse samples. The signal is quite faint as the fragment identified is large and may not have transfered fully during blotting. When the blot was washed at intermediate stringency (68°C and 0.5xSSC) a band was observed in the rat DNA lane, a band was also observed in the hamster lane indicating a degree of conservation. This supported the idea that the lack of hybridisation with the previous Zoo analysis was a result of poor signal rather than a lack of conservation. When the blot was washed at 68°C and 0.1xSSC these bands were no longer seen. It is difficult to decide whether there are any bands observed in the human samples, at any of the washing conditions. There is a smear at the top of the blot suggesting that a large homologous fragment may be present but it is difficult to tell whether it is a real signal or not. These results taken together with the previous Zoo analysis suggest that the 1.3kb fragment is moderately conserved across species. Further digests with different enzymes would clarify this situation.

The Zoo blot analysis was repeated for other fragments derived from the isolated 32.8kb of mouse genomic DNA. The other fragments that produced signals on Northern and Reverse Northern Blots were from Phage 1, Phage 4 and Phage 7. The results of the Zoo analysis of the 6kb *EcoR1* fragment, the repetitive 3kb *EcoR I* fragment from Phage 7 and the 3.9kb *EcoR I* fragment of Phage 4 at low stringency are shown in Figure 15(a). These fragments were further restriction

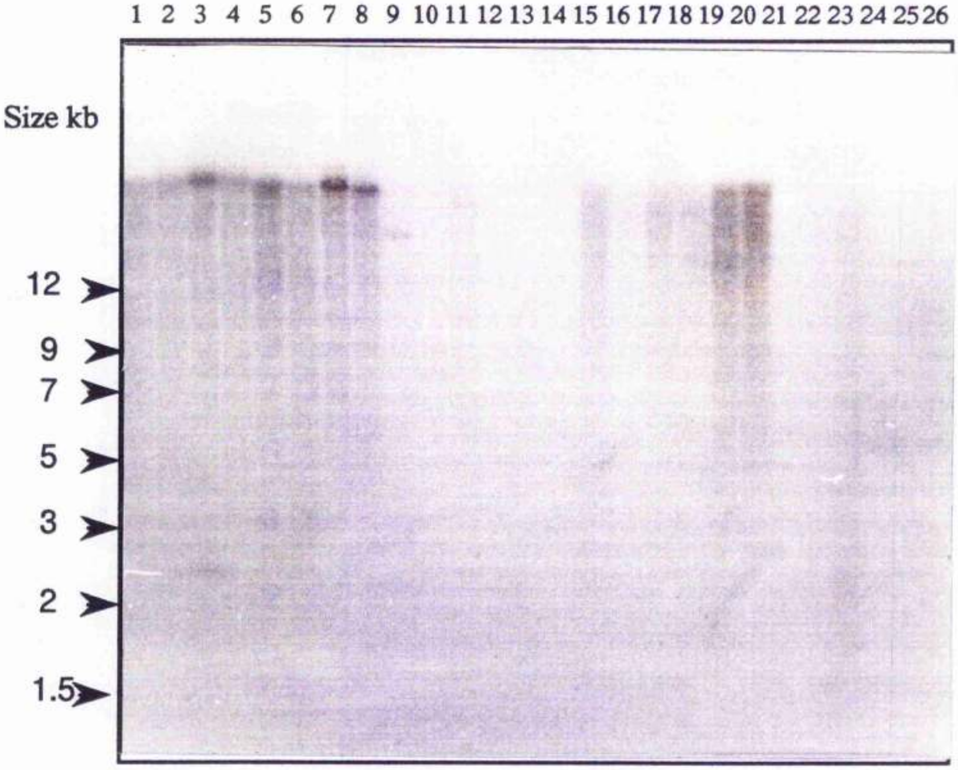


Figure 14: Evolutionary conservation of the 1.3kb *Apa I+ Sst I* subfragment of the 4.1kb *EcoR I* fragment from Phage 4.

4μg of genomic DNAs (listed below) were digested with *EcoR V*, electrophoresed, Southern blotted and hybridised with the 1.3kb *Apa I + Sst I* subfragment from the 4.1kb *EcoR I* fragment from Phage 4. Wash stringency of the blot shown was 0.5xSSC, 0.1% SDS, 65°C.

DNAs were derived from the tissues of; lane 1, ***mus domesticus***, female; lane 2, ***mus domesticus***, male; lane 3, ***mus spretus***, female; lane 4, ***mus spretus***, male; lane 5, ***mus macedonicus***, female; lane 6, ***mus macedonicus***, male; lane 7, ***mus hortulanus***, female; lane 8, ***mus cookii***, male, lane 9, rat; lane 10 & 11 hamster; lane 12, dog; lane 13, rabbit; lane 14, cow; lane 15 & 16, primate; lane 17, 18, 19 & 20, human; lane 21, chicken; lane 22, axolotl; lane 23, salmon; lane 24, drosophila; lane 25, *C. elegans*; lane 26, *S. cerevisiae*.

digested and used in the Zooblot analysis, but all the fragments and subfragments were found to be repetitive in the Zoo analysis.

The 6kb *EcoR I* fragment was restriction digested with *BamH I* and *Hind III* producing several fragments, 2.2kb, 1.5kb, 1.2kb and 1.0kb. These fragments were used to probe Zoo blots containing mouse, rat, hamster, marmoset and human DNAs. (The probes were subject to 'quashing' prior to addition to hybridisation buffer, see below). Figure 15(b) shows examples of the results. At low stringency (55°C and 1xSSC) the 2.2kb and the 1.0kb *BamH I* + *Hind III* fragments were found to be highly repetitive but only in the mouse. The other DNAs exhibited a very low background of hybridisation but no clearly defined bands are observable as it contains a mouse specific repeat and the rest of the sequence may not be conserved. The 1.5kb subfragment is highly repetitive across all species as shown by the Zooblot in Figure 15(b). There are no bands discernible at all in any of the DNA samples. The 1.2kb *BamH I* + *Hind III* subfragment from the 6kb *EcoR I* fragment produced a clearly discernible band in mouse DNA when hybridised to the Zoo blot see Figure 15(b). As expected this band is 6kb in size. There is a slightly smaller band present in the rat DNA sample just less than 6kb in size. There appear to be no bands in either marmoset or human DNA samples and the sequences are moderately conserved in rodents. When the temperature of the washes was increased to 68°C and the stringency increased to 1xSSC, the pattern of hybridisation for all the blots remained unchanged (data not shown). These results suggest that the 2.2kb and the 1.0kb *BamH I* + *Hind III* subfragments of Phage 7 6kb *EcoR I* fragment are composed of repetitive sequence that is present at high levels in the mouse genome. These repetitive sequences are not well conserved between different species. The pattern of hybridisation observed with the 1.5kb *BamH I* + *Hind III* subfragment on a Zoo blot suggests that it consists of repetitive sequence also. This repetitive sequence though is conserved across species. The 1.2kb *BamH I* + *Hind III* subfragment of the Phage 7 6kb *EcoR I* fragment contains a small amount of

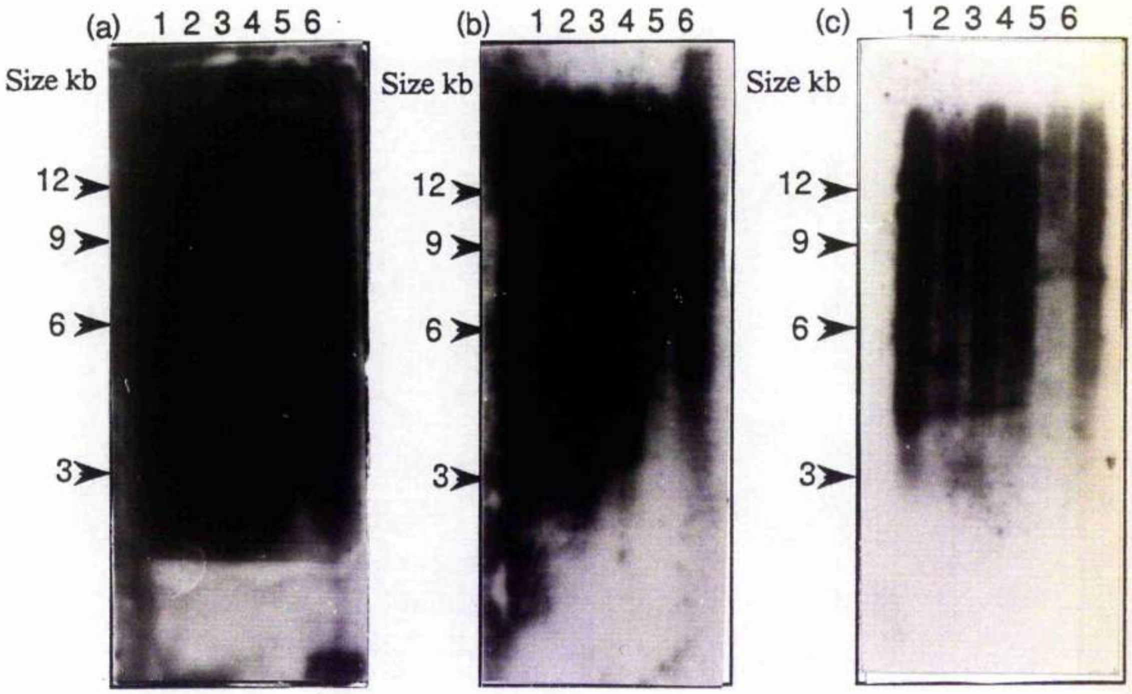


Figure 15 (a) : Evolutionary conservation of several *EcoR I* DNA fragments isolated from cloned genomic DNA.

4μg of genomic DNAs (listed below) were digested with *EcoR I*, electrophoresed, Southern blotted and hybridised with (a) the repetitive 3kb *EcoR I* fragment from Phage 7, (b) the 6kb *EcoR I* fragment from Phage 7 and (c) the 3.9kb *EcoR I* fragment from Phage 4. Wash stringency of the blots shown is 1xSSC, 0.1% SDS, 65ºC.

DNAs were derived from the tissues of; lane 1, mouse female; lane 2, mouse, male; lane 3, rat; lane 4, cow; lane 5, marmoset and lane 6, Human.

Note that all three fragments appear to be repetitive at this wash stringency.

Figure 15 (b): Evolutionary conservation of the *Apa I+Acc I* DNA subfragments of the 6kb *EcoR I* DNA fragment from Phage 7 .

4µg of genomic DNAs (listed below) were digested with *EcoR I*, electrophoresed, Southern blotted and hybridised with (i) the 1.0kb *Apa I+ Acc I* subfragment, (ii) the 1.2kb *Apa I+Acc I* subfragment, (iii) the 1.5kb *Apa I+Acc I* and (iv) the 2.2kb *Apa I+Acc I* subfragment. Wash stringency of the blots shown are 1xSSC, 0.1% SDS, 65°C. All of these probes were "quashed" with 100µg of sonicated unlabelled mouse DNA (for details of quashing see section 2.4.1).

DNAs were derived from the tissues of; lane 1, mouse female; lane 2, mouse, male; lane 3, rat; lane 4, hamster and lane 5, human.

Note that the 1.2kb fragment appears to to hybridise to a faint band at about 6kb in mouse DNA and a slightly smaller band in rat DNA.

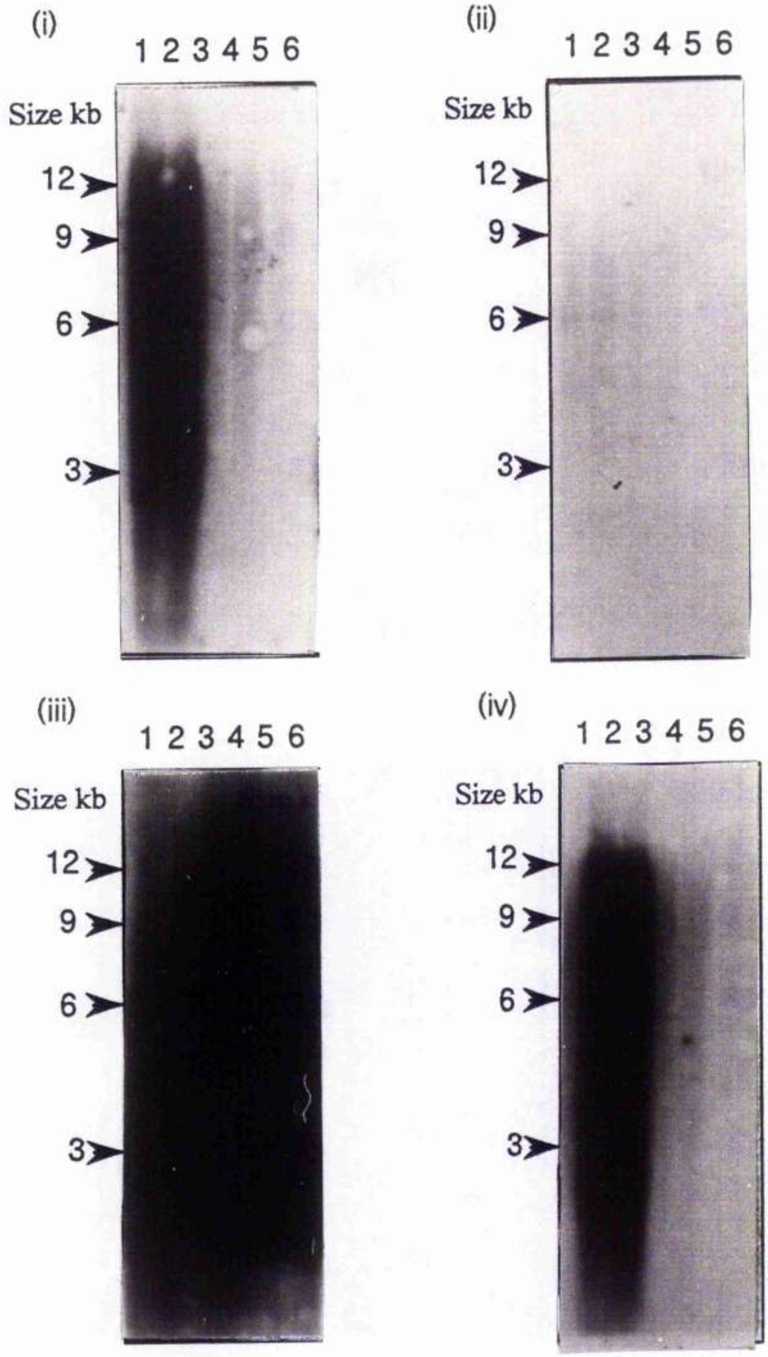


Figure 15(b): Evolutionary conservation of the *Apa I* + *Acc I* subfragments of the 6kb *EcoR I* DNA fragment from Phage 7.

repetitive sequence as shown by Zoo blot analysis. This sequence is conserved into rat and hamster at high stringency of washing.

All of the other *EcoR I* fragments identified from the isolated DNA were characterised by Zoo blot analysis. These data are summarised in Figure 16. The summary figure details the degree of conservation of DNA across species for the entire 32.8kb of isolated DNA.

The Phage 7, 3kb *EcoR I* fragment, 1.6kb *EcoR I* fragment and 1.4kb *EcoR I* fragment were found to be repetitive across all species at all stringencies of washing. The 2.7kb *EcoR I+Not I* fragment (that represents the 3kb *EcoR I* fragment disrupted by transgene insertion in the mutant mouse line) was found to be conserved across rat, hamster and marmoset at intermediate stringency (see Introduction). The 1.9kb *EcoR I* fragment, contained unique sequences in the mouse that were not conserved across species. These fragments are shared with the Phage 1 isolate. In Phage 7 the 1.3kb *EcoR I+Not I* fragment (not found on Phage 1) appeared repetitive at low stringencies of washing. With increasing temperature and stringency the repetitive nature of the signal decreased but is still present and a band was clearly visible in the mouse DNA lane. This band is present as two copies in the female genome and only one in the male genome, due to the fragment being present on the X chromosome. This can be clearly seen in Figure 17. At intermediate washing stringency of (55°C and 0.1xSSC) a band was clearly observable in rat, cow, primate and yeast. Several bands are observable in the human DNA samples, but it exhibits a repetitive nature, as do the primate DNA samples (Figure 17). From this blot the sequence of the 1.3kb *EcoR I + Not I,* from Phage 7, contains sequence that is repetitive in human but not repetitive in mouse. It is possible that a repetitive DNA element is present in this DNA fragment that has a low copy number in the mouse but that copy number is greatly expanded in the human genome or is functionally homologous to a human repeat.

When the temperature of the wash was increased to 68°C and the stringency maintained at 0.1xSSC the pattern of hybridisation changed only slightly. The mouse

Figure 16: Summary of Zoo blot analysis.

The legends below summarize the information obtained from the Zoo blot analysis.for the *EcoR I* fragments obtained from the cloned genomic DNA. The information shown summarizes the results of Zoo blot analysis without quashing.
Restriction sites shown are *EcoR I*, E1 and *Not I*, N1.

▧ Slightly repetitive with unique sequence at high stringency.

▦ Unique at medium stringency

▦ Unique at high stringency

▨ Slightly repetitive medium stringency

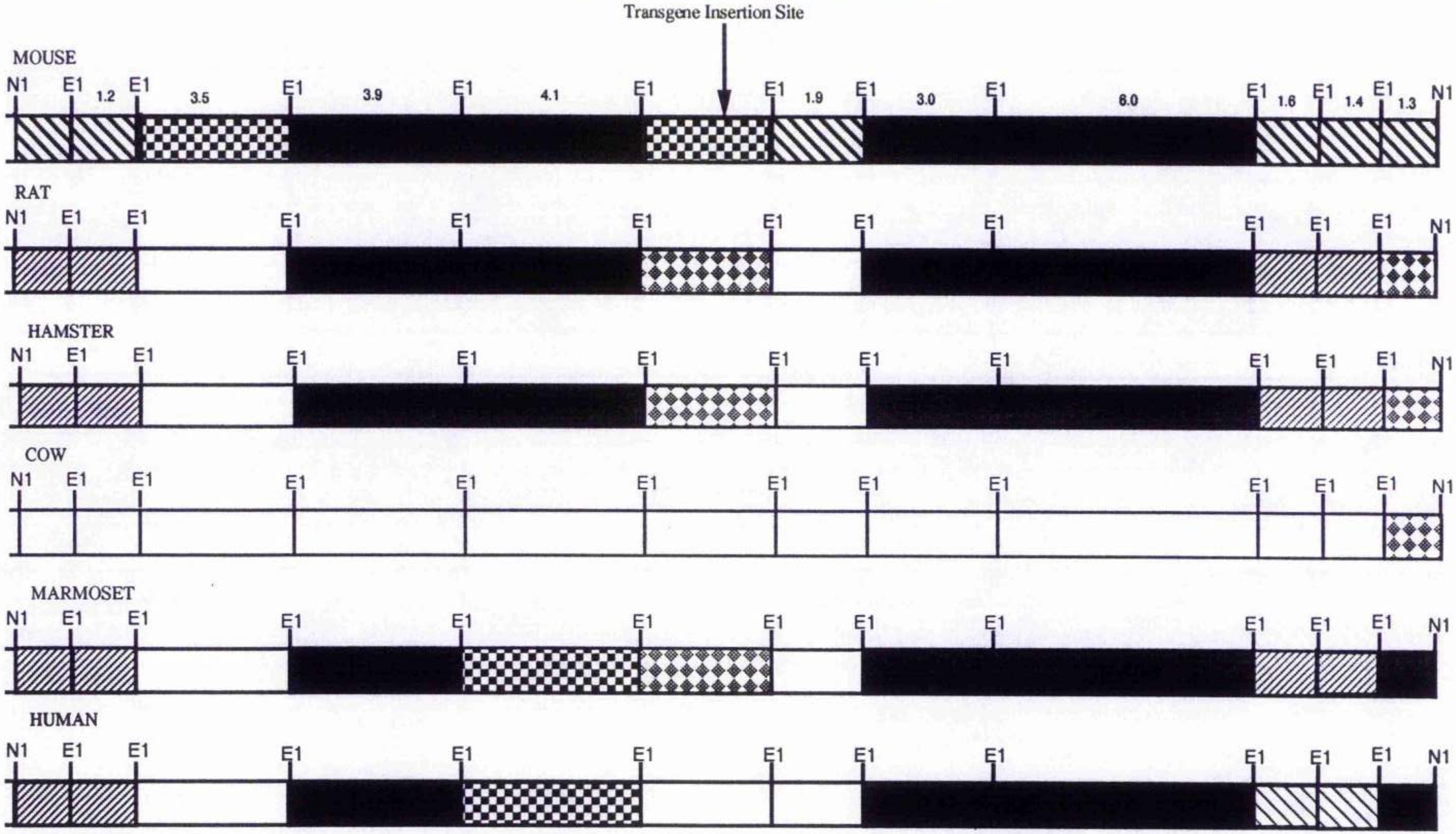
▧ Slightly repetitive at high stringency

■ Repetitve at high stringency

■ Repetitve at medium stringency

□ No signal at any stringency

EVOLUTIONARY CONSERVATION OF CLONED DNA


Transgene Insertion Site
MOUSE
N1 E1 1.2 E1 3.5 E1 3.9 E1 4.1 E1 E1 1.9 E1 3.0 E1 6.0 E1 1.6 E1 1.4 E1 1.3 N1
RAT
N1 E1 E1 E1 E1 E1 E1 E1 E1 E1 E1 E1 N1
HAMSTER
N1 E1 E1 E1 E1 E1 E1 E1 E1 E1 E1 E1 N1
COW
N1 E1 E1 E1 E1 E1 E1 E1 E1 E1 E1 E1 N1
MARMOSET
N1 E1 E1 E1 E1 E1 E1 E1 E1 E1 E1 E1 N1
HUMAN
N1 E1 E1 E1 E1 E1 E1 E1 E1 E1 E1 E1 N1

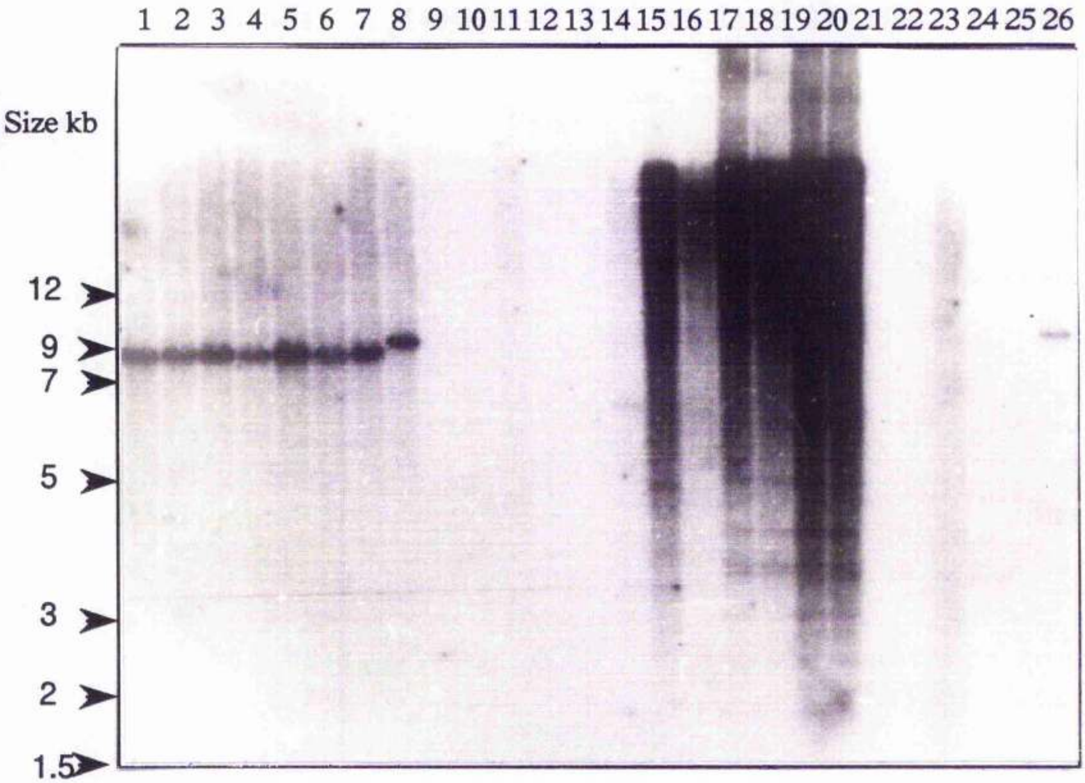


Figure 17: Evolutionary conservation of the 1.3kb *EcoR I + Not I* DNA fragment from Phage 7.

4μg of genomic DNAs (listed below) were digested with *EcoR V*, electrophoresed, Southern blotted and hybridised with the 1.3kb *EcoR I + Not I* fragment from Phage 7. Wash stringency of the blot shown was 0.1xSSC, 0.1% SDS, 55°C.

DNAs were derived from the tissues of; lane 1, *mus domesticus*, female; lane 2, *mus domesticus*, male; lane 3, *mus spretus*, female; lane 4, *mus spretus*, male; lane 5, *mus macedonicus*, female; lane 6, *mus macedonicus*, male; lane 7, *mus hortulanus*, female; lane 8, *mus cookii*, male, lane 9, rat; lane 10 & 11 hamster; lane 12, dog; lane 13, rabbit; lane 14, cow; lane 15 & 16, primate; lane 17, 18, 19 & 20, human; lane 21, chicken; lane 22, axolotl; lane 23, salmon; lane 24, drosophila; lane 25, *C. elegans*; lane 26, *S. cerevisiae* .

signal was unchanged and the signals observed with other species DNA decreased in intensity. The pattern of hybridisation observed in the primate and human DNA sample was unchanged. It is not possible to tell whether any unique sequences are conserved from the mouse DNA fragment into human DNA due to the highly repetitive nature of the signal observed.

All of the other *EcoR I* fragments from Phage 4 were characterised by Zoo blot analysis. Their pattern of hybridisation to, and conservation across DNA from different species are depicted in Figure 16. The highly repetitive 3.9kb *EcoR I* fragment from Phage 4 exhibits strong conservation of its repetitive sequence between DNAs from different species, demonstrated by high stringency hybridisations. While the 1.25kb *EcoR I+Not I* fragment demonstrated that it contained highly repetitive sequences also, this repetitive sequence was not conserved between species. The 1.2kb *EcoR I* fragment and the 1.4kb *EcoR I + Not I* fragment from Phage 4 harbour some repetitive sequence, observable at low stringency, suggesting that the sequence is not repetitve but has homology to a repeat. This repetitive sequence hybridised to the DNA of other species under these conditions. At high stringency, the patterns of hybridisation observed with these fragments revealed that they contained sequences that were homologous to repetitive sequences. These sequences were not conserved across species at any of the high stringency washing conditions used. The 3.5kb *EcoR I* fragment from Phage 4 was unique in mouse and did not display any conservation of sequence across species even at low stringencies of washing used.

2.4.1 Analysis of unique sequence conservation by "quashing" repetitive elements

A considerable problem with the analysis of this locus has been that it is composed extensively of repeated sequence. There was the possibility that short stretches of unique sequence were hidden within these repetitive sequences. In order to address the conservation of small stretches of unique sequence within repetitive

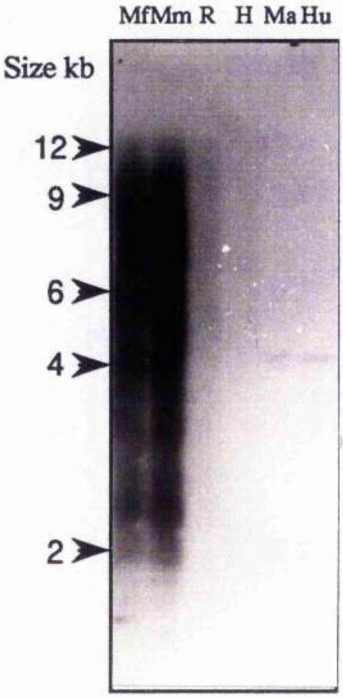


Figure 18: Evolutionary conservation of the 4.1kb *EcoR I* DNA fragment from Phage 4.

4μg of genomic DNAs (listed below) were digested with *EcoR I*, electrophoresed, Southern blotted and hybridised with a 'quashed' 4.1kb *EcoR I* fragment from Phage 4. Wash stringency of the blot shown was 0.1xSSC, 0.1% SDS, 60°C.

DNAs were derived from the tissues of; Mm, *mus domesticus*, male; Mf, *mus domesticus*, female; R, rat; H, hamster; Ma, marmoset and Hu, human.

A band at about 4.1kb is visible in the marmoset and human samples. Comparison of the pattern observed in this blot, with a quashed 4.1kb *EcoR I* probe, to those of subfragments of the 4.1kb fragment in Figure12 (washed at a similar stringency) reveals the absence of repetitive signal from all the DNA samples except the mouse.

stretches, highly repetitive stretches were quashed. Labelled probe fragments were prehybridised with a 2,000-fold excess of cold unlabelled sonicated genomic DNA thereby competing out (quashing) highly repeated sequences from the probe and preventing their hybridisation to a blot. Any unique sequence present in the probe fragment has an extremely low probability of binding to the sonicated DNA and remains competent (single stranded) to hybridise to the blot.

The Zoo blot analysis was repeated using the 'quashing' technique, for those fragments identified as being highly repetitive and highly expressed. For the 6kb and repetitive 3kb *EcoR I* fragments from Phage 1 and Phage 7 'quashing' of the repetitive sequences removed all signal from the Zooblots upon washing, except for the 1.2kb *BamH I+Hind III* subfragment from the 6kb *EcoR I* fragment see Figure 14(b). This result indicates that most of these fragments are composed solely of repetitive sequences, or that if unique sequences were present (as indicated by the results of the 1.2kb *BamH I + Hind III* from the 6kb fragment exhibiting unique sequence in Figure 15(b)) they were so small and interspersed with repeats as to evade detection by this technique (data not shown).

Using a quashed Phage 4, 4.1kb *EcoR I* probe fragment unique sequences were detected in marmoset and human DNA. Although the pattern of hybridisation is still very repetitive in mouse the repetitive signal that was observed in the rat and hamster lanes is greatly reduced, though no single band is observed in these lanes. No repetitive signal is observed at all in the marmoset and human lanes. These two DNA samples both exhibit a single band sized at about 4.1kb. The results of this 'quashed' Zoo blot indicate that the 4.1kb *EcoR I* fragment of Phage 4 contains unique sequence shrouded in repetitive sequences. It also reveals that this unique sequence is conserved into marmosets and humans (Figure 18).

2.5 Sequence analysis of selected isolated DNA fragments

The analyses of the isolated locus have shown that the DNA at this locus is highly complex. It was clear that in order to identify any DNA sequences contained in a transcript an alternative approach of investigation was required. From the information obtained about patterns of expressed sequence and the highly repetitive nature of these sequences, it was felt that dideoxy sequencing of some of the isolated fragments might yield more information regarding a potential gene. The fragments selected for sequencing were some of those that were shown to be expressed at high levels and/or shown to exhibit conservation of sequence across species. Some of the fragments were only partially sequenced and not all of the sequence data is presented here.

Figure 19 shows the sequence data obtained for the Phage 7, 3kb *EcoR I*, 0.7kb *BamH I + Hind III* subfragment, which was highly repetitive and expressed. A search for Open Reading Frames (ORF) using the MacVector^{TM} program identified several ORFs. These ORFs were either incomplete, as they resided at the ends of the sequence or composed of only about 100 nucleotides. When the sequence was scanned for putative splice acceptor and donor sequences, using MacVector^{TM}, none was found to be located at the ends of the small ORFs. A database search was carried out using the 0.7kb subfragment sequence against those sequences present in the BLASTN database. Due to the repetitive nature of the sequence many matches were found in the search e.g., 100 nucleotide long stretches of the complete 700 nucleotides of sequence were found to be 80% identical to sequences spread across the mouse beta globin locus (Shehee *et al.* 1989). The mouse beta globin locus is over 55,000 base pairs long. Another match was to the *mus musculus* long interspersed sequence LINE-1 (L1Md-6) (Burton *et al.*, 1991), this match exhibited 78% across 400 nucleotides. A large number of other matches were obtained from the database using the 0.7kb subfragment sequence, they all shared the repeats of the two matches detailed here. The database information confirmed that the sequences present in the 0.7kb subfragment are repetitive and are wide spread in the mouse genome.

Sequence information was obtained for the Phage 7, 1.3kb *EcoR I + Not I* fragment which had been found to be expressed and contain repetitive sequences.

Figure 19: Sequence of the 0.7kb *BamH I* + *Hind III* subfragment from the repetitive 3kb *EcoR 1* fragment from Phage 7.

```
          10         20         30         40         50
GCTCCACCGC  GGTGCGGCCG  CTCTAGAACT  AGTGGATCCA  TCCCATATCA
CGAGGTGGCG  CCACGCCGGC  GAGCTCTTGA  TCACCTAGGT  AGGGTATAGT
          60         70         80         90        100
GCCTCCAAAC  GCCTGACACC  ATTGCATACC  CCAGCAAGAT  TTTGCTGAAA
CGGAGGTTTG  CGGACTGTGG  TAACGTATGG  GGTCGTTCTA  AAACGACTTT
         110        120        130        140        150
GGACCCTGAT  ATAGCTGTCT  CTTGTGAGGC  TATGCTGGTG  CCTGGCAAAC
CCTGGGACTA  TATCGACAGA  GAACACTCCG  ATACGACCAC  GGACCGTTTG
         160        170        180        190        200
ACAGAAGTGG  ATGCTCACAG  TCATCTATAG  ATGGAACACA  GGGCCCCCAG
TGTCTTCACC  TACGAGTGTC  AGTAGATATC  TACCTTGTGT  CCCGGGGGTC
         210        220        230        240        250
TGGAGGAGCT  AGAGAAAGTA  CCAAAGGAGC  TAAAGGGATC  TGCAACCCTA
ACCTCCTCGA  TCTCTTTCAT  GGTTTCCTCG  ATTTCCCTAG  ACGTTGGGAT
         260        270        280        290        300
TAGATGGAAC  AACAATATGA  ACTAATCAGT  ACCCCCAGGA  GCTCGTGTCT
ATCTACCTTG  TTGTTATACT  TGATTAGTCA  TGGGGGTCCT  CGAGCACAGA
         310        320        330        340        350
TAGCTGCATA  TGTATCAGAA  GA*TGGCCTA  GTCTGCCATC  ACTGGGAAGA
ATCGACGTAT  ACATAGTCTT  CT*ACCGGAT  CAGACGGTAG  TGACCCTTCT
         360        370        380        390        400
GAGGCCCTTT  GGTCTTGCAA  ACTTTATATG  CCTCAGTACA  GGGGGAATGC
CTCCGGGAAA  CCAGAACGTT  TGAAATATAC  GGAGTCATGT  CCCCCTTACG
         410        420        430        440        450
AGGGCCAAGA  AGGTAGGTGG  GTATGGGAGT  GGGGGGGGGG  TTAGGGTATG
TCCCGGTTCT  TCCATCCACC  CATACCCTCA  CCCCCCCCCC  AATCCCATAC
         460        470        480        490        500
GGGGACTTTT  GGAATATCAT  TTGAAATATA  AATGAGGAAA  ACAGGTTGTA
CCCCTGAAAA  CCTTATAGTA  AACTTTATAT  TTACTCCTTT  TGTCCAACTA
         510        520        530        540        550
GTGGCAATCA  CACACACAGA  CACACACAGA  CACACAGACA  CACAGACACA
CACCGTTAGT  GTGTGTGTCT  GTGTGTGTCT  GTGTGTCTGT  GTGTCTGTGT
         560        570        580        590        600
CAGACACACA  GACACACACG  TTAAGCAGTT  AGTATTCACT  GTAGATCAAT
GTCTGTGTGT  CTGTGTGTGC  AATTCGTCAA  TCATAAGTGA  CATCTAGTTA
         610        620        630        640        650
ATTGTCATAT  AGGAACCAGG  CATTCTTTTC  CAAGATTGAA  GATGATGTTT
TAACAGTATA  TCCTTGGTCC  GTAAGAAAAG  GTTCTAACTT  CTACTACAAA
         610        620        630        640        650
TCCCTATAAA  ACCCTAGAGT  AAATTCTAAC  TGCCTTACAA  AAATCAGCCA
AGGGATATTT  TGGGATCTCA  TTTAAGATTG  ACGGAATGTT  TTTAGTCGGT
         660        670        680        690        700
TTAAGTAGTT  CACATTCAGA  TTTTAAAAAC  AGGTTGTA
AATTCATCAA  GTGTAAGTCT  AAAATTTTTG  TCCAACAT
```

The complete sequence of this fragment could not be obtained due to an internal region through which sequencing could not pass, possibly due to a conformation block. However, over 1kb of sequence information was obtained from the ends of this fragment. Several small putative ORFs were identified using MacVector^{TM} program but no splice sites were identified at either end of the ORFs. The internal block to sequencing was a $(GA)_{20}$ repeat, this was possibly larger but the repeat sequence prevented the polymerase from progressing properly across it. The database search with this sequence revealed that this $(GA)_n$ sequence motif is common in sequences present in the database. It is found in the sequences of the *Rattus norvegius* renin gene, exons 1-9 (Fukamizu *et al.,* 1988), the mouse beta globin complex (Shehee *et al.*, 1989), the 5' flanking sequence of the *homo sapiens* activin beta-A subunit (Tanimoto *et al.*, 1992) and the mouse microsatellite marker DNA for IL-2 gene encoding Interleukin 2 (Hearn *et al.*, unpublished 1992) to mention but a few.

When this $(GA)_n$ repeat is removed from the database analysis only two matches were found. The first of these was to sequence from a *Homo sapiens* DNA repeat region (Weber, unpublished 1994), this match consisted of 62% identity over only 86 nucleic acid residues. The second match was with sequence from the Drosophila retrotransposon 412 genome. Again it exhibited a moderate level of identity, 63%, over a small region of 72 nucleic acids. This sequence analysis is very interesting in the light of the Zoo blot analysis of the Phage 7, 1.3kb *EcoR I* + *Not I* fragment. It was found to consist mainly of unique sequence in mouse but to be highly repetitive in primates and humans. The identification of these repetitive elements coupled with the Zoo blot data supports the idea that these elements are present in the mouse genome at a very low level, but at high levels in the genomes of primates and humans.

The partially conserved 4.1kb *EcoR I* fragment from Phage 4 was also sequenced. It had previously being subcloned and undergone further restriction digests. Digestion with *BamH I* + *Hind III* yielded four fragments, accounting for the complete 4.1kb fragment which were subcloned into pBluescript. These four subfragments were partially sequenced. The sequences obtained for the Phage 4, 4.1kb *EcoR I*, 1.3kb *BamH I + Hind III* subfragment were found to contain several putative ORFs. Sequence analysis failed to reveal splice sites present at the boundaries of these putative ORFs. The search of the BLASTN database revealed that the sequence was similar to known repetitive sequences. The most obvious similarity was with *Homo sapiens* (haplotype 1, allele MS32) minisatellite (Jeffrey, unpublished, 1990), with which it shared 80% identity at the nucleotide level. Another interesting match from the database search was to a sequence from *Dictyostelium discoideum* that encoded a nucleoside diphosphate kinase (*ndkm*) gene (Troll *et al.*, 1993). Over a region of 114 nucleic acids it exhibited 56% identity to the sequence from the 1.3kb *BamH I + Hind III* subfragment. The sequence of the 1.3kb *Bam H I + Hind III* subfragment was shown to contain short stretches of sequence similar to that of repetitive elements from other organisms. No matches were found at high identity or to span the entire sequence of the 1.3kb *BamH I + Hind III* subfragment.

The partial sequence obtained from the ends of the 1.4kb *BamH I + Hind III* subfragment of Phage 4 was found to contain putative ORFs using the MacVector^{TM} program but no splice sites were shown to flank these putative ORFs. Database searches using this sequence indicated that the sequence was composed of B2 like repetitive sequences. This B2 family of repeats is widely dispersed in the genome of many organisms. The partial sequence of the 1.4kb *BamH I + Hind III* subfragment exhibited 82% identity to the consensus sequence of the *Rattus norvegicus* B1- and B2 repeats across 150bp of sequence (Den Dunnen and Schoenmakers, 1987). This B2 repeat is present in the mouse genome also. The database search revealed homology to the sequence of a small poly adenylated B2 RNA gene present in the mouse genome (Tillib *et al.*, 1985). The region of homology between the partial sequence of 1.4kb subfragment and the sequence of the B2 RNA was 150bp long and exhibited 77% identity. This repetitive sequence is often closely associated with gene sequences. The 150bp of sequence found in the 1.4kb subfragment that was shown to be similar to sequence from the B2 repeat sequence was also found to be similar to a sequence from another mouse gene. The similar sequence was found in sequence from exon 4 and 5 of the gene encoding myotonic dystrophy protein kinase (DM-PK) from *Mus musculus* (Mahadevan *et al.*, 1993).

The sequence data gained for some of the fragments confirmed and explained the results obtained with the blotting analyses of expression and conservation of the sequences at this locus. The presence of so many repetitive elements and their distribution between different species supported the Zoo blot data. The presence of the defined repetitive sequences in the transcripts of many different genes explains why such high levels of expression are observed as well as the lack of discrete bands when the isolated fragment is used as a probe on a Northern blot of total RNAs.

The sequence data, while aiding our understanding of the genomic analysis did not reveal a putative ORF with splice sites flanking it. So it was still not possible to identify any sequences that may be part of a relevant transcript. Even though much of the sequence characterised showed some relation to expressed sequences we were unable to determine whether we were observing the expression from the locus, or the hybridisation of the isolated DNA to other expressed sequences. The problem remained as to whether a gene was present within the 30Kb of isolated DNA or not. To unravel the highly complex nature of this locus required an alternative method of analysis.

2.5.1. Critique I

The data presented in Section 2 indicated that there were extensive repetitive sequences present in the DNA cloned from the locus (exemplified in Figure 4 (c) and summarised in Figure 7). Several probes when hybridised to genomic DNA Southern blots resulted in a heavy smear of hybridisation (exemplified by Figure 4(c)). Which was interpreted as the probe fragment containing repetitive sequence. Alternative interpretations could be that the probe DNA had degraded or that there was degraded genomic DNA on the blots (both of which would result in a smear of hybridisation). The latter was clearly not the case. Restriction digested genomic DNA samples were visualised by Ethidium Bromide staining and UV illumination before blotting and showed no evidence of degradation as satellite bands were present and the DNA was digested to a large range of fragment sizes. With regard to the probe fragments, although these were not verified as being intact (by gel analysis) subsequent to this use as probes, several pieces of data suggest that these were intact; (i), certain probes hybridised selectively to DNAs of different species suggesting the presence of repeat sequences in one species not present in another. A degraded probe would not demonstrate this selectivity (a more detailed explanation of this is given below); (ii), the probes hybridised to the predicted bands on plasmid blots (data not shown), and again degraded probes might hybridise to more fragments; (iii), in certain instances the repetitive signal could be quashed by the addition of excess cold genomic DNA to reveal specific bands (Figure 18), which would not happen with a degraded probe. The probe fragments were routinely isolated by electrophoresing restriction digested plasmid DNA on low melting point agarose gels and were observed to be intact. Probes that had been isolated from low melting point agarose on different occasions produced the same results. Taken together these observations demonstrate that the probes were intact and that the interpretation of repetitive sequence content must be correct.

There is still a possibility that the interpretation of these results is flawed. This difficulty would have been overcome by the inclusion of positive controls on all of the blots probed and presented. If a probe was found to hybridise to the control in the expected manner then this would have answered some of the questions posed by the presence of such extensive repeats. It would also have helped ascertain whether there were any problems with the hybridisation conditions and the quality of the blots. The presence of positive controls on all Southern, Zoo and Northern blots would have aided our interpretation of results greatly.

The results of the Southern analysis are corroborated by the results of the Zoo blot analysis of the cloned DNA which revealed species specific repeats. This can be clearly seen in Figure 13 (b) i, which shows the presence of repeat sequence in the mouse genomic DNA samples but none in the human or hamster genomic DNA samples. This situation is reversed in Figure 17, which represents a Zoo blot probed with a 1.3kb *EcoR I + Not I* fragment. In this instance the probe fragment hybridises to a 9kb *EcoR V* fragment in mouse genomic DNA samples indicating that the fragment is unique in mouse DNA. Interestingly, however, the murine fragment contains sequences which hybridise to extensive repeat sequences in primate and human DNA samples. These results indicate that there are repeat containing sequences present in the cloned DNA and that these repeat sequences are different from each other and are species specific.

The findings from the Southern analysis distinguishing repetitive and unique fragments are completely supported by those of the Reverse Southern analysis and sequencing analysis of cloned DNA. Those fragments that had been shown to contain repeats by conventional Southern analysis (as in Figure 4) also hybridised in Reverse Southern analysis (Figure 5) suggesting high degree of representation in the genome. Moreover, in an attempt to identify small regions of coding sequence several candidate fragments of cloned mouse genomic DNA were sequenced. Database analysis revealed multiple repeat elements as was expected from the previous results, (described in detail in section 2.5). The results of the

Southern, Reverse Southern and sequence analysis when viewed together support the conclusion that the locus studied in this thesis is replete with repeat containing sequence.

These conclusions also hold for the Northern analyses, section 2.3. Again the probe fragments must have been intact and the smear of hybridisation signal can only be due to RNA containing repeats. The RNAs used for these Northern blots were total RNAs and these were electrophoresed on denaturing gels, stained with ethidium bromide and illuminated with UV light. In so doing it was possible to identify the 28S ribosomal band which indicated that the RNA was intact as any degradation would have resulted in smearing or absence of the 28S ribosomal band.

With many of the fragments used as probes in the Zoo blot analysis repetitive signals were observed as discussed above. It is usually more informative to use probe fragments in the size range of 0.5kb to 1kb in Zoo blot analysis to reveal conserved motifs. However, most of the probes used in the analysis presented in this thesis were slightly larger than this, in the range 1kb to 2kb. However, in the absence of useful restriction sites to obtain smaller probe fragments, an attempt was made to prevent the repetitive sequences of the larger probe fragment from hybridising by using the technique of quashing (Figure 18).

At this point this approach (Zoo blotting to detect conserved motifs) was not pursued further as the exon-trapping approach was considered as providing a more effective means of isolating the exon sequences. In retrospect, in consideration of the small size of the exons and the extensive repeat sequences within the region it was a fortuitous decision as other methods available may well have failed to yield gene sequences.

2.6 EXON-TRAPPING OF ISOLATED DNA

2.6.1 Prologue

Exon amplification or 'trapping' is a method for detecting genes, independent of their transcriptional status. In exon-trapping experiments a genomic fragment is cloned into the intron of a reporter vector, which provides splice donor and acceptor sites flanking the inserted sequence. After processing of the primary transcript in tissue culture, any exon contained within the insert will be included (trapped) between the vector exons. Genomic fragments not containing exons will be simply spliced out during processing of the vector transcript. RNA-based PCR amplification using oligodeoxynucleotide primers specific for the reporter gene sequences is then used to detect the presence of any genomic derived exons. The method can only trap exons that possess both splice donor and acceptor sites. As the first and last exons of a gene do not contain both splice sites these will not be trapped and therefore genes must contain at least two introns to be isolated by exon-trapping.

Exon-trapping is ideally suited to screening large, complex sources of DNA for functional exons (Church *et al.*, 1994). Church *et al.*, have reported that when pools of 9-12 cosmids from human chromosome 9 were screened, an average of 1.4 exons was obtained per 20-25Kb of genomic DNA screened. This Figure increases when cosmids were screened individually. Exon-trapping was used recently to identify a copper transporter gene defective in individuals with Menkes disease (Vulpe *et al.*, 1993) ,the neurofibromatosis type 2 tumour suppressor gene (Trofatter *et al.*, 1993) and the candidate gene for Huntington's disease (The Huntington's Disease Collaborative Research Group, 1993). Having been unable to identify any gene encoding regions using a strategy based on the conservation of DNA sequence across species or by expression analyses, exon trapping was employed for the *DXRib 1* locus.

2.6.2 Preparation of phage DNA for exon-trapping

The cloned 32.8kb of DNA from the *DXRib 1* locus was used for exon-trapping as detailed in the Materials and Methods. The exon-trapping vector, pSPL3, was digested with *EcoR I* and treated with Calf Intestinal Phosphatase (CIP). The three isolated recombinant phage containing the 32.8kb of mouse genomic DNA were also digested with *EcoR I*. No *EcoR I* sites are present in the phage arms ensuring that only mouse genomic DNA fragments were subcloned. The genomic fragments from the three phage were then shotgun cloned into the vector, pSPL3, (Materials and Methods) and the three ligation mixes transformed into XL1-Blue seperately. The non-recombinant frequency of each transformation was estimated by plating 10% of each transformation on medium containing 100μg/ml of carbenicillin. The remaining 90% of each transformation was used to inoculate 50ml of selective liquid medium. In the case of each transformation, the non-recombinant frequency was found to be less than 10% and all colonies analysed by mini-prep analysis and subsequent restriction digest analysis were found to contain inserts. Plasmid DNAs from the three pooled clones were used for COS-7 cell transfections. By this method, only exons spanning an *EcoR I* site would be missed.

2.6.3 Expression of subcloned DNA in COS-7 cells

10^5 x COS-7 cells were transfected with 1μg of the vector pSPL3 containing the mouse genomic DNA fragments. Three separate transfections were performed, one for each pool of clones derived from the three individual recombinant phage isolates. The cells were incubated at 37°C for 24hrs after which total RNA was prepared yielding 40-50μg of RNA (Materials and Methods). 3μg of this total RNA was then used for cDNA synthesis.The cDNA synthesis reaction was primed using an oligomer (SA2) from the 3' end of the reporter construct, ensuring that cDNA synthesis only occurred from vector derived RNA transcripts. Primary PCR was performed on this cDNA using primers homologous to the reporter exons. This PCR produced a double stranded product, composed of the control exons flanking any exons that had been trapped by splicing events. To eliminate any products resulting from cryptic splice sites or single splice donors/acceptors, or vector alone splicing events the primary PCR product was treated with *Bst XI*, as described in Materials and Methods. This restricted all products that did not contain a functionally spliced exon product and exons that contained a *Bst XI* site. A secondary PCR reaction was performed on the *Bst Xi* treated primary PCR product. The primers used for this reaction are nested in relation to the initial primers. This helped ensure the specificity of the amplified exons. The primers used in the second round of PCR also contained dUMP residues that allow efficient uracil DNA glycosylase (UDG) cloning into the pAMP10 cloning vector (Gibco BRL)(Materials and Methods). The secondary PCR reaction was also carried out in parallel on samples that had not undergone *Bst XI* treatment as a control.

After the secondary PCR reaction, 5μl of each 50μl reaction was electrophoresed in a 2% agarose gel (Figure 20). As predicted a major product of 177bp is present in the vector control lane. This vector derived product is also present in the *Bst XI* untreated lanes of mouse genomic derived exons. This vector product is removed by treatment with *Bst XI*. Several other bands are also present in those samples that did not undergo *Bst XI* treatment, these may be vector derived or the result of cryptic splicing events in the samples. There appears to be a band at 220bp and another at 480bp common to all untreated samples, removed by treatment with *Bst XI*. Phage 1 derived exon-trapped products consist mainly of a single band 270bps in size, with several fainter products running at 600-700bp. Exon-trapped products from Phage 4 appear more numerous and discrete. Again there is a 270bp product, as observed with Phage 1, but bands are also present at 300bp, 340bp, 390bp and 450bp. Fainter bands are also present at 600-700bp. The products observable in the Phage 7 sample again consist of the 270bp shared with the other phage, in addition to a band at 350bp. This PCR product is apparently slightly larger than the 340bp band seen with Phage 4. As with the exon-trapped products derived from the

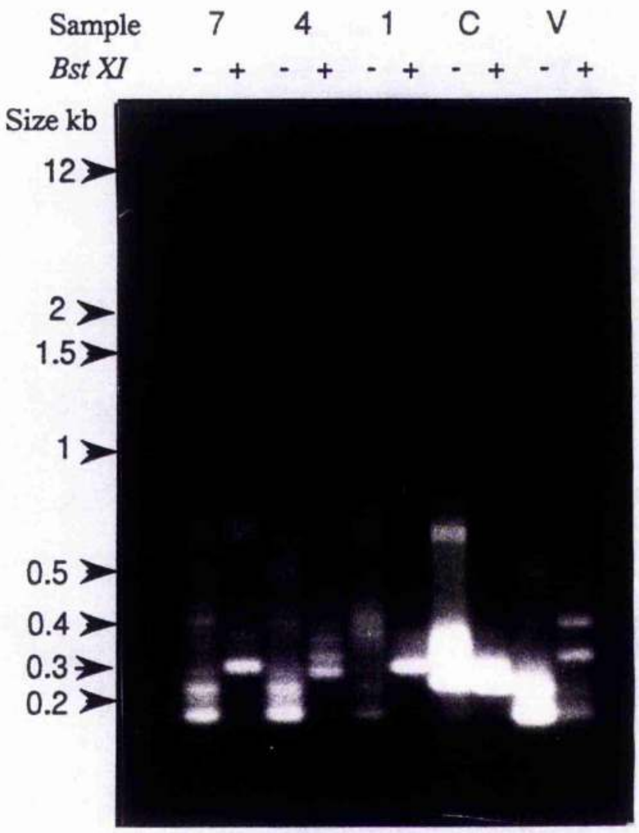


Figure 20: Agarose gel analysis of secondary PCR products from exon-trapping.

Exon-trapping products from Phage 1, 4 , 7, the control exon (C) and the vector only control (V) were either treated with (+) or without (-) *Bst XI*. A fifth of the secondary PCR reaction produced by exon-trapping from each sample was electrophoresed on a 2.0% agarose gel and visualised with UV light by ethidium bromide staining.

The 100bp and 200bp bands present in all untreated samples that correspond to vector derived products are removed by *Bst XI* treatment. A band smaller than 300bp product is present in the *Bst XI* treated samples from the cloned genomic DNA. Several other products are present in the cloned DNA samples, Phage 7 exhibits several products at 700bp, Phage 4 exhibits discrete products ranging in size from 300bp to 450bp. There are clear differences between *Bst XI* treated and untreated samples as well as between controls and genomic samples.

other phage, bands are observable at 600-700bp. In the case of Phage 7, these products are discrete and three can readily be identified. All these PCR products are resistant to *Bst XI* treatment.

When comparing the exon-trapped products from the three phage with products derived from the vector control or control exon, a clear difference was observable (Figure 20). This result suggested that the exon-trapping had been successful. The total *Bst XI* treated, secondary PCR products were then used to confirm their derivation from the 3 phage isolates and further characterised. This was achieved by probing Southern blots of recombinant phage DNA digests, mouse genomic DNA digests and Northern blots of mouse embryonic total RNA with the secondary PCR products as the probe.

2.6.4 Analysis of the products of exon-trapping to confirm authenticity

To confirm if any of the trapped product is derived from mouse genomic DNA, total exon-trapped PCR product from the individual phage was radio-labelled by random priming (Materials and Methods) and then hybridised to Southern blots of Phage DNA. Each phage blot consisted of DNA from the original three phage isolates digested with a restriction enzyme, in this case *EcoR I*. This probing of phage DNA was repeated for each of the three samples of exon-trapped products. Figure 21 demonstrates that several sequences from the original phage, have been trapped. Hybridisation with PCR product from Phage 1 and Phage 7 is consistent with the high level of shared sequence between these phage. It is also interesting to note that the fragments of DNA that exhibited the highest level of expression by Northern analysis and Reverse Northern analysis, i.e., the 6kb and 3kb *EcoR I* fragments (see Figure 11), appear to contain functional exons. The 1.9kb *EcoR I* fragment that is shared between all three phage also appears to contain a functional exon. The consistency between the three individual experiments is suggestive of valid product, not cryptic or artifact products.

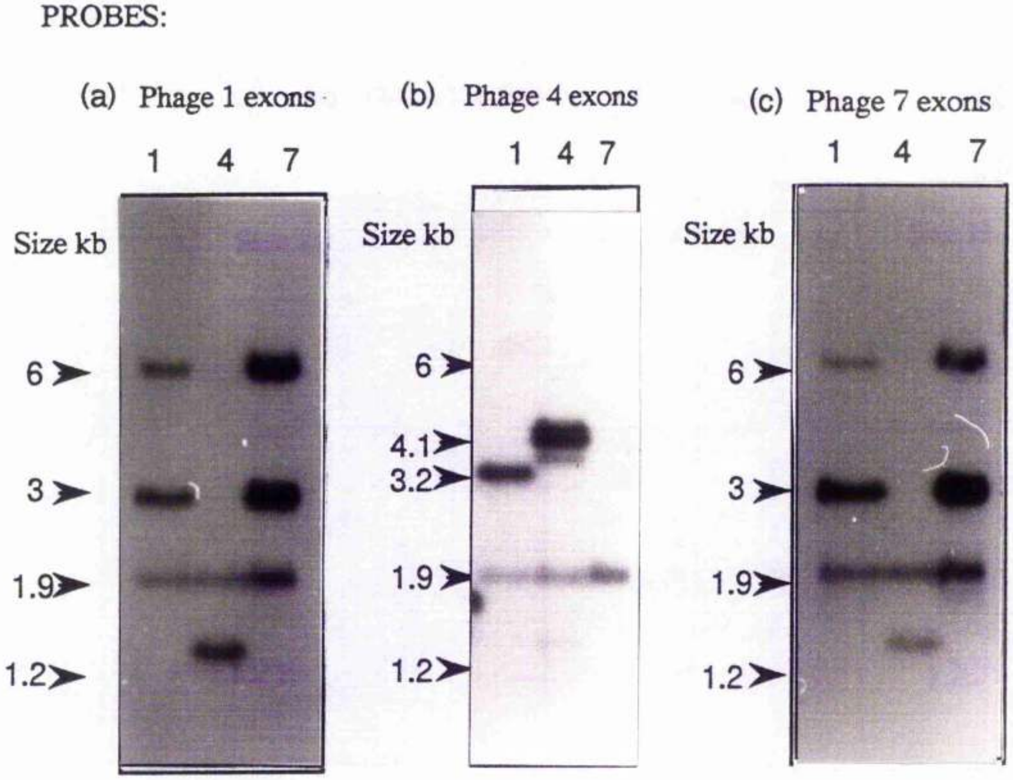


Figure 21: Location of trapped exons in cloned genomic DNA.

Southern blots of cloned genomic DNA probed with total secondary PCR products from exon-trapping. Each Southern blot represents 1μg of each genomic clone digested with *EcoR I* and *Not I*; lane 1, Phage 1; lane 2, Phage 4; lane 3, Phage 7, electrophoresed through 0.8% agarose gel, Southern blotted and probed with total secondary PCR products derived from Phage 1, Phage 4 and Phage 7 as indicated above.

Note: that the 4.1kb fragment is unique to Phage 4 and only detected by Phage 4 PCR products. The 3.2kb *EcoR I + Not I* of Phage 1 which overlaps the 4.1kb fragment of Phage 4, is detected by Phage 4 PCR products but not by either Phage 1 or 7 PCR products. Phage 1 and Phage 7 PCR products detect similar fragments as the sequence of both overlap. They also detect the 1.25kb *EcoR I +Not I* fragment from Phage 4.

The pattern of hybridisation with Phage 4 is quite dissimilar to that of Phage 1 and Phage 7. Although it shares the signals seen with the 1.9kb and the 1.2kb *EcoR I* fragments, it also hybridises to several unique bands. The 4.1kb , 3.9kb and the 3.5kb *EcoR I* fragments are identified from Phage 4. Northern and Reverse Northern analysis indicated high expression of the 4.1kb fragment (see section 2.3). The 3.2kb *EcoR I +Not I* junctional fragment of Phage 1 (equivalent to the 4.1kb *EcoR I* fragment of Phage 4) also hybridised to the total Phage 4 exon-trapped product. As would be predicted, this Phage 1 fragment did not hybridise to the PCR products of exon-trapping derived from Phage 1 or Phage 7. This is a result of the fragment not being present in Phage 7 and as only *EcoR I* fragments have been ligated into the exon-trapping vector the *EcoR I +Not I* junctional fragment of Phage 1 was not cloned into the exon-trapping vector. Therefore any functional exon within this Phage 1 junctional fragment would only be isolated from the Phage 4 4.1kb *EcoR I* fragment. In all, the probing of the phage blots with the PCR products of exon-trapping suggests that there are functional exons spread across the entire 32.8kb of isolated DNA.

2.6.5 Analysis of expression of total exon products

In order to gain information on the expression patterns of these isolated exons, Northern blots of embryonic RNA were probed first. Northern blots of total RNA from the head and body of embryonic day 15.5 mice, were probed with randomly labelled total secondary PCR product. The Northern blots presented in Figure 22 represent one week long exposures. Phage 7 derived exon-trapping products hybridise to several transcripts. A signal can be clearly seen at 500 nucleotides and another clearly defined at 800 nucleotides. Both of these transcripts are found in both head and body RNA. There are also several other signals on the blot. One appears to shadow the 28S Ribosomal RNA band in both head and body, the probe also hybridises to a large transcript.

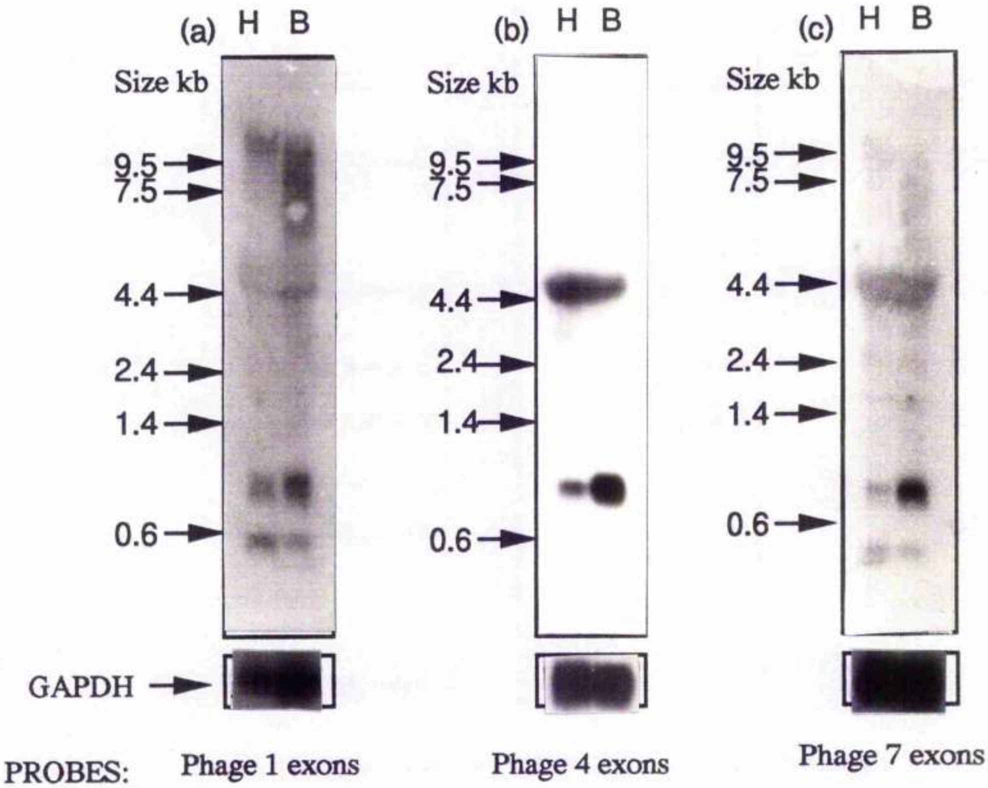


Figure 22: Expressional analysis of exons trapped from the cloned genomic DNA in development.

Northern blots of day 15.5 embyonic RNA probed with the total secondary PCR products from exon-trapping. Each lane represents 10μg of total RNA prepared from either embryonic day 15.5 head or body electrophoresed through a formaldehyde denaturing gel, Northern blotted and hybridised with; (a) Total secondary PCR products derived from Phage 1, (b) Total secondary PCR products derived from Phage 4, (c) total secondary PCR products derived from Phage 7. Lane 1 is embryonic day 15.5 head RNA and lane 2 is embryonic day 15.5 body RNA.

Note: Phage 1 and Phage 7 exons hybridise to a 0.8kb and 0.5kb transcript, while exons from phage 4 hybridise to the 0.8kb transcript. Exons from all Phage appear to hybridise to transcripts at about the 28S ribosomal RNA band.

The pattern of hybridisation seen with Phage 1 is similar to that of Phage 7. The bands seen at 500 nucleotides and 800 nucleotides are clearly visible as are the signals shadowing the 28S Ribosomal band and those found higher up on the blot. However, here there appears to be a doublet at 800 nucleotides. This may be an artifact of the Northern blot or it could represent two RNA transcripts, which hybridise with exons with sequence derived from Phage 1 *EcoR I* fragments. Phage 4 exon-trapped products hybridise to the 800 nucleotide transcript much more intensely than either Phage 1 or Phage 7. The 800 nucleotide message may be a doublet, but the resolution of the gel makes this difficult to determine. Like Phage 1 and Phage 7; Phage 4 derived exons also hybridise to the 500 nucleotide message but less intensely than observed with Phage 1 and Phage 7 (this is not apparent from the photograph in Figure 22). Also like Phage 1 and Phage 7 products there is a strong signal that corresponds to the 28S Ribosomal RNA, but no signal is seen at 10kb .

In order to rule out the possibility that the vector sequences within the PCR products are fortuitously hybridising to the blots, the secondary PCR product produced by the vector control was used to probe a Northern blot. This Northern blot contained total RNA from embryonic mice at different stages of development and from several post-natal tissues. The signal which appears to run with the 28S Ribosomal RNA, can be readily discerned (data not shown). Therefore this signal appears to be of vector origin as the probing 'conditions' are specific. However, none of the other signals is detected. This demonstrates that the transcripts at 500 nucleotides and 800 nucleotides observed with the mouse genomic derived secondary PCR products are real and not artifacts of the exon-trapping vector. Thus the products from all three phage contain functionally expressed exon(s) and are expressed during development.

The pattern of expression during development and in post-natal tissues was analysed for the exons from Phage 7. Figure 23(a) shows a blot of embryonic RNAs and post-natal RNAs probed with total exons from Phage 7. The 800 nucleotide message can be seen in the body sample from day 12.5, it is also present in the RNA samples from the bodies of day 13.5, day 14.5 and day 15.5 mouse embryos. The 800 nucleotide message also appears to be present in the head sample of day 14.5, but the signal is quite faint. The 800 nucleotide message is present in post-natal heart and lung, the signal does not appear to be present in the skin, kidney and brain sample. From this blot it is difficult to say whether the 500 nucleotide message is present in any of the samples but there appears to be a diffuse signal at about 500 nucleotides in size in some of the samples. Whether this is a *bona fide* message or not will require further analysis.

There is a band observable at 5kb in the body RNA sample from day 13.5 embryos with the Phage 7 exons, this message is not observed in any of the post-natal tissue samples. This indicates that sequences within the Phage 7 exons hybridise to a 5kb message that is expressed only at day 13.5 in the body of embryonic mice (in this analysis). This 5kb message may be expressed at other stages of development or at levels not detected in this analysis which would require a more rigorous analysis than that attempted here to detect it.

The pattern of expression in development and post-natally of the Phage 4 exons was also analysed as shown in Figure 23(b). As can be seen in Figure 23(b) the 800 nucleotide message is detected in the body of day 12.5 RNA sample and appears to increase in the level of its expression in day 13.5, day 14.5 and 15.5 body RNA samples. The 800 nucleotide message is also present in the day 16.5 and 20.5 body RNA samples but due to loading differences it is difficult to state whether there is a change in the level of expression. The 800 nucleotide message is also detected in the head RNA samples from day 13.5 onwards, it also appears to show an increase in the level of expression with an increase in age to day 15.5 of embryonic development.

In the post-natal tissues the 800 nucleotide message was detected in the bone, spleen and stomach RNA samples. Due to great variations in the amount of RNA in these post-natal samples the 800 nucleotide message may also be expressed in other tissues but it will not be detected in this analysis. None of the other two

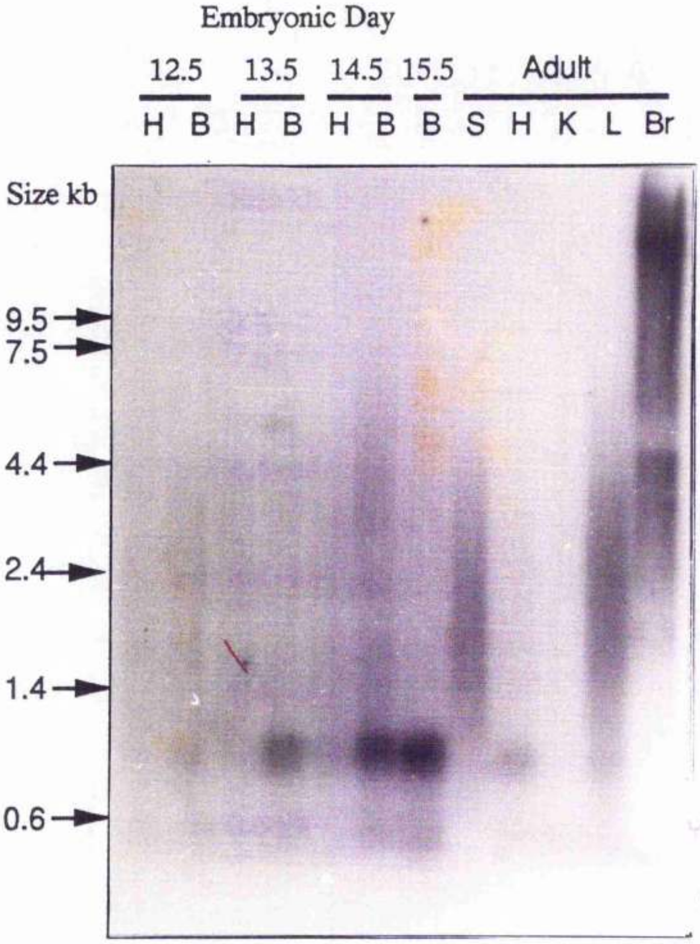


Figure 23 (a): Analysis of expression of the secondary PCR products from Phage 7 during mouse embryonic development and post-natally.

Total RNA was prepared from the heads (H) and bodies (B) of mouse embryos at different stages of embryonic development indicated above, and tissue samples of adult mice. Tissues represented are S, adult skin; H, adult heart; K, adult kidney; L, adult lung; and Br, adult brain.

10μg samples of total RNA were electrophoresed through a denaturing formaldehyde gel, blotted and probed with radio-labelled total secondary PCR products of exon-trapping from Phage 7.

Note: the 0.8kb transcript seen in Figure18 is present, however due to the poor quality of the blot,it is difficult to state whether the 0.5kb transcript is present or not.

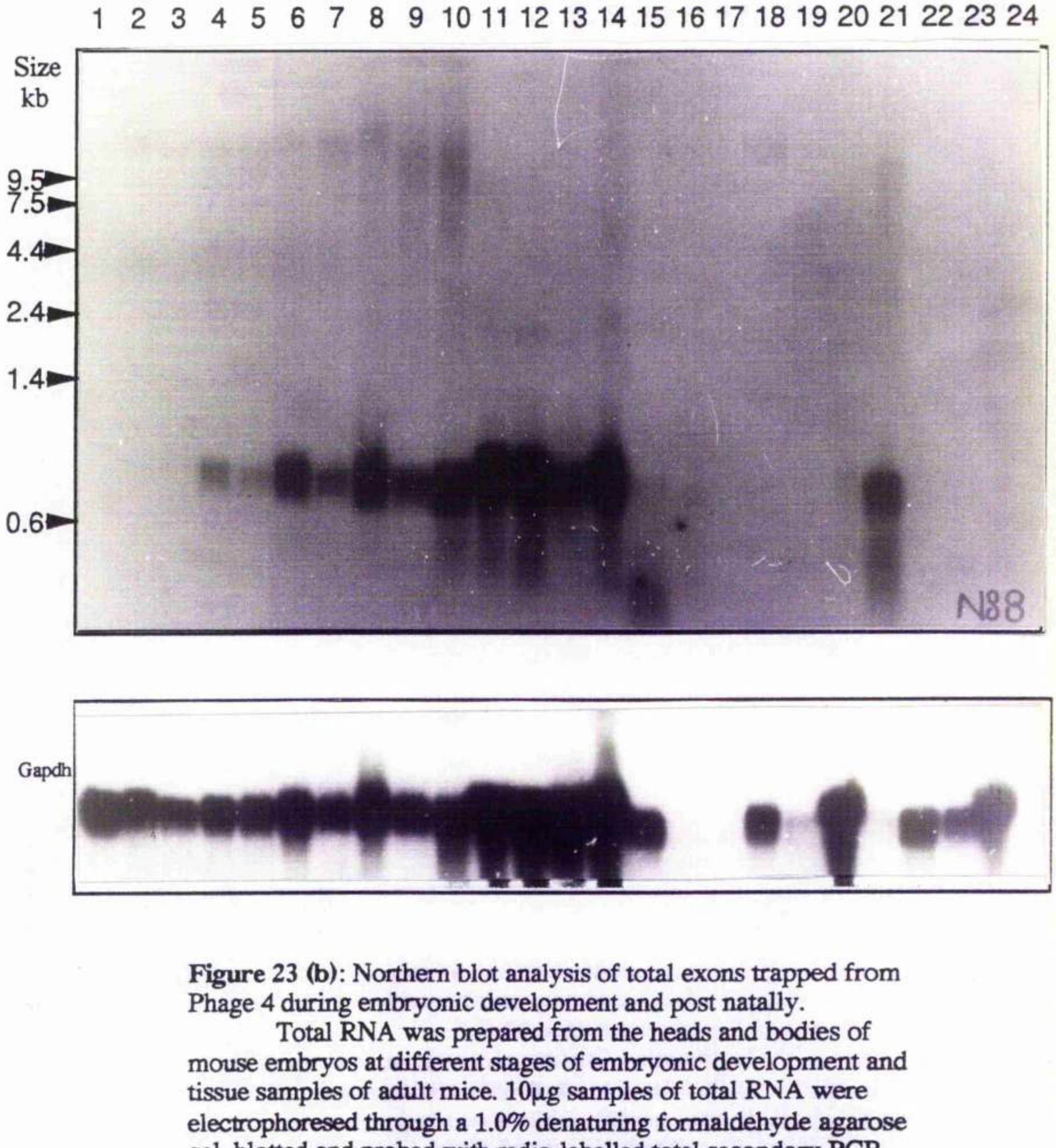


Figure 23 (b): Northern blot analysis of total exons trapped from Phage 4 during embryonic development and post natally.

Total RNA was prepared from the heads and bodies of mouse embryos at different stages of embryonic development and tissue samples of adult mice. 10μg samples of total RNA were electrophoresed through a 1.0% denaturing formaldehyde agarose gel, blotted and probed with radio-labelled total secondary PCR products of exon-trapping from Phage 4. Lanes 1, 3, 5, 7, 9, 11 and 13 represent total RNA from the heads of day 11.5, day 12.5, day 13.5, day 14.5, day 15.5, day 16.5 and day 20.5 embryonic mice respectively. Lanes 2, 4, 6, 8, 10, 12 and 14 represent total RNA from the bodies of day 11.5, day 12.5, day 13.5, day 14.5, day 15.5, day 16.5 and day 20.5 embryonic mice respectively. Lane 15, adult bone; lane 16, adult kidney; lane 17, adult liver, lane 18, adult lung; lane19, adult skin and lane 20, adult spleen; lane 21, adult stomach; lane 22, adult testes; lane 23, adult thymus and lane 24, adult spleen.

Lower panel shows Northern blot probed with Gapdh as a loading control.

messages (5kb and 500 nucleotides) were detected in this analysis, indicating that exons trapped from Phage 4 only hybridise to the 800 nucleotide message.

2.6.6 Differences in expression of exons between transgenically positive and transgenically negative RNAs

The 32.8kb stretch of DNA from which the exons are derived was isolated on the basis that it was disrupted in the cleft palate transgenic mouse. Two novel transcripts were identified at 4.5kb and 10kb in transgenically positive tissues. One of these was too large to be transgene sequence alone and must therefore be a hybrid of transgene and cellular sequence. This suggested that a cellular locus near to the site of transgene insertion was affected in transgenically positive mice.

In order to address if these cellular sequences within the 10kb hybrid RNA were present or if the secondary PCR products were affected in the transgenic line, transgenically positive RNAs were tested by hybridisation with labelled exon-trapped products. Figure 24 shows total RNAs derived from the kidney and the spleen of transgenically positive and transgenically negative mice, hybridised with the total exon-trapped products derived from Phage 7. An 800 nucleotide product is seen in the spleen of both transgenically positive and negative mice. It does not appear as a doublet as the transcript identified is not discrete. No change in size of the transcript is detectable between the transgenically positive or transgenically negative spleen RNA samples. Neither of the transgenically negative or transgenically positive kidney samples appear to express the 800 nucleotide transcript. The smaller 500 nucleotide transcript is not detected in any of the samples. Several larger transcripts are observed in some of the samples, though their pattern of expression is not consistent in the different samples between transgenically positive and negative RNAs. This erratic pattern of expression means that nothing conclusive can be drawn from their presence on the blot. The hybrid transcripts were not detected, this may be a result of the low level of expression of the hybrid transcripts. It may also be a result of none of the

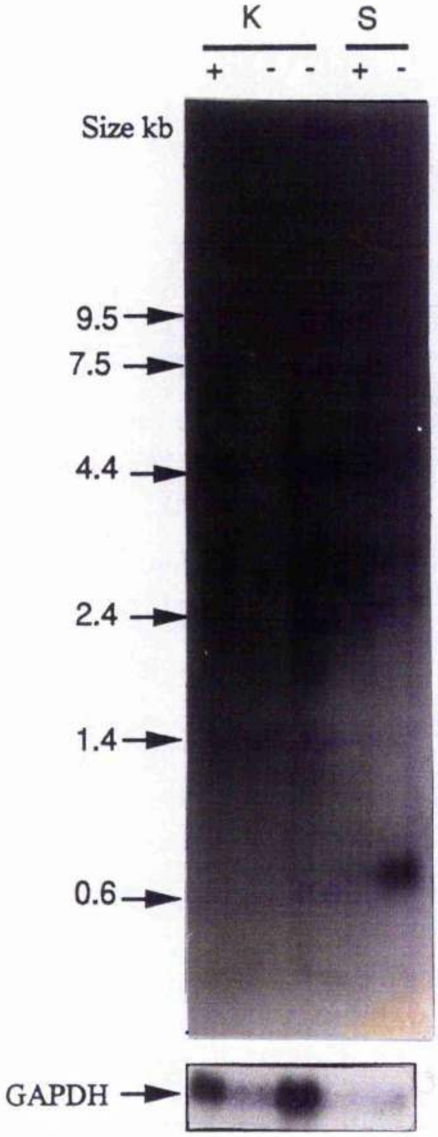


Figure 24: Analysis of expression of secondary PCR products between post-natal transgenically positive and transgenically negative tissues.

Total RNA was prepared from the kidneys (K) and spleens (S) from transgenically positive(+) and transgenically negative (-) age matched mice. 10μg samples of total RNA were electrophoresed through denaturing formaldehyde gels, Northern blotted and probed with radio-labelled total secondary PCR products of exon-trapping from Phage 7. Two transgenically negative kidney RNA samples are shown above.

Loading was controlled for by probing the Northern with GAPDH. Note: the 0.8kb transcript detected in heart RNA in Figure 18 is detected in both the transgenically positive and transgenically negative Spleen samples, but is not detected in any of the kidney samples.

trapped exons being part of the hybrid transcripts, this may indicate that there are different transcripts derived from this locus.

The secondary PCR products from the exon-trapping, when hybridised to total RNAs from transgenically positive and transgenically negative tissue, revealed an upregulation in the level of the 800 nucleotide transcript in the transgenically positive spleen. Neither of the 10kb or 4.5kb hybrid transcripts was found in either transgenically positive RNA sample and no signal was observed in either the transgenically positive or transgenically negative kidney samples.

2.7 Characterisation of individual exons from total PCR products

Having shown that sequences contained in the secondary PCR products are part of a transcription unit expressed from mouse genomic DNA, individual exons from the pools were isolated. The secondary PCR products from the individual phage were cloned into pAMP10 (Materials and Methods). Ligated products were transformed into XL1-Blue, for each pool of ligated products about 100 transformed colonies were plated and 10-20% of these transformants were analysed. Thirteen white colonies (i.e. recombinants) from Phage 7, fourteen white colonies from Phage 1 and nineteen white colonies from Phage 4 were selected at random initially for analysis by Colony PCR (Colony PCR was carried out as in Materials and Methods). By this method the presence of inserts and their size can be discerned. The products of Colony PCR were electrophoresed on 2% agarose gels, then blotted. To identify those products that contained sequence derived from the 30kb of mouse genomic DNA, these blots were then hybridised with entire radiolabelled recombinant phage DNA. Each set of subcloned PCR products was hybridised with the respective phage from which they were derived.

Figure 25 (a) shows the result of hybridisation of Phage 4 to Colony PCR products from Phage 4. Out of the nineteen colonies from Phage 4, fifteen hybridised with mouse genomic sequences. In this case fourteen of the PCR products were about 250bp in size. 177bp of the 250bp is vector sequence due to the location of the

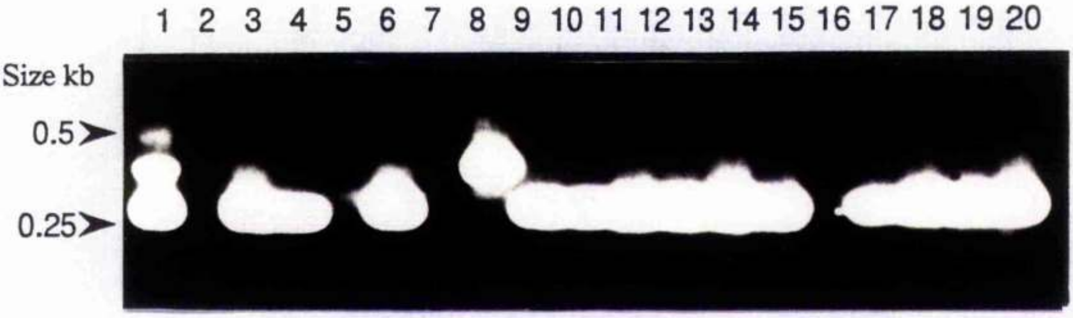


Figure 25 (a): Isolation of individual exons trapped from Phage 4 cloned genomic DNA.

Total secondary PCR products trapped from Phage 4 were cloned and transformants analysed using colony PCR. A fifth of the PCR reaction was electrophoresed on 2% agarose gels, Southern blotted and hybridised with radio-labelled Phage 4 total cloned genomic DNA.

The contact print of the autoradiograph shown represents some of the cloned exon-trapping PCR products form Phage 4 *EcoR I* fragments probed with Phage 4 total genomic DNA. Each lane is an individual colony PCR product from the cloned products of exon -trapping. On the basis of this analysis individual colonies were chosen and their plasmids sequenced.

The product from lane 8 was found to be a 206bp exon and designated exon p51.The PCR product in lane 1, 3, 4, 6, 9, 10, 11, 12, 13, 14, 16, 17, 18 and 19 were sequenced and found to be identical to each other, this was designated exon p52.

primers used for the Colony PCR this would result from an exon of approximately 80bp in size. The remaining PCR product hybridizing with mouse genomic sequences was just under 400bp. Again allowing for the 177bp of vector sequences this would result from an exon of 220bp in size. Out of the fourteen colonies from Phage 1 tested, only one appeared to be positive. The PCR fragment is 600bp in size, 177bp of which is vector sequence, therefore, this PCR fragment contains a putative exon of approximately 450bp. Similarly for the thirteen colonies tested from Phage 7, only one hybridised with mouse genomic sequences. This PCR fragment was also 600bp, similar to that isolated from Phage 1 which again would mean a putative exon of 450bp (data not shown).

It was possible that not all the trapped exons were being identified. A total of about 300 colony PCR products were screened using individual phage fragments. Those phage fragments which had been identified previously as containing exon sequences were used. Using this approach it was possible to identify another six clones. Figure 25 (b) shows the same Southern blot of PCR products probed with two different Phage derived fragments. Although they reveal PCR products of a similar size, the pattern of hybridisation is completely different. The pattern of hybridisations in Figure 25 (b) indicated that a single Colony PCR product was derived from the 1.9kb *EcoR I* fragment of Phage 7, while four of the Colony PCR products are derived from the 1.25kb *EcoR I* + *Not I* fragment from Phage 4.

Since different exons may be the same size, several of the clones of each size were sequenced. Twenty-five subcloned secondary PCR products were sequenced in total. Out of these twenty-five sequences five novel exons were identified. For each exon the location on the 32.8kb of DNA was ascertained, patterns of expression characterised, and conservation across species investigated. The novel exons were assigned the names of p50, p51, p52, p56 and p57 for the purposes of reference only.

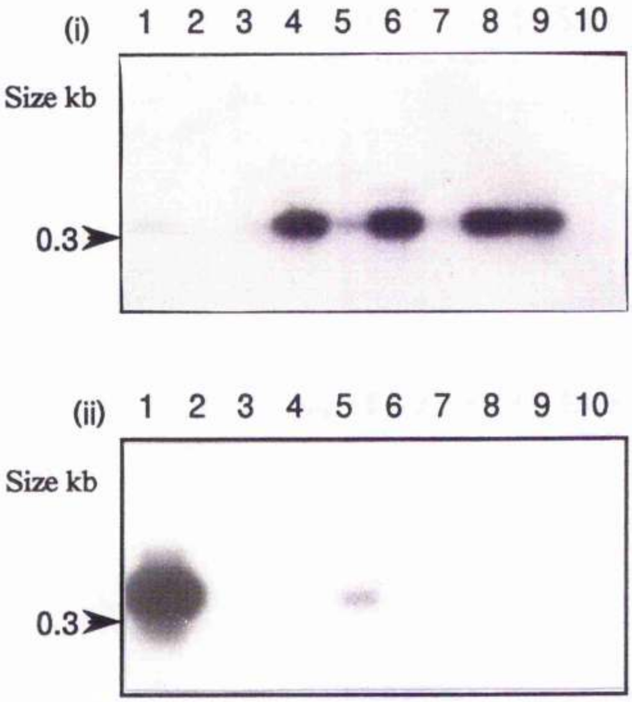


Figure 25 (b):(i) Isolation of exon p56, **(ii)** Isolation of exon p57.

Colony PCR was performed on cells transformed with ligations of total secondary PCR product from Phage 1. One fifth of each PCR reaction was electrophoresed on a 2% agarose gel, blotted and hybridised with (i) the 1.2kb *EcoR I* and 1.25kb *EcoR I+Not I* fragments from Phage 4 and (ii) the 1.9kb *EcoR I* fragment that is found in all Phage clones. The autoradiographs of these hybridisations are shown above.

Several PCR products were found to hybridise to these probes, isolation and sequencing of all the identified PCR products in (i) revealed that they were the same sequence and this sequence was designated exon p56. One PCR product was found to hybridise to the 1.9kb *EcoR I* fragment from Phage 7 (ii), it was isolated, sequenced and designated exon p57.

2.8 Characterisation of trapped exons

The following results have been sub-divided into several sections. Each section deals with an individual exon and the results contained within each section follow the same layout. Firstly, the sequence analysis for each exon and any database information. The sequence analysis for each exon includes motif searches on the GCG package. This was carried out for both nucleic acid sequence and predicted amino acid sequence of ORFs. Motif searches are, of course, limited by the motifs present in the database and the only matches acknowledged are those which exhibit 100% homology. The sequence analysis will be followed by the mapping of the exon onto the phage map. There then follows Southern analysis of conservation of the exon across species and expression analysis. All the techniques used are detailed in the Materials and Methods.

2.8.1 Characterisation and expressional analysis of exon p50

Exon p50, isolated from the Phage 4 exon pool, was sequenced using double stranded dideoxy sequencing (Sanger *et al.*, 1977) with primers to the T7 and SP6 promoters present in the vector flanking the multiple cloning site. Exon p50 is composed of 446bps (see Figure 26). When analysed using MacVector^{Tm} or the GCG package several Open Reading Frames (ORF) were found. None, however, was complete across the exon. This analysis did not reveal any nucleic acid motifs. Translation of the ORFs did not reveal any protein motifs in the deduced amino acid sequence when compared to the GCG package Motifs program. A search of the BLASTN database revealed that the sequence of exon p50 contained two major LINE-1 sequence repeats that constitute over 95% of the exons sequence. LINE-1 repeat sequences are widespread in the mouse genome (Charlesworth *et al.*, 1994). They are found in the sequence of some retrotransposable elements. The database

Figure 26: Sequence analysis of Exon p50

Nucleotide Sequence of Exon p50: oligos used to prime PCR are indicated by bold type and underlining.

```
        10           20           30           40           50
TCATGGAAGA   AGTACAGAGA   CAAAGTTTGG   AGCTGAGATG   AAAGATGACC
AGTACCTTCT   TCATGTCTCT   GTTTCAAACC   TCGACTCTAC   TTTCTACTGG
        60           70           80           90          100
ATCTAGAGAC   TGCCACACCC   GGGGATCCAT   CCCATAATCA   GCCTCCAAAC
TAGATCTCTG   ACGGTGTGGG   CCCCTAGGTA   GGGTATTAGT   CGGAGGTTTG
       110          120          130          140          150
GCTGACACCA   TTGCATACGC   CAGCAAGATT   TTGCTGAAAG   GACCCTGATA
CGACTGTGGT   AACGTATGCG   GTCGTTCTAA   AACGACTTTC   CTGGGACTAT
       160          170          180          190          200
TAGCTGTCTC   TTGTGAGGCT   ATCTGGTGCC   TGGCAAACAC   AAGTGGATGC
ATCGACAGAG   AACACTCCGA   TAGACCACGG   ACCGTTTGTG   TTCACCTACG
       210          220          230          240          250
TCACAGTCAT   CTATAAGATG   GAACACAGGG   CCCCCAGTGG   AGGAGCTAGA
AGTGTCAGTA   GATATTCTAC   CTTGTGTCCC   GGGGGTCACC   TCCTCGATCT
       260          270          280          290          300
GAAAGTACCA   AAGGAGCTAA   AGGGATCTGC   AACCCTATAG   ATGAACAACA
CTTTCATGGT   TTCCTCGATT   TCCCTAGAGG   TTGGGATATC   TACTTGTTGT
       310          320          330          340          350
ATATGAACTA   ATCAGTACCC   CCGAGATCGT   GTCTCTAGCT   GCATATGTAT
TATACTTGAT   TAGTCATGGG   GGCTCGAGCA   CAGAGATCGA   CGTATACATA
       360          370          380          390          400
CAGAAGATGG   CCTAGTCTGC   CATCACTGGG   AAGAGAGGCC   CTTTGGTCTT
GTCTTCTACC   GGATCAGACG   GTAGTGACCC   TTCTCTCCGG   GAAACCAGAA
       410          420          430          440          450
GCAAACTTTA   TATGCCTCAG   TACAGGGGAA   TGCCAGGGCC   AAGAAG
CGTTTGAAAT   ATACGGAGTC   ATGTCCCCTT   ACGGTCCCGG   TTCTTC
```

search revealed that these LINE-1 sequences are also present in the 5' and 3' untranslated regions of several genes, e.g. *Mus muscaris* bradykinin B2 Receptor gene sequence (Ma *et al.* 1994) and the mouse interleukin 6 (IL-6) gene (Tanabe *et al.,* 1988).

Exon p50 was mapped onto the restriction map of the 32.8kb of isolated mouse genomic DNA (Figure 27). This was achieved by probing a Southern blot of phage DNA digested with *EcoR I* with an exon p50 probe. The exon p50 probe hybridised to the 6kb and the repetitive 3kb *EcoR I* fragments found in Phage 1 and Phage 7, it also hybridised to the 4.1kb *EcoR I* fragment found in Phage 4. The same fragments were identified by Reverse Southern analysis and by Reverse Northern analysis. The fragments identified appeared to be the most highly repetitive and also the most highly expressed. The fact that these fragments probably contain LINE-1 repetitive sequence helps to explain previous observations.

Mapping is, however, complicated by this result and by this approach alone it is not possible to ascertain which of these two fragments the exons are derived from. Exons derived from Phage 4 did not exhibit the pattern of hybridisation seen with exon p50, when used as a probe on a Southern blot of phage DNA digested with *EcoR I* (Figure 21), suggesting that although LINE-1 sequences are present in the Phage 4 they are not contained in an exon trapped in this analysis. Exons derived from Phage 1 and Phage 7 did exhibit this pattern of hybridisation. It appears then, that exon p50 resides in either the 6kb *EcoR I* fragment or the 3kb *EcoR I* fragment of Phage 1 and Phage 7.

If the sequence of exon p50 is part of an 'important' transcript, then it may show conservation across species. The conservation of the sequence of exon p50 was analysed by hybridisation to a Zoo blot (Figure 28). Figure 28 shows that this sequence is highly repetitive in mouse and no discrete band is observable. It is also highly repetitive across species as rat, hamster and human all exhibit a repetitive pattern at low temperature and low stringency. At a higher temperature of 65°C, the repetitive signal is still apparent in the mouse and human DNA, though the rat and

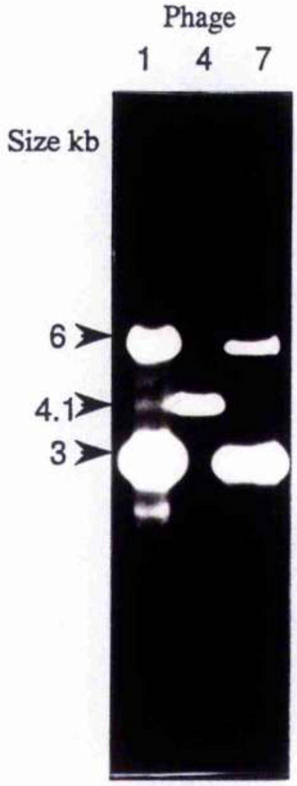


Figure 27: Location of exon p50 in cloned genomic DNA.

Contact print of the autoradiograph of a Southern blot of cloned genomic DNA probed with radio-labelled exon p50. Each lane represents 1μg of each genomic clone digested with *EcoR I* + *Not I*; lane 1, Phage 1; lane 2, Phage 4; lane 3, Phage 7. Samples were electrophoresed through an 0.8% agarose gel, Southern blotted and probed.

Note: exon p50 identifies the 3kb and 6kb *EcoR I* fragments from Phage 1 and Phage 7 and the 4.1kb *EcoR I* fragment from Phage 4.

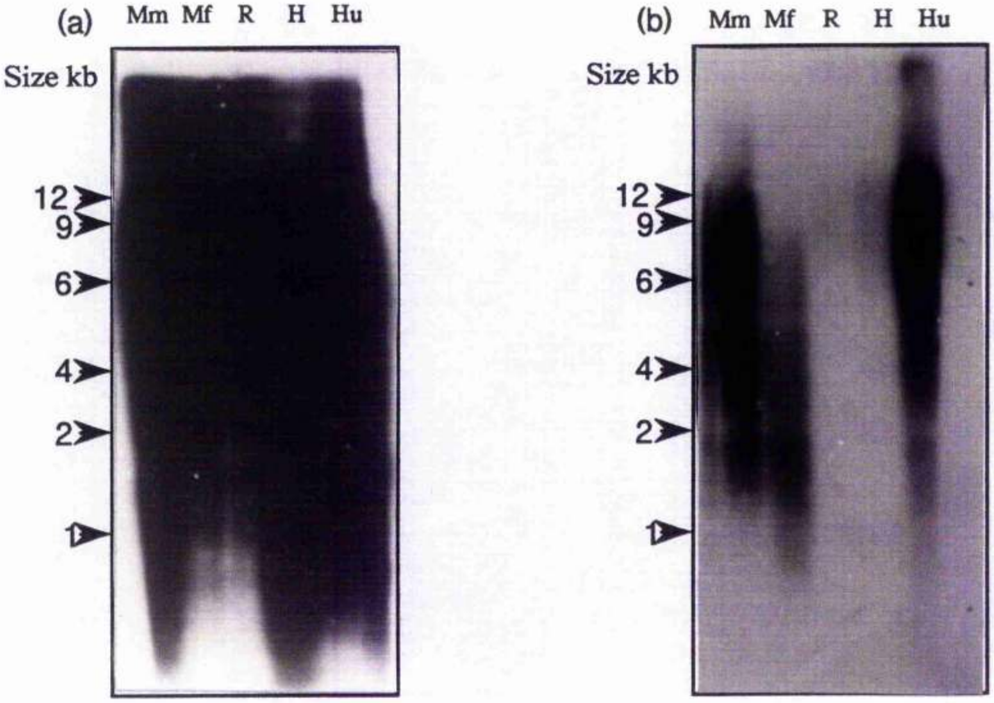


Figure 28: Zoo blot analysis of DNA isolated from several species hybridised with exon p50.

5μg of DNA from each species was digested with *EcoR I*, electrophoresed through a 0.7% agarose gel, Southern blotted and probed with exon p50. The samples used are Mm, mouse male; Mf, mouse female; R, rat; H, hamster and Hu, human. Hybridisations were carried out at 55°C and 1x Church Buffer overnight. Blots shown were washed to a final stringency of (a) 55°C and 1xSSC, and (b) 65°C and 1xSSC.

Note: that exon p50 hybridises to the mouse and human DNA samples at the higher stringency, exon p50 does still hybridise to the rat and hamster DNA samples but the signal is greatly reduced at this higher stringency.

hamster samples show a much reduced signal. In the hamster lane there appears to be a band at about 6kb . When the stringency was increased, although the high level of repetitive signal in mouse and human DNA decreased, no obvious bands are observable. This result is consistent with the sequence data.

Expressional analysis was carried out. The exon p50 sequences were hybridised to Northern blots of mouse embryonic day 15.5 total RNA. As Figure 29 (a) shows, the sequences contained within exon p50 are highly expressed at day 15.5 of mouse embryonic development. However, no single transcript is detectable. This is not surprising considering the number of LINE-1-like repetitive elements dispersed throughout the mouse genome. The association of these LINE-1 sequences with many genes explains the high level of expression detected by Northern analysis. When a similar Northern blot of mouse embryonic RNAs was probed using the total exon-trapped secondary PCR products derived from the original phage, several signals were detected at >10kb (see section). These large signals were only observed on the Northern blots probed with the exon-trapped products of Phage 1 and Phage 7. The pattern of expression exhibited by exon p50 at day 15.5 of mouse embryonic development identifies these larger transcripts, although the intensity is much stronger. This difference is probably due to the reduced representation of exon p50 in the total secondary PCR product. This would reduce the specific activity of the labelled fragment and consequently produce a much reduced signal on a Northern blot

2.9.1 Characterisation of exon p51 sequences

That exon p51 from Phage 4 was confirmed by hybridisation of Phage 4 sequences to exon-trapped sequences that had been subcloned and analysed by colony PCR. Both strands were sequenced by dideoxy sequencing (Sanger *et al.*, 1977), using primers to the flanking T7 and SP6 promoters present in the cloning vector. Sequencing of exon p51 showed that it is 205bp (Figure 30). When this 205bp sequence was analysed using MacVector^{TM} and the GCG package no nucleic

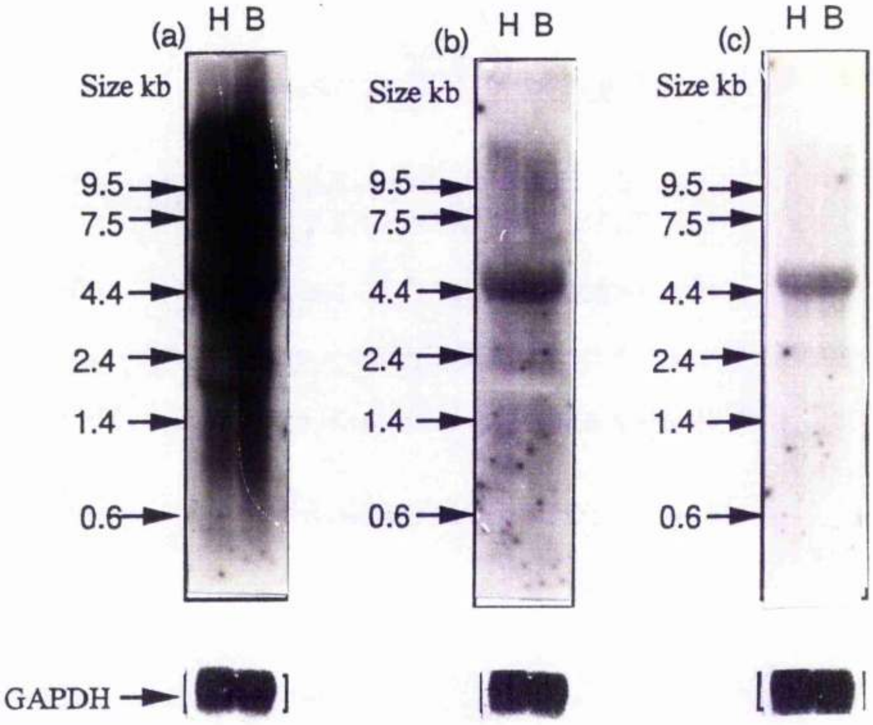


Figure 29: Analysis of exon p50, exon p51 and exon p52 expression during mouse embryonic development.

Total RNA isolated from the head (H) and body (B) of embryonic day 15.5 mice was electrophoresed and hybridised at 65°C in 1x Church Buffer overnight with radio-labelled (a) exon p50, (b) exon p51 and (c) exon p52. Blots were washed at 65°C and 0.1xSSC and 0.1% SDS.

Note: exon p50 hybridises strongly with many RNA species, while exon p51 and p52 hybridise to an RNA species that runs at the 28S ribosomal RNA band.

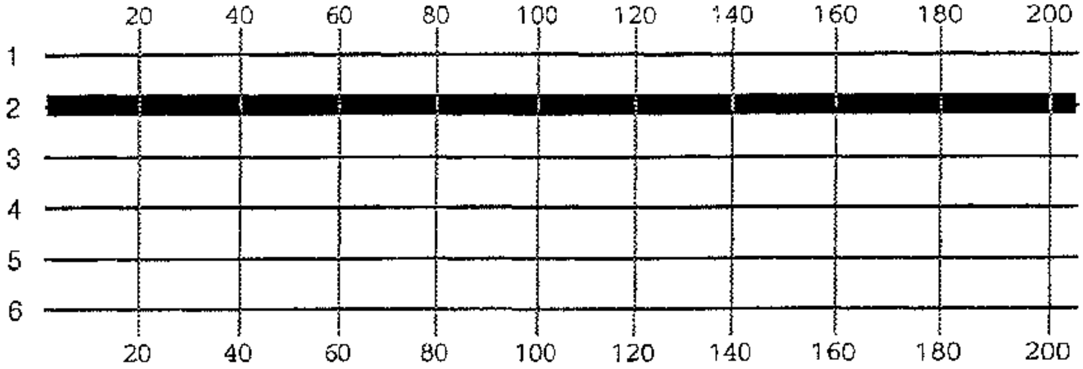
The lower panel shows the same blot probed with glyceraldehyde 3-phosphate dehydrogenase (Gapdh) as a loading control.

Figure 30: Sequence analysis of Exon p51

(a) Nucleotide Sequence: oligos used to prime PCR are indicated by bold type and underlining.

```
         10         20         30         40         50
GAGAATACGA  ACACGAACTA  TAACTTTCTC  GTCAAAAAAT  AGAGAGGAAA
CTCTTATGCT  TGTGCTTGAT  ATTGAAAGAG  CAGTTTTTTA  TCTCTCCTTT
         60         70         80         90        100
GAGTAATAGT  AAGAGGCGAT  GATGATAACC  ATAATCATAG  TAAGAAGTTT
CTCATTATCA  TTCTCCGCTA  CTACTATTGG  TATTAGTATC  ATTCTTCAAA
        110        120        130        140        150
AGTCACGTGA  AATTTGAATT  GAAATCAAAA  GAGGAACGTA  ACTTAAAGAG
TCAGTGCACT  TTAAACTTAA  CTTTAGTTTT  CTCCTTGCAT  TGAATTTCTC
        160        170        180        190        200
TCAACAAGAT  GACAAAGACC  TTCTACCTAA  AACATCCGTA  TTGCTAGGCG
AGTTGTTCTA  CTGTTTCTGG  AAGATGGATT  TTGTAGGCAT  AACGATCCGC
        210
AGACG
TCTGC
```

(b) Open Reading Frames of Exon p51



(c) Amino Acid Sequence Of Translated ORF From Exon p51

```
         10         20         30         40         50
RIRTRTITFS  SKNREERVIV  RGDDDNHNHS  KKFSHVKFEL  KSKEERNLKS

         60         70
QQDDKDLLPK  TSVLLGET
```

acid sequence motifs were identified. A search of the BLASTN database with the nucleic acid sequence revealed several similarities to the sequence of exon p51. No other matches produced 100% identity, at the nucleotide level, though a number of sequences exhibited a level of 70-80% homology across short stretches of nucleic acid sequence. Of these sequences, a *T.brucei* variable surface glycoprotein gene with 80% identity across 60 bases (Young *et al.*, 1987), a rice partial cDNA with 71% identity across a stretch of 48 bases (Sasaki and Minobe, unpublished) and the rat *Map2* gene for microtubule associated protein with 76% identity across 41 bases (Kindler *et al.*, 1990) were the best matches found in the databases.

Using MacVector^{TM} a single ORF that spanned the sequence was identified (Figure 30 (b)). The ORF was translated using the GCG package (see Figure 30 (c)). The predicted amino acid sequence of the ORF is 68 amino acids in length. A search of the BLASTP databases using this peptide sequence did not reveal any high level of similarity to other peptide sequences across its entire length. Short stretches of <10amino acids did show very high levels of identity. These matches were too numerous and the types of genes too diverse for anything to be concluded. Several larger peptide sequences however did exhibit a low level of similarity. Macrogolgin, a human structural protein, exhibited a 25% identity across 51 amino acids of the translated exon (Seelig *et al.*, 1994). When similar amino acid substitutions are included in the analysis, the level of similarity increased slightly to 28%. Another amino acid sequence which exhibited a low level of similarity was neurofilament triplet M protein from the Pacific electric ray (Linial and Scheller, 1990). Over a 50 amino acid sequence it showed a 29% identity and a similarity of 43% when similar amino acid substitions are included.

2.9.2 Mapping of exon p51 onto mouse genomic DNA

To identify where on the mouse genomic DNA exon p51 is located, it was hybridised to a Southern blot of phage DNA digested with *EcoR I* (Figure 31). Exon p51 was found to reside on the 3.9kb fragment of Phage 4, though it was quite

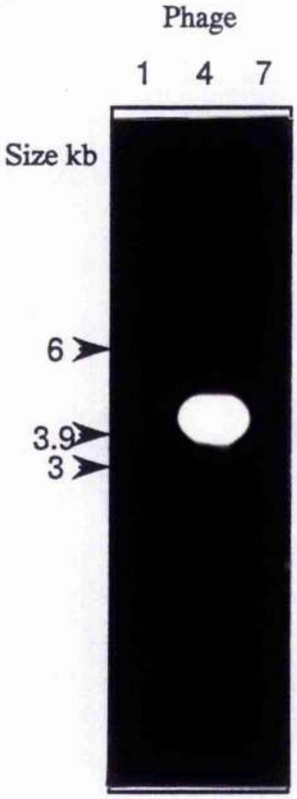


Figure 31: Location of exon p51 in cloned genomic DNA.

Contact print of the autoradiograph of a Southern blot of cloned genomic DNA probed with radio-labelled exon p51. Each lane represents 1μg of each genomic clone digested with *EcoR I* + *Not I*; lane 1, Phage 1; lane 2, Phage 4; lane 3, Phage 7. Samples were electrophoresed through an 0.8% agarose gel, Southern blotted and probed.

Exon p51 hybridises to the 3.9kb *EcoR I* fragment from Phage 4.

difficult to be sure that it only hybridised to the 3.9kb fragment due to the strength of the signal. The 3.9kb *EcoR I* and the 4.1kb *EcoR I* subclones were digested with several restriction enzymes, electrophesed on an agarose gel, blotted and probed with exon p51. Exon p51 was found to hybridise to the 3.9kb *EcoR I* fragments subfragment and not to the 4.1kb *EcoR I* fragment digested with several restriction enzymes. Exon p51 is trapped from the 3.9kb and not the 4.1kb *EcoR I* fragment.

2.9.3 Analysis of conservation of the exon p51 sequence across species

Exon p51 was hybridised to a Zoo blot (data not shown). The Zoo blot contained DNA digested with *EcoR I* from mouse, rat, hamster and human. Although random labelling of the exon produced probes of high specific activity, when these probes were hybridised to the Zoo blots they produced no clear signal upon washing. In view of the ability of the exon p51 successfully to hybridise to Southern blots of the phage fragments that it was derived from, it was unclear as to why this should be. 177bp of the exon-trapped products is derived from the vector. This possibly results in a probe too small for genomic blots (though fine for blots of cloned DNA). This added to the number of sequences available for hybridisation present in the genomic DNA samples may explain the lack of hybridisation.

2.9.4 Expressional analysis of exon p51

Exon p51 was hybridised to a Northern blot of mouse embryonic day 15.5 total RNA (Figure 29(b)). Exon p51 did not hybridise to the 500 and 800 nucleotide transcripts detected when the total secondary PCR product was hybridised to Northern blots of embryonic RNA, even after the blot had been exposed for over seven days. The nonspecific signal is seen at about the 28S Ribosomal RNA band. A fainter signal is observed at >10kb, in both head and body RNA. This is also observed in other Northern blots probed with total PCR products. It may be that this is a real transcript and that it contains exon p51. Exon p51 is either not part of the 500 or the 800 nucleotide transcripts observed with the total secondary PCR product or the fragment is too small to make a good probe by random priming. This would seem to be true as it did not hybridise to the 3.9kb *EcoR I* on a Southern blot. The Northern blots that exhibited the 500 and 800 nucleotide transcripts had been probed with total secondary PCR products. Exon p51 may belong to the primary transcript that produces the 500 or 800 nucleotide mature mRNA and splicing removes it from the final transcript. Or exon p51 may belong to a completely separate transcript that has not been observed at this stage of development.

In view of the inability to detect any hybridisation of exon p51 to the Zoo blot there was the possibility that the probes were too small to detect the sequence on the blot. Each exon fragment used as a probe contains 177bp of vector sequence. Due to the nature of exon-trapping this 177bp of vector sequence cannot be easily removed by restriction digestion. To overcome this problem two other strategies were attempted to produce high specific activity probes. Primers were designed to the 5' and 3' ends of the exon sequence, PCR was then carried out (Materials and Methods). This PCR product should consist solely of exon p51 sequence with the vector sequences removed. Several templates known to contain the exon p51 were used in PCR reactions. No matter the reaction conditions used no PCR product could be identified, either by ethidium bromide staining of agarose gels or by hybridisation of exon p51 to blots of exon p51 PCR reactions. The reason why the PCR analysis failed is unknown. It is possible that there were tertiary structures in the DNA molecules that prevented *Taq* polymerase from progressing through the molecule, this could also be the reason for the poor probes generated. However exon p51 was sequenced using T7 polymerase with no difficulties and therefore it is unlikely that tertiary DNA structure is the reason why PCR analysis failed.

The oligonucleotides designed to the 3' and 5' ends of exon p51 for PCR (Figure 30) were end labelled and hybridised to Northern blots of embryonic RNAs and Zoo blots, in an attempt to overcome the problems caused by the small size of the probes used previously. No matter the hybridisation conditions nor the washing conditions used with the oligonucleotide probes, no signal could be detected on any of the blots. Due to pressures of time all the conditions used for this analysis could not be investigated.

2.10.1 Characterisation of exon p52

Exon p52, isolated from Phage 4, was identified by hybridisation of Phage 4 sequences to the exon-trapped products derived from Phage 4. Exon p52 was sequenced using the dideoxy sequencing method (Sanger *et al.*, 1977). Both strands were sequenced (Materials and Methods). The sequence information obtained showed that it is composed of 81 nucleotides, see Figure 32. Subjecting this 81 nucleotides of sequence to sequence analysis using MacVector^{TM} and the GCG package revealed no nucleic acid sequence motifs. Comparison of this sequence to those present in the BLASTN database revealed a small number of sequences with a high degree of similarity at the nucleotide level. Most of these sequences were derived from the Calcium/calmodulin-dependent protein kinase II gene (CaM kinase II), from three different species. They were, *Xenopus laevis* (Taylor, unpublished), *Mus musculus* (Karls *et al.*, 1992) and *Rattus rattus* (Bennett *et al.*, 1987). A comparison of the sequence of CaM kinase II gene from *Mus musculus* to that of *Rattus rattus* revealed that they were 97% identical at the nucleotide level across the entire length of the gene including the 3' untranslated region (Karls *et al.*, 1992).

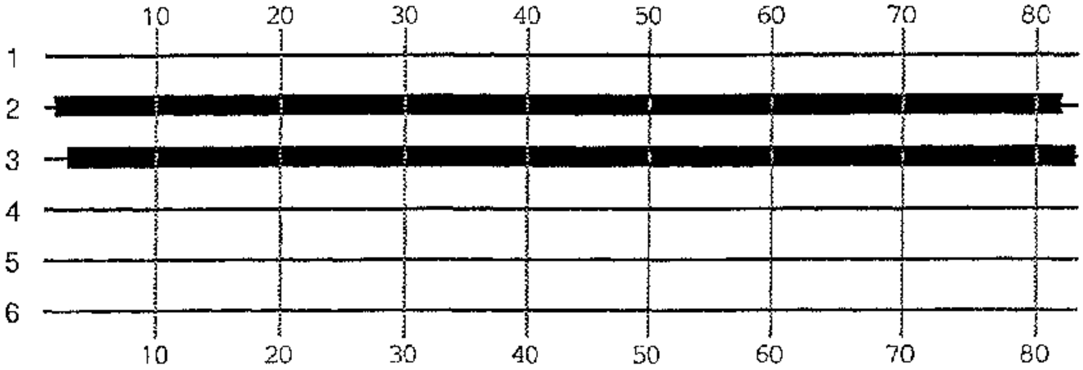
The sequences of CaM kinase II from the three species had exactly the same percentage similarity to the sequence of exon p52. Over a stretch of 70 bases, 47 were identical (68%) between the exon sequence and the database sequence. Although the size of exon p52 is very small, the level of identity is reasonably high. This might suggest that exon p52 sequence is part of a protein kinase gene sequence. Another match in the database search was to a human cDNA clone, from an unknown gene located as an expressed sequence tag (EST) from human brain (Adams *et al.,* 1993). The region of similarity was smaller, 60 bases, but the level of similarity was higher, 40 identical bases or 75%.

Figure 32: Sequence analysis of Exon p52

(a) Nucleotide Sequence: oligos used to prime PCR are indicated by bold type and underlining.

```
         10         20         30         40         50
GGGGATGGTT ATGCCTACAA AATCCATCTT CCAGAAACAG TAGAACAACT
CCCCTACCAA TACGGATGTT TTAGGTAGAA GGTCTTTGTC ATCTTGTTGA
         60         70         80
GAGGAAATTC AATGCAAGGA GAAAACTAAA G
CTCCTTTAAG TTACGTTCCT CTTTTGATTT C
```

(b) Open Reading Frames of Exon p52



(C) Amino Acid Sequence of Exon p52 ORFs

ORF2

```
         10         20
SGSLCLQNPS SRNSRTTEEI QCKAKTK
```

ORF3

```
         10         20
GDGYAYKIHL PETVEQLRKF NARRKLK
```

Using MacVector^{TM} two ORFs were found that spanned the entire exon see Figure 32(b), both of these ORFs shared the same orientation. Translation of these two ORFs produced two very different peptide sequences see Figure 32(c). The peptide sequences were named after the ORF they were derived from see Figure 32(b)&(c). Neither of these two peptide sequences contain structural protein motifs as shown by a peptide sequence motif search on the GCG package. A comparison of each peptide sequence to peptide sequences in the BLASTP database was performed.

ORF 2 was found to have a high degree of similarity at the amino acid level to the peptide sequence of calcium/calmodulin-dependent protein kinase II. Out of the twenty-seven amino acids present in the peptide ORF 2, sixteen of these were identical to part of the sequence of calcium/calmodulin-dependent protein kinase II (59% identity). This level of identity was found with CaM kinase II peptide sequences from Drosophilia, Human, rat, mouse and Xenopus (Hanson and Schulman, 1992) over the same region. This region of CaM kinase from each of the different species showed 90% identity to each other. The level of similarity increased to 97% when similar amino acid substitions are included.

A search of the BLASTP database with the deduced amino acid sequence of ORF 3 revealed that the deduced amino acid sequence of ORF 3 was similar to the Sex Determining Region Y protein. ORF 3 showed a lower level of identity, 43% identity across 23 amino acids to SRY amino acid sequence, than that exhibited by ORF 2 to calcium/calmodulin-dependent protein kinase II. The level of similarity of the deduced amino acid sequence for ORF 3 to SRY sequence, increased to 73% when similar amino acid substitutions are included. Again this level of similarity was consistent when ORF 3 was compared to the SRY peptide sequence from different species (Bianchi *et al.,* 1993 and Pailhouse *et al.,* unpublished). The region of the SRY protein that was found to be similar to the deduced amino acid of ORF 3 by the database searches, appears to be the HMG box of the protein. The deduced amino acid sequence of ORF 3 was found to be 36% identical and 63% similar across 25 amino acids of the 55 amino acids of the HMG box of the SRY protein from *Sus scrofa* (domestic pig) (Pailhouse *et al.*, unpublished).

The database searches using the nucleotide sequence and the deduced amino acid sequences from ORF 2 and ORF 3 are both informative and interesting in regards to the nature of exon p52 and the gene that exon p52 belongs to. Though the nucleic acid sequence is highly conserved in other species the presence of two complete ORFs gives us several possibilities regarding the nature of the gene(s) present at this locus. The results of the database search imply that either ORF may be part of a single gene or that each belongs to two separate transcripts, both sharing the same sequence at the nucleic acid level but different ORFs.

2.10.2 Mapping of exon p52 onto mouse genomic DNA

Exon p52 was labelled by the random priming method (Materials and Methods), then hybridised to a Southern blot of phage DNA digested with *EcoR I* see Figure 33. Exon p52 was found to be located on the 3.9kb *EcoR I* fragment of Phage 4. As with exon p51, The 3.9kb *EcoR I* fragment was digested with several restriction enzymes, run on an agarose gel and blotted, then probed with exon p52. Exon p52 resides on the 3.9kb *EcoR I* (data not shown). Exon p52 appeared to reside on the same restriction fragment as exon p51 on the 3.9kb *EcoR I* fragment. A comparison of the sequences of exon p51 and that of exon p52 reveals no similarity. These results suggest that exon p51 and exon p52 are within the same *EcoR I* DNA fragment in mouse.

2.10.3 Conservation of exon p52 sequence across species

The conservation of exon p52 sequence was examined via hybridisation to a Southern blot of DNAs from several species. As was observed with exon p51, no signal was apparent on the blot after washing, no matter the washing conditions used. It was considered that the size of the exon p52 sequence was too small to be a good probe under the conditions used, hence the lack of signal even with long exposure

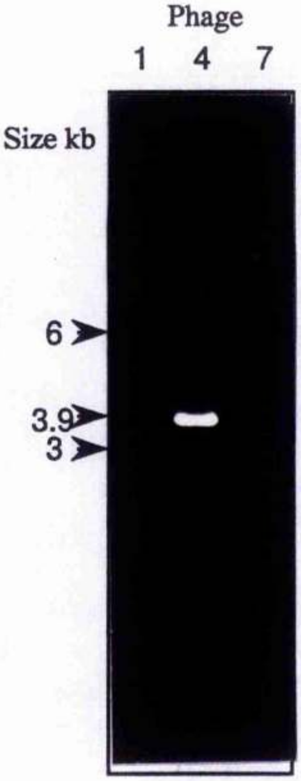


Figure 33: Location of exon p52 in cloned genomic DNA.

Contact print of the autoradiograph of a Southern blot of cloned genomic DNA probed with radio-labelled exon p52. Each lane represents 1μg of each genomic clone digested with *EcoR I* + *Not I*; lane 1, Phage 1; lane 2, Phage 4; and lane 3, Phage 7. Samples were electrophoresed through an 0.8% agarose gel, Southern blotted and probed.

Exon p52 hybridises to the 3.9kb *EcoR I* fragment from Phage 4.

times of this blot. Due to a lack of time not all the conditions of hybridisation could be tested and so nothing can be said about the conservation of exon p52 across species.

Primers were designed to the 5' and 3' ends of the exon p52 sequence for PCR (Materials and Methods). These oligos were end-labelled using polynucleotide kinase, and hybridised to the Zoo blots. This analysis failed to produce any signal on the Zoo blot under the conditions used for hybridisation and subsequent washes. Again due to lack of time not all of the conditions used could be tested and so this analysis does not tell us any thing about the conservation of exon p52 across species.

Using the same oligonucleotide primers described in Figure 32(a) PCR was attempted to show that the exon p52 sequences were present in the mouse genome. The conditions for PCR were tested and when the optimum PCR conditions were established for the primers, the exon p52 sequences were amplified from mouse DNA. Male and female samples of mouse genomic DNA were analysed by PCR. Figure 34 shows the Southern blot of these samples probed with exon p52 sequences. From this blot Exon p52 appears to be present in the mouse genome, the other samples are from human DNA and it appears that the exon p52 primers can amplify sequences from the human DNA. Quite what the sequence of this PCR product from human DNA is, is unknown. It will require the cloning and sequencing of this product.

PCR was used to investigate the conservation of exon p52 sequences across species. If the oligonucleotides designed to exon p52 were sufficiently conserved then it should be possible to use PCR to detect the presence of exon p52 related sequences in the DNA of other species. It was possible that in other species the primer sequences had themselves diverged from the mouse sequence and that the PCR would be unable to amplify any sequence from these DNAs. DNA from mouse, rat, hamster, cow and human were analysed. Samples of the PCR reactions were analysed by electrophoresis, blotted and probed with exon p52. Figure 35 shows that the PCR reaction amplified a product from each species tested that was the same size

Figure 34:

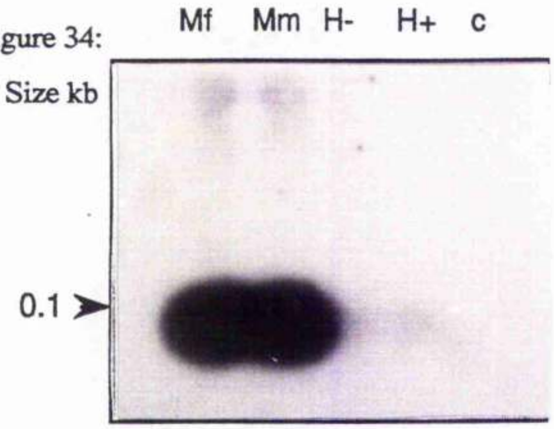


Figure 34: PCR analysis of DNA from mouse and human

PCR was carried out using DNAs as template from; Mf, Mouse female; Mm, Mouse male; H-, unaffected human; H+, human exhibiting sex-linked cleft palate; c,no template control. Reactions were primed using exon p52 forward and reverse primers. 300ng of genomic DNA was used as template and the PCR carried out as described in Materials and Methods.

A fifth of the PCR products were electrophoresed through a 2% agarose gel and blotted onto a Nylon membrane. This was probed with radio-labelled exon p52 sequences.

Note: the presence of product in all the DNA samples indicates the presence of primer sequences in the DNA of mouse and human. The presence of the PCR product in both human samples suggests that exon p52 sequences are present in both affected and unaffected human DNA samples.

Figure 35:

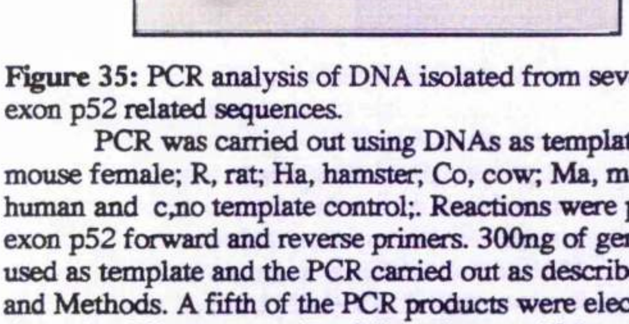


Figure 35: PCR analysis of DNA isolated from several species for exon p52 related sequences.

PCR was carried out using DNAs as template from; Mf, mouse female; R, rat; Ha, hamster; Co, cow; Ma, marmoset; Hu, human and c,no template control;. Reactions were primed using exon p52 forward and reverse primers. 300ng of genomic DNA was used as template and the PCR carried out as described in Materials and Methods. A fifth of the PCR products were electrophoresed through a 2% agarose gel and blotted onto a Nylon membrane. This was probed with radio-labelled exon p52 sequences.

Note: A product is observed in all samples including the no template control, indicating contamination of this sample.

as the product obtained from the mouse sample. At first it appeared that the sequence was present in the DNA of other species. Unfortunately it also amplified a product from the water only control. This contamination should have rendered this analysis meaningless. However, the products amplified, when blotted and probed, exhibit a great difference in their strength of hybridisation to exon p52 sequences. This could mean that the sequences amplified in this PCR analysis are real. However, this PCR analysis will have to be repeated to remove the contaminant from the control.

2.10.4 Expressional analysis of exon p52

To analyse whether exon p52 sequence is expressed at the RNA level during mouse embryonic development, exon p52 was hybridised to a Northern blot of embryonic total RNAs. Figure 29(c) shows that exon p52 hybridises to a sequence around the 28S Ribosomal RNA on this blot. The vector control secondary PCR products produce this signal when hybridised to day 15.5 embryonic total RNA. The fainter signal seen at >10kb found on the Northern blots hybridised with exon p50 and exon p51 is also present. This Northern analysis does not confirm whether exon p52 is part of a transcript or not. The fact that exon p52 does not hybridise to a Southern blot of mouse genomic DNA means that nothing conclusive can be drawn from this Northern analysis. The 500bp and the 800bp transcripts detected by the total exon-trapped products are not observed. It is possible that several exons are present on a single fragment of cloned DNA used in the exon-trapping. These may be spliced together during the processing of the RNA transcript, from the vector and trapped sequences, to form larger products that for whatever reason are resistant to cloning. These larger exon products would be a better template for generating probes and it may be the hybridisation of these larger products that produces the signal seen on the blots probed with total secondary PCR product.

Again lack of signal on the Northerns may have been due to the size of the exon sequences. The oligonucleotide primers designed for PCR were end labelled and used to probe Northern blots of embryonic RNA. This analysis failed to produce

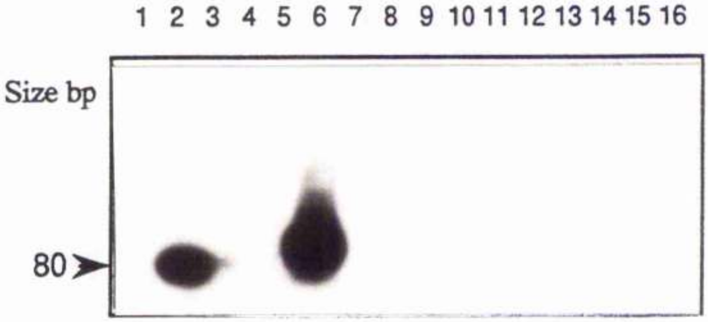


Figure 36: RT-PCR detection of exon p52 expression during embryonic development.

1μg of total RNA from embryonic day 13.5 mice was DNase treated, then reverse transcribed to cDNA primed with oligo [dT]. PCR was then carried out using forward (F) and reverse (R) primers designed to exon p51, p52, p56 and p57. The basic PCR protocol outlined in Materials and Methods was followed.

The combination of primers used were:
Lane 1; Exon p51 (F) + (R)
Lane 2; Exon p52 (F) + (R)
Lane 3; Exon p56 (F) + (R)
Lane 4; Exon p57 (F) + (R)
Lane 5; Exon p52 (R) + Exon p57 (R)
Lane 6; Exon p52 (F) + Exon p57 (R)
Lane 7; Exon p52 (R) + Exon p57 (F)
Lane 8; Exon p52 (F) + Exon p57 (F)

The template used for lane 1-8 was cDNA reverse transcribed from day 13.5 mouse embryonic RNA

Lane 9-lane 16, is a repeat of the PCR reactions described for lane 1-8, except that the template used was DNase treated RNA that had not been reverse transcribed to cDNA,as controls for DNA contamination.

A fifth of the PCR reaction was electrophoresed through a 2% agarose gel, blotted onto Nylon membranes and probed with exon p52.

Note: detection of product in lane 2 demonstrates that exon p52 is expressed at the RNA level at day 13.5 of embryonic development. The detection of product in lane 6 demonstrates the correct orientation of the oligos to give the linked product between exon p52 and exon p57.

any signal on a Northern blot (data not shown). As before a complete analysis could not be attempted and not all of the conditions were tested due to lack of time.

The primers designed to the 5' and the 3' ends of exon p52 were used to analyse expression during embryonic development. Reverse transcribed-PCR (RT-PCR) was used to detect expression of exon p52 during mouse embryonic development (Materials and Methods). RT-PCR is a highly sensitive method of detecting expressed sequences. It is particularly useful for detecting transcripts present at very low levels. The first stage of RT-PCR is cDNA synthesis from the RNA under analysis. The day 13.5 mouse embryonic total RNA sample was treated with DNase to ensure that no DNA sequences would contaminate the PCR reaction, then cDNA was synthesized from this RNA sample. PCR was then carried out on this cDNA template using the primers designed to the 5' and 3' ends of exon p52. The PCR products were then electrophoresed on a 2% agarose gel, blotted and probed with exon p52 sequences. As can be seen from Figure 36, exon p52 sequences were detected in day 13.5 mouse embryonic RNA. The water only control and the RNA only control confirm that the product observed with RT-PCR is a product of expressed sequences.

The detection of exon p52 sequences by RT-PCR demonstrates that exon p52 is expressed and part of a transcript(s) during embryonic development. However, it does not indicate what size the transcript is nor in what tissues it is expressed. In the light of this RT-PCR analysis the lack of detection by Northern analysis of transcripts using exon p52 as a probe is probably due to experimental conditions, which need to be optimised using hybridisation to, and detection on, a Southern blot as a control.

2.11.1 Characterisation of exon p56

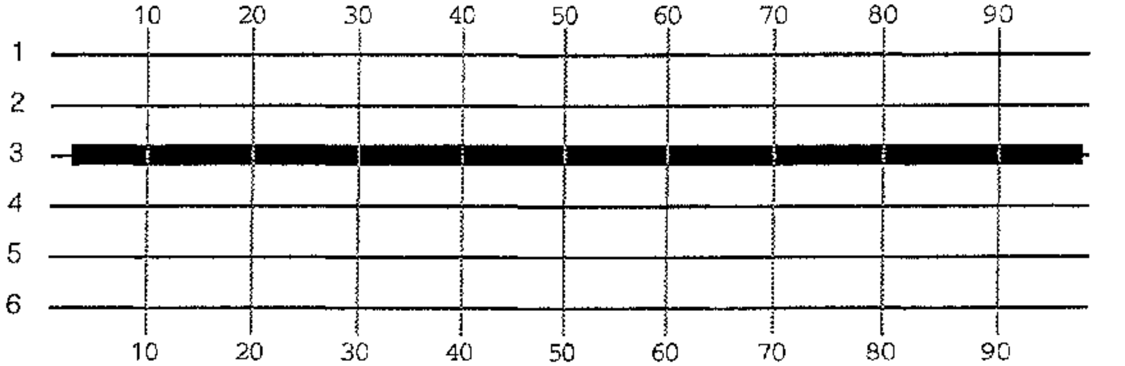
Hybridisation of the Phage 4, 1.2kb *EcoR I* fragment to the secondary PCR products of exon-trapping from Phage 4 identified exon p56. Sequencing of this PCR product showed that it was composed of 99 nucleotides (see Figure 37). Using the

Figure 37 Sequence analysis of Exon p56

(a) Nucleotide Sequence: oligos used to prime PCR are indicated by bold type and underlining.

```
         10         20         30         40         50
TACTCCTAAG ATGGCAGCAA CTCTGGATGG ATGACAAAGC TTCTCTCCTA
ATGAGGATTC TACCGTCGTT GAGACCTACC TACTGTTTCG AAGAGAGGAT
         60         70         80         90        100
TGGTTCAGTG GGTCTTCTCT GTGGCTTCTA GAACCCATCT GTAACCAG
ACCAAGTCAC CCAGAAGAGA CACCGAAGAT CTTGGGTAGA CATTGGTC
```

(b) Open Reading Frames of Exon p56



(c) Amino Acid Sequence of Exon p56 ORF

```
         10         20         30
LLRWQQLWMD DKASLLWFSG SSLWLLEPIC NQ
```

GCG package and the MacVector^{TM} program to analyse the nucleotide sequence no nucleic acid motifs were found. This exon's nucleic acid sequence was compared to the sequences present in the BLASTN database. The search revealed that there were no similar sequences present in the database. This confirms that the sequence of this trapped exon is not derived from the exon-trapping vector and that this exon sequence is unique. The GCG package and the MacVector^{TM} identified a single ORF spanning the complete exon.

The deduced amino acid sequence of the exon p56 ORF was then compared to the sequences present in the BLASTP database. No matches of high identity were found. The matches that were found between the predicted amino acid sequence and the sequences present in the database were either high identity over a very short region, or a low level of identity across the entire sequence. A low level of identity to the G alpha-7 and G alpha-8 subunit genes of *Dictyostelium discoideum* (Wu *et al.,* 1994) was found. Over 28 amino acids 7 were identical (28%) increasing to 42% when similar amino acid substitutions were taken into account. Another protein sequence that exhibited a low level of identity was GATA binding factor-3 of chicken, a general switch factor for cell specific development (Yamamoto *et al.*,1990 and Ko *et al.*, 1991). It exhibited a 32% level of identity over 28 amino acid residues. The level of similarity increased to 42% when similar amino acids substitutions are included. A tyrosine 10 receptor kinase from mouse was found to have a 35% level of identity, increasing to 50% similarity across 18 amino acid residues (accession number S42796 EMBL databank). Another match from this search was PHSE protein from *E.coli*, which is a penicillin binding protein. Although it exhibited a low level of identity at 28% across 24 amino acid residues, the level of similarity was 54% (Accession number P33013 EMBL databank).

These matches from the database reveal that the deduced amino acid sequence of exon p56 does not share a high level of identity to any amino acid sequences present in the databases. This suggests that the sequence of exon p56 is unique and does not reveal a high level of relatedness to other proteins in the database. This protein database search supports the results of the nucleotide sequence database search.

2.11.2 Mapping of exon p56 onto mouse genomic DNA

To ensure that exon p56 had been exon-trapped from Phage 4, exon p56 was hybridised back to the phage clone it was derived from. Figure 38 shows a Southern blot of Phage 4 DNA digested with *EcoR I* probed with exon p56. It can be clearly seen that exon p56 is indeed derived from the Phage 4 1.2kb *EcoR I* fragment. This confirms the isolation of exon p56 by hybridisation of the Phage 4 1.2kb *EcoR I* fragment to subcloned secondary PCR products.

2.11.3 Conservation of exon p56 across species

Exon p56 was hybridised to a Zoo blot of DNAs from various species. The blot included DNA taken from mouse, rat, hamster, cow and human. The exon p56 sequences were labelled using the random priming method. As with the other exon sequences no signal was observed when the blot was washed. This was true no matter the washing stringency. The probes used had been of high specific activity and had already been shown to hybridise to mouse DNA sequences present in recombinant phage. As with the previous exon probes derived from the exon-trapping vector it was believed the probe was too small to detect sequences in genomic DNA under the conditions used.

Oligonucleotide primers were designed to the 5' and 3' ends of exon p56. These primers were used for two experiments. Firstly they were end labelled and used to try and produce probes of high specific activity. They were also used for PCR reactions to detect the presence of exon p56 sequences in DNA samples and also RNA samples. Zoo blots hybridised with oligonucleotides from exon p56 exhibited no signal on short exposures. On longer exposures (in excess of 7 days), all that could be observed was a high level of background. There are many reasons as to why the end labelled oligonucleotides failed to hybridise. The lack of specific signal is

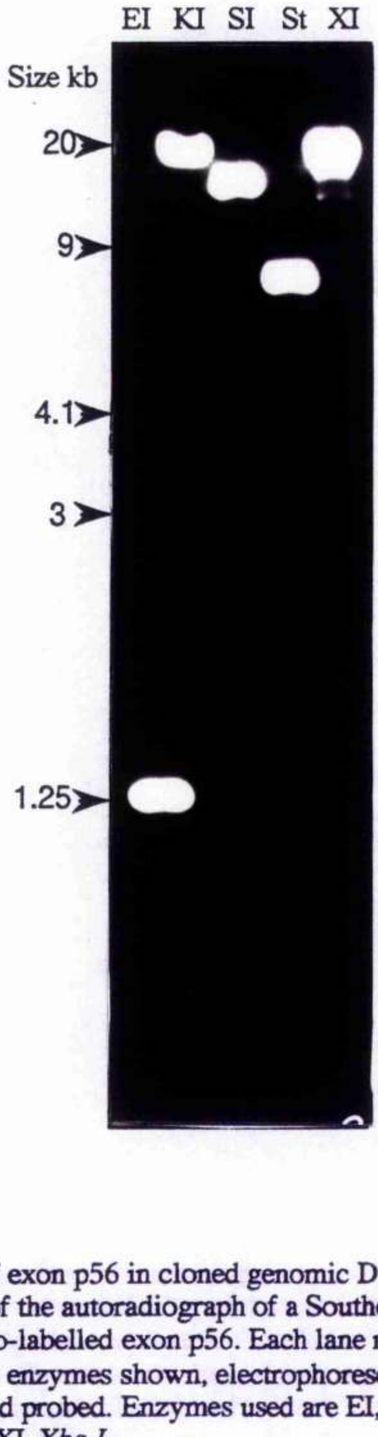


Figure 38: Location of exon p56 in cloned genomic DNA

Contact print of the autoradiograph of a Southern blot of cloned genomic DNA probed with radio-labelled exon p56. Each lane represents 1μg of Phage 4 DNA digested with the enzymes shown, electrophoresed through an 0.7% agarose gel, blotted and probed. Enzymes used are EI, *EcoR I + Not I*; KI, *Kpn I*; SI, *Sal I*; St, *Sst I* and XI, *Xba I*.

Note:exon p56 hybridises to the 1.25kb *EcoR I + Not I* fragment.

probably a result of the optimal hybridisation conditions not being used. Due to the lack of time the optimal conditions were not defined.

PCR was attempted to confirm that exon p56 was present in mouse genomic DNA samples. Using the oligonucleotide primers designed to the 5' and 3' ends of exon p56. Conditions for PCR were tested varying magnesium concentration and primer concentration, (see Figure 39(a)), when these PCR conditions had been optimised using plasmid DNA as template, PCR was conducted using mouse genomic DNA. From Figure 39(a) exon p56 is present in mouse genomic DNA, this confirms that the analysis of Zoo blots using end labelled oligonucleotides requires different conditions to those used. Due to the results of expression analysis (see below), the conservation of exon p56 across species was not addressed using PCR.

2.11.4 Expressional analysis of exon p56

To analyse whether exon p56 was expressed during mouse embryonic development, exon p56 was hybridised to a Northern blot of embryonic mouse total RNAs. No transcript of any nature was revealed by this analysis. It was also possible that the failure of hybridisation resulted from the small size of the exon used to probe the filter and the hybridisation conditions used, as experienced with other exon hybridisations. The oligonucleotide primers, used to show the presence of exon p56 in mouse genomic DNA above, were used for oligonucleotide hybridisation to the Northerns. The hybridisation of end labelled oligonucleotide primers to Northern blots of embryonic RNA also proved negative.

RT-PCR was then attempted to address expression of exon p56 sequence in day 13.5 embryonic total RNA. The RT-PCR reaction samples were electrophoresed on agarose gels. No product could be detected by ethidium bromide, therefore the gels were electroblotted and probed with exon p56 sequence, which also was negative (data not shown). The 500bp control product was clearly visible, this means that the cDNA synthesis step of the reaction was successful. It is possible that the

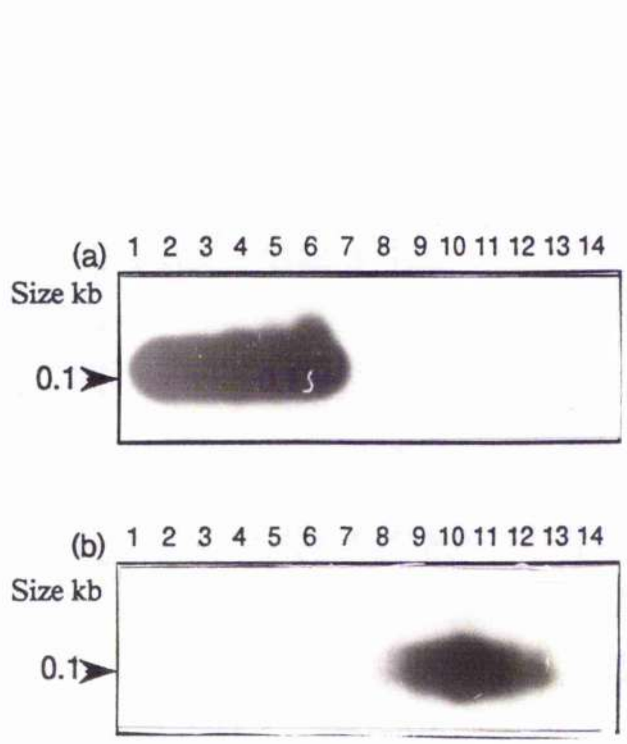


Figure 39: PCR detection of (a) exon p56 and (b) exon p57 in mouse genomic DNA, varying the Mg^{2+} concentration.

PCR was carried out using primers to exon p56, lane 1-6 and 13, exon p56 forward and reverse primers;and exon p57, lane 7-12 and 14, exon p57 forward and reverse primers. Lane 13 and 14 are no template controls.

$MgCl_2$ concentration was varied between reactions, lane 1 and 7, 0mM; lane 2 and 8, 0.5mM; lane 3 and 9, 1.0mM; lane 4 and 10, 1.5 mM; lane 5 and 11, 2.0mM; lane 6 and 12, 2.5mM. The no template control reactions were carried out at 1.5mM $MgCl_2$.

A fifth of each PCR reaction was electrophoresed through a 2% agarose gel, Southern blotted and probed with (a) exon p56 and (b) exon p57.

Note: the primers to exon p56 are capable of priming PCR at all of the Mg^{2+} concentrations used, except in the absence of Mg^{2+}. The primers to exon p57 are optimal at 1.5mM Mg^{2+}, lane 10.

secondary structure of RNA containing exon p56 sequence restricts the synthesis of cDNA or alternatively that it is simply not expressed at this stage of development. Exon p56 may still be expressed at another stage of embryonic development or even postnataly. Thus it can not be deduced from these results that exon p56 is not part of an authentic transcript.

The results of analysis of exon p56 only reveal that it is present in mouse DNA. Due to time limitations the question of exon p56 expression and the temporal and spatial aspects of that expression will have to be addressed at a later date.

2.12.1 Characterisation of exon p57

The hybridisation of Phage 1 and Phage 7 sequences to the secondary PCR products produced by exon-trapping, resulted in the identification of exon p57 from the exons-trapped from Phage 7. Both strands were sequenced and exon p57 was found to be 101 base pairs long (Figure 40). Nucleic acid sequence analysis using MacVector^{TM} and the GCG package revealed no sequence motifs. A search of the sequence to identify ORF revealed several, one of which spanned the entire sequence. The putative amino acid sequence of the ORF was determined using the GCG package. Analysis of this amino acid sequence using the motif option of the GCG package failed to reveal any motifs.

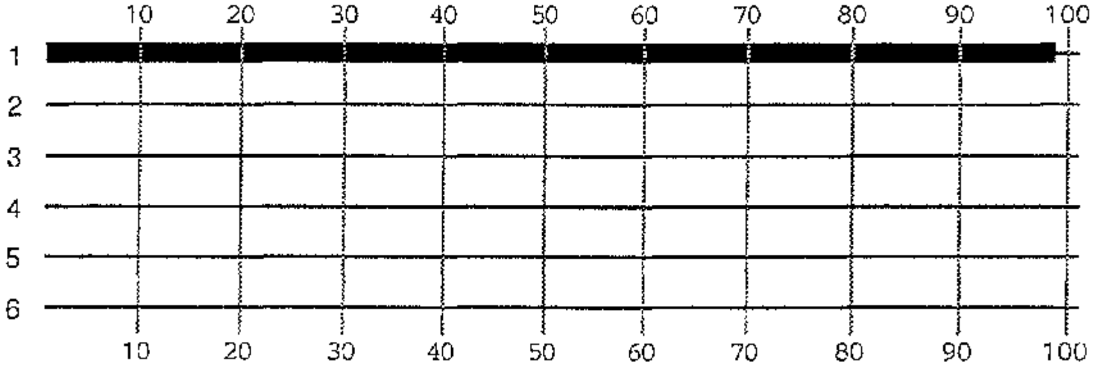
The amino acid sequence was then compared to amino acid sequences present in the BLASTP database. No high levels of similarity at the amino acid level were found for long stretches of the sequence of exon p57s' translated ORF. Small regions of the amino acid sequence revealed high levels of similarity. One peptide sequence showed a high degree of similarity to amino acid sequence from a diverse range of genes. The identical peptide sequence, FSEDPTS, occurred in a range of sequences from the mouse homeobox protein HOX-D4 (*Hox-4.2*), to the yeast AFG1 protein, which belongs to the CDC48/PAS1/SEC18 family. Several other amino acid sequences in the database exhibited a lower level of similarity to the entire exon p57 amino acid sequence. The *E.coli* nucleoside specific channel-forming protein TSX

Figure 40: Sequence analysis of Exon p57

(a) Nucleotide Sequence of Exon p57: oligos used to prime PCR are indicated by bold type and underlining.

```
        10         20         30         40         50
GGTGCAGTAC TAGCTGCTGT GTCAAGTCAC AAATTCAATT CCTTCTATGG
CCACGTCATG ATCGACGACA CAGTTCAGTG TTTAAGTTAA GGAAGATACC
        60         70         80         90        100
GGACCCTCCT GAAGAGTTGC CAGATTTCTC CGAAGACCCT ACCTCCTCAG
CCTGGGAGGA CTTCTCAACG GTCTAAAGAG GCTTCTGGGA TGGAGGAGTC
       110
G
C
```

(b) Open Reading Frames of Exon p57



(c) Amino Acid Sequence of Exon p57 ORF

```
        10         20         30
GAVLAAVSSH KFNSFYGDPP EELPDFSEDP TSS
```

(Bremer *et al.,* 1990) exhibited a 46% level of identity at the amino acid level (over the entire 33 amino acid residues of translated sequence). When similar amino acid substitutions were included the level of similarity increased to 57%. A human transcription factor GATA-2 (Dorfman *et al.,* 1992) showed a lower level of identity (28%) across the 33 amino acid residues. The degree of similarity increased to 53% when like amino acids substitions were included in the analysis. The final amino acid sequence from this comparison was the cucumber mosaic virus (CMV) 1A protein (Hayakawa *et al.*, 1989). It displayed a 36% identity to the deduced amino acid sequence of exon p57. It also displayed a 51% level of similarity when similar amino acid substitions were included.

The database searches carried out do not reveal much information about the nature of exon p57. The database searches reveal that the nucleotide sequence of exon p57 has no similar sequences present in the database. It also proves that the sequence of exon p57 is not derived from vector sequences. Although a search of the BLASTP database with the translated ORF revealed several sequences with a low level of identity, the levels of similarity are quite low.

2.12.2 Mapping of exon p57 onto mouse genomic DNA

Exon p57 was mapped onto mouse genomic DNA. Exon p57 sequences were random prime radio-labelled, then hybridised to a Southern blot of Phage DNA digested with *EcoR I* (Figure 41). Figure 41 confirms that exon p57 is derived from the 1.9kb *EcoR I* fragment that is common to all phage. Exon p57 clones were isolated from the pools of exon-trapped products from all three phage. The results of section 2.6.4, do indicate that some secondary PCR products are shared between the exon-trapped products from the different phage as seen in Figure 21. Exon p57 is one of these exon-trapped products. The subcloned 1.9kb *EcoR I* fragment was restriction digested, electrophoresed through an agarose gel, and electroblotted. Exon p57 was hybridised to this blot. Exon p57 is mapped to a 0.2kb *EcoR V-BamH I* restriction fragment present in the 1.9kb *EcoR I* fragment (data not shown).

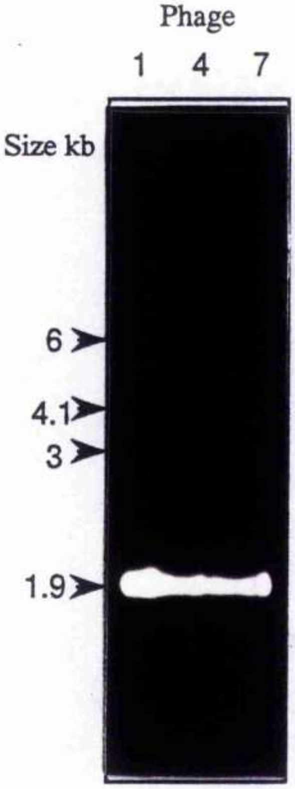


Figure 41: Location of exon p57 in cloned genomic DNA.

Contact print of the autoradiograph of a Southern blot of cloned genomic DNA probed with radio-labelled exon p57. Each lane represents 1µg of each genomic clone digested with *EcoR I* + *Not I*; lane 1, Phage 1; lane 2, Phage 4; lane 3, Phage 7. Samples were electrophoresed through an 0.8% agarose gel, Southern blotted and probed.

Note: that exon p57 hybridises to the 1.9kb *EcoR I* fragment that is present in all the Phage clones.

2.12.3 Conservation of exon p57 across species

To analyse the evolutionary conservation of exon p57 sequences, exon p57 was hybridised to a Zoo blot of DNAs from a variety of species. As with the previous exon hybridizations to Zoo blots, no signal was observed with any of the washing conditions used. This lack of hybridisation detected with the random labelling of the exon sequences is probably due to the small size of the exon sequences. Oligonucleotide primers were designed to the 5' and the 3' ends of the exon p57 sequence. These primers were used in end labelling reactions and for PCR reactions. The end labelling of the oligonucleotides was an attempt to produce probes of high specific activity. As described earlier, it was hoped that these labelled oligonucleotides would overcome the problems found with random labelling.

The oligonucleotide primers to exon p57 were used in PCR reactions to investigate the presence of exon p57 in DNA samples. The conditions for PCR were defined using mouse DNA as template, see Figure 39(b). These results demonstrate that exon p57 sequences are present in the mouse genome. They were also used to define the PCR conditions required for subsequent analysis.

PCR reactions were used to investigate the conservation of exon p57 sequences in different species. If the sequence of the oligonucleotides is not sufficiently conserved then we might expect that the PCR will not work. The PCR products were electrophoresed, blotted and probed with exon p57. This revealed the presence of products from the different species the same size as exon p57 (Figure 42). As with the PCR on the sequences from exon p52, there appears to be contamination that can be clearly seen in the water only (no template) control. This means that no meaningful analysis can be made of these results. The blot shows that the more divergent the species from mouse the greater the reduction in ability of the PCR product to hybridise with the exon p57 probe. This suggests that the sequences though conserved are sufficiently diverged to produce a difference in ability to hybridise to the original mouse sequence. The water control exhibits an intensity of

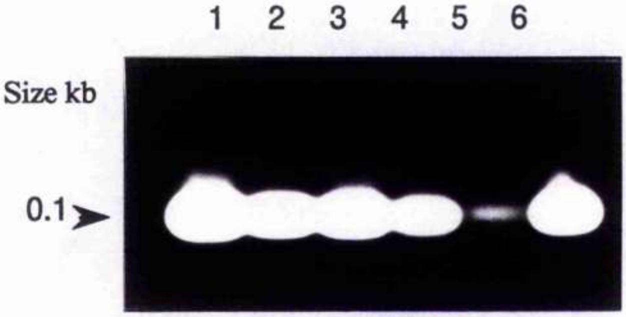


Figure 42: PCR analysis of DNAs isolated from several species for exon p57 related sequences.

PCR was carried out using DNAs as template from, lane 1, male mouse; lane 2, rat; lane 3, hamster; lane 4, cow; lane 5, human and lane 6, no template control. Reactions were primed using exon p57 forward and reverse primers. 300ng of genomic DNA was used as template and the PCR carried out as given in Materials and Methods.

A fifth of the PCR products were electrophoresed through a 2% agarose gel and blotted onto a Nylon membrane. This was probed with radio-labelled exon p57 sequences.

hybridisation equivalent to that of the mouse DNA control. This suggests that the contamination observed is restricted to the water only control and that the PCRs in the other DNA samples are real and not artifacts. This can be only proven by repeating the PCR.

2.12.4 Expressional analysis of exon p57

To determine whether exon p57 is part of a transcript, exon p57 was hybridised to a Northern blot of mouse embryonic total RNA. Absolutely no signal was detected on any Northern blot of embryonic RNAs derived from different stages of devlopment. Reprobing of these blots with the total secondary PCR product revealed the 800 nucleotide and 500 nucleotide transcripts, which showed that the RNAs present on the blot were intact. As with the other exons used as probes the conditions of hybridisation were not optimised for exon p57. As the Southern blot analysis failed to show the presence of exon p57 in genomic DNA but showed that it was present in the cloned genomic DNA then the lack of hybridisation is probably a result of the conditions used.

The lack of hybridisation with exon p57 may also be a result of the transcript that exon p57 is part of being present in RNA at a very low level, or present in an RNA that was not tested. RT-PCR, as mentioned earlier, is a highly sensitive method for detecting the presence of low level transcripts. The primers that were used to detect the presence of exon p57 in DNA, were used in RT-PCR analysis of embryonic RNA. RNA from day 13.5 embryonic mice was reverse transcribed to cDNA. This cDNA was then used as a template for PCR. The products were electrophoresed through a 2% agarose gel, blotted and probed with exon p57. Figure 43 shows that exon p57 is transcribed at day 13.5 of mouse embryonic development. RT-PCR analysis cannot reveal the size of the transcript that exon p57 belongs to. RT-PCR on total RNA from the body of embryonic mice also does not reveal whether the transcript is ubiquitous or confined to one cell type or organ.

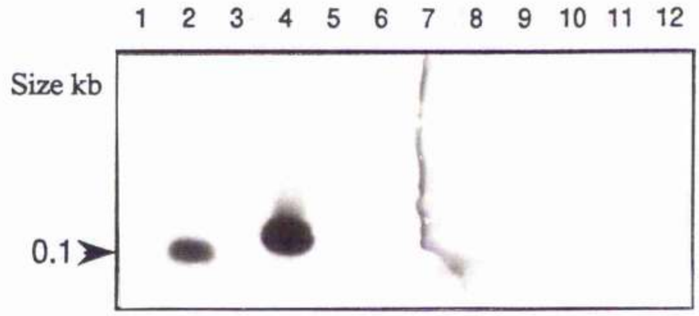


Figure 43: RT-PCR detection of exon p57 expression during embryonic development.

1μg of total RNA from embryonic day 13.5 mice was DNase treated, then reverse transcribed to cDNA primed with oligo [dT]. PCR was then carried out using forward (F) and reverse (R) primers designed to exon p51, p52, p56 and p57. The basic PCR protocol outlined in Materials and Methods was followed.

The combination of primers used were:

Lane 1; Exon p52 (F) + (R)
Lane 2; Exon p57 (F) + (R)
Lane 3; Exon p52 (R) + Exon p57 (R)
Lane 4; Exon p52 (F) + Exon p57 (R)
Lane 5; Exon p52 (R) + Exon p57 (F)
Lane 6; Exon p52 (F) + Exon p57 (F)
Lane 7-lane 12, RNA controls for the above.

The template used for lane 1-6 was cDNA reverse transcribed from day 13.5 mouse embryonic RNA.

Lane 7-lane 12, is a repeat of the PCR reactions described for lane 1-6, except that the template used was DNase treated RNA that had not been reverse transcribed to cDNA,as controls for DNA contamination.

A fifth of the PCR reaction was electrophoresed through a 2% agarose gel, blotted and probed with exon p57.

Note: exon p57 hybridises to the RT-PCR product in lane 2 indicating that exon p57 is expressed at day 13.5 of embryonic development. Exon p57 hybridises to the RT-PCR product in lane 4 confirming that exon p52 and exon p57 are linked together.

Characterisation of the expression of exon p57 fully using RT-PCR was not attempted due to lack of time.

As two exons have been shown to be expressed using RT-PCR it is probable that the hybridisation conditions used for Northern blot analysis, using the exon sequences as probes, are not optimal and this results in the lack of hybridisation observed with the exon probes.

2.13.1 Orientation of the trapped exons to each other

The method used to subclone the secondary PCR products is non-directional. This results in being unable to either orientate the exons relative to each other or orientate them on the DNA from which they are derived. If the exons are part of the same transcript then the cloning of the cDNA of that transcript would reveal their relative orientation to each other. Another approach is to use PCR to amplify between the individual exons from the genomic DNA or the cloned recombinant phage DNA. As the exons identified could not be shown to be expressed by Northern analysis, no attempt was made to screen a cDNA library prepared from mouse embryonic RNA. Instead an attempt was made to use PCR to amplify between the individual exons. There were some problems with this approach. Firstly, one (p51) of the exons was resistant to PCR analysis. The reason for this remains unclear, so PCR analysis using exon p51 was not attempted. Secondly, exon p50 is composed of repetitive sequence, therefore PCR using exon p50 primers within the repeat would produce uninterpretable results. Thirdly, a potential problem with the PCR approach was the relatively large distances that separated the exons, exon p50 and exon p52 are separated by about 10kb DNA. Nevertheless, this approach was attempted for the exons which were deemed least problematic.

Exon p56 and exon p57 are separated by at least 8kb of genomic DNA. Contained within this 8kb of DNA is some very highly repetitive sequence, as shown previously by Southern analysis. Individual primers from one exon were coupled with a single primer from the other exon, in each PCR reaction. This resulted in four possible combinations of primers. The PCR products were electrophoresed through a 2% agarose gel, blotted and probed with exon p56 and exon p57 sequences. The results of the PCR are inconclusive (data not shown). Although several products hybridise to the separate exon sequences, only a single band hybridises with sequences from both exons. This single band though is only 1kb in size and is not compatible with the restriction map produced for the 30kb of cloned DNA. There is the possibility that repeat sequences were deleted out early in the PCR resulting in the 8kb reduced to 1kb PCR product, this may therefore suggest the orientation of exon p56 to exon p57.

2.13.2 Analysis of expression by RT-PCR across different exons

RT-PCR is a very sensitive method for detecting the presence of transcripts. In the case of the exons that have been studied here, RT-PCR has been used to determine whether those sequences are expressed individually. The analysis so far has shown that at least three of the exon sequences are expressed, exon p50, exon p52 and exon p57. Because the RT-PCR was carried out on individual exons it could not be determined whether the exons were part of one transcript or separate transcripts

RT-PCR was carried out between individual exons to determine whether the separate exons were part of a single transcript. If the exons did belong to a single transcript this RT-PCR would reveal if any other exons are intervening in the transcript. The same primers were used for this inter-exon RT-PCR as had been used to analyse the expression of the individual exons. The primers that were available for exon p50 were thought be unsuitable for RT-PCR analysis of exon p50, as the entire sequence of exon p50 is highly repetitive. There was a high probability that the primers to exon p50 would anneal to other LINE-1 sequences in the RNA. Attempted RT-PCR using primers to exon p50 confirmed the redundancy of the primers, so exon p50 primers were not used for inter-exon RT-PCR.

RT-PCR was attempted between exon p52 and exon p57 (Figure 36 and Figure 43). The cDNA templates used were day 13.5, 15.5 and 16.5 mouse embryonic total RNA. Figure 44 shows the blot of the RT-PCR reactions probed with exon p57. A product of about 177bp was observed to hybridize with exon p57. The blot was reprobed with exon p52 and a similar pattern of hybridisation was observed. The 177bp product was amplified from all three RNA samples, for different embryonic time points, tested. The controls for cDNA synthesis appear to have worked, as visualised by ethidium bromide staining (data not shown). The product is also the same size as the sum of the two exons added together. To confirm that, the RT-PCR product was authentic, it was subcloned and sequenced (Figure 45).

Analysis of the sequence of the RT-PCR product revealed that the observed 177bp product was indeed derived from the two exons (see Figures 32 & 40). A database search of BLASTN, unsurprisingly, identified the same sequences that the individual exon sequences identified. When this nucleic acid sequence was analysed to identify any ORFs it was found that several were present. Only one ORF was complete across the entire sequence of both exons. So the ORF of exon p57 was maintained. One of the two ORFs (ORF3) present in exon p52 reaches a stop codon in exon p57 and was not identified in this search, while the other (ORF2) is open with the exon p57 ORF. This is partly due to slightly different parameters in the search. The ORF analysis was weighted towards longer ORFs to identify one which covered the entire nucleic acid sequence. The sequence of the larger ORF that extends across both exons shows a level of similarity to the sequence of CaM kinase II at the nucleotide level observed with exon p52. The region of similarity extended into exon p57 from exon p52 giving new values for the extent of homology. The orientation of the ORF is 5' in exon p52 to 3' in exon p57.

Over a 30 amino acid sequence (6 of which are derived from exon p57 sequence) 20 amino acid residues (66%) were found to be identical to residues in the sequence of CaM kinase II. This was found to be true no matter the species the CaM kinase II was derived from. When similar amino acid substitutions were taken into

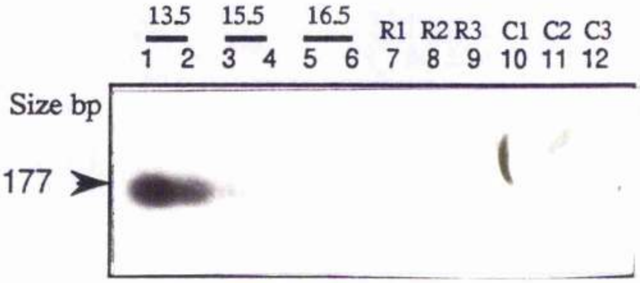


Figure 44: Analysis of expression by RT-PCR of exon p52-exon p57 sequences during development.

1µg each of total RNA from embryonic day 13.5 , 15.5 and 16.5 mice were DNase treated, then reverse transcribed to cDNA primed with either oligo [dT], lanes 1,3 and 5 or random primers, lanes 2,4 and 6. PCR was then carried out using the combination of primers found to produce the exon p52-exon p57 RT-PCR product in Figure 33, lane 5. The basic PCR protocol outlined in Materials and Methods was followed.

RT-PCR controls are; R1, day 13.5 RNA only; R2, day15.5 RNA only and R3, day 16.5 RNA only. C1, C2 and C3 are cDNA synthesis controls to ensure that cDNA was synthesised as template for the RT-PCR. RT-PCR products were electrophoresed through a 2% agarose gel, blotted onto Nylon membrane and hybridised with exon p57.

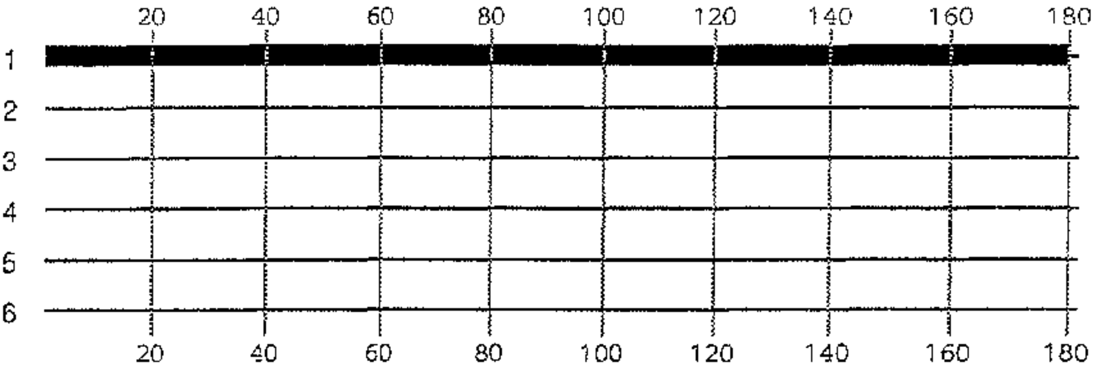
Note: exon p52-exon p57 RT-PCR product is detectable in day 13.5 cDNA, day 15.5 cDNA and on longer exposures in day 16.5 cDNA. The variation in intensity of hybridisation reflects the efficiency of the PCR reaction.

Figure 45: Sequence Analysis of the Exon p52-p57 RT-PCR Product

(a) Nucleotide Sequence: oligos used to prime RT-PCR are indicated by bold type and underlining.

```
         10         20         30         40         50
GGGGATGGTT ATGCCTACAA AATCCATCTT CCAGAAACAG TAGAACAACT
CCCCTACCAA TACGGATGTT TTAGGTAGAA GGTCTTTGTC ATCTTGTTGA
         60         70         80         90        100
GAGGAAATTC AATGCAAGGA GAAAACTAAA GGGTGCAGTA CTAGCTGCTG
CTCCTTTAAG TTACGTTCCT CTTTTGATTT CCCACGTCAT GATCGACGAC
        110        120        130        140        150
TGTCAAGTCA CAAATTCAAT TCCTTCTATG GGGACCCTCC TGAAGAGTTG
ACAGTTCAGT GTTTAAGTTA AGGAAGATAC CCCTGGGAGG ACTTCTCAAC
        160        170        180
CCAGATTTCT CCGAAGACCC TACCTCC
GGTCTAAAGA GGCTTCTGGG ATGGAGG
```

(b) Open Reading Frames of the Exon p52-p57 RT-PCR product



(c)Amino Acid Sequence Of Exon p52-p57 ORF

```
         10         20         30         40
GDGYAYKIHL PETVEQLRKF NARRKLKGA VLAAVSSHKF NSFYGDPPE
         60
PDFSEDPT
```

account the level of similarity rose to 76% (23 out of 30 residues). This analysis of ORFs and deduced amino acid sequence suggests that the transcript which contains exon p52 and exon p57 shares a protein motif when translated with CaM kinase II.

The result of the RT-PCR suggests that in this transcript exon p52 and exon p57 are spliced together and no other exons intervene between exons p52 and p57. Although other exons may lie between exon p52 and exon p57 at the genomic level, that were not detected by the exon-trapping analysis. These other exons may be expressed as part of other transcripts that do not include exon p52 and exon p57 at different stages of development, or they may include exon p52 and exon p57 but be differentially spliced. To understand the transcripts expressed at this locus will require further analysis.

2.14 Disruption of RT-PCR product in transgenically positive mice

From the restriction map of this cellular locus, exon p52 and exon p57 are found to reside on separate *EcoR I* fragments that flank the transgene site of insertion. Evidence suggested that a cellular locus was affected by transgene insertion in the mutant line that exhibited sex-linked cleft palate. The nature of the cellular disruption is unclear. It is not known whether there is a physical disruption of the cellular locus, or if the pattern of expression is perturbed by the presence of the transgene. The total secondary PCR products were hybridised to a Northern blot of RNAs derived from transgenically positive and transgenically negative mice (see Figure 24). In this analysis no physical disruption of the transcript was observed. RT-PCR was used to test whether the transcript(s) that exon p52 and exon p57 belong to was affected in this transgenic line.

RT-PCR, using the primers that produced the 177bp fragment containing exon p52 and exon p57, was repeated on transgenically positive and transgenically negative RNA. Figure 46 shows the result of this RT-PCR. The transgenically negative spleen sample appears to contain the expected 177bp PCR product. It also contains a slightly larger PCR product at about 250bp in size. It is difficult to decide at this stage whether this latter PCR product is from an alternatively spliced transcript or an artifact. The RT-PCR products from the transgenically positive sample exhibit a clear difference from those of the negative sample (Figure 46(a)). The 177bp product is again the predominant product and the slightly larger product at 250bp is also detectable. However two other PCR products are also present. One of these is again slightly larger than the 250bp PCR product at about 350bp. The other product detectable is a band of about 1kb.

These two RT-PCR products are clearly not detectable in the transgenically negative sample. The transgenic construct contained the EBV LMP-1 gene, and it has been shown to be expressed in transgenically positive mice. It is a multiply spliced transcript. It is possible that part of the LMP-1 transcript has become integrated into that of the exon p52-exon p57 transcript. To test this the blot of transgenically positive and transgenically negative RT-PCR products was probed with LMP-1 sequences.

Probing of the RT-PCR products from transgenically positive and transgenically negative RNA with LMP-1 (transgene) identified a 250bp product, a 350bp product (shown to hybridise with exon p52 and exon p57) and a product that is about 1.5kb in size (Figure 47(b)). No products were found to hybridise to transgenic sequences in the transgenically negative sample. The 1.5kb product that hybridises to the LMP-1 sequences fails to hybridise to either exon p52 or exon p57, whereas the 1.0kb fragment that hybridises to exon sequences fails to hybridise with the LMP-1 probe. This result indicates that exon p52 and exon p57 splice to LMP-1 sequences in the transgenically positive sample. The exact sequence of the hybrid RT-PCR products is unknown at this stage. This will require the cloning and sequencing of these RT-PCR products to determine which LMP-1 sequences are present in the hybrid RT-PCR products.

What is clear from this result, is that the splicing of the exon p52-exon p57 product is affected in transgenically positive spleen RNA, compared to the transgenically negative sample. This disruption is probably as a result of the presence

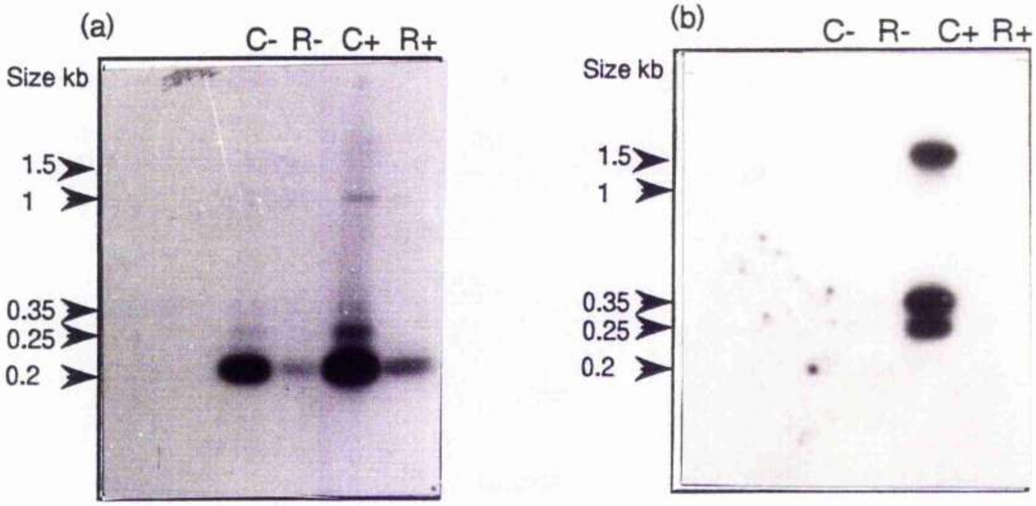


Figure 46 (a): Disruption of the exon p52-exon p57 RT-PCR product in transgenically positive spleen; **(b)**: Identification of cellular-transgene hybrid transcripts in transgenically positive spleen.

1µg of total RNA prepared from the spleens of transgenically positive (+) and negative (-) mice were separately DNase treated, then reverse transcribed to cDNA primed with oligo [dT]. PCR was then carried out using the primers that produced the 177 bp RT-PCR product spanning exon p52 and exon p57 in embryonic RNA. The basic PCR protocol outlined in Materials and Methods was followed.

'C' denotes cDNA and 'R' denotes RNA control. PCR products were electrophoresed on a 2% agarose gel, blotted and hybridised with (a) exon p57 and (b) the LMP-1 transgene construct.

Note: that four products are identified by hybridisation with exon p57 in the transgenically positive sample at 177bp, just under 250bp, about 350bp and just under 1kb. Compared to the 177bp and about 250bp in the transgenically negative sample.

Probing the same blot with the LMP-1 transgene construct identified three RT-PCR products only in the transgenically positive cDNA sample the 250bp product, the 350bp product and a product >1kb.

of the transgene and transgene sequences are spliced to the exons present at the wild type locus.

2.14.1. Critique II

As the analysis of genomic DNA described in sections 2.1 to 2.5 revealed a high degree of repetitive sequence the approach of exon-trapping was used to isolate gene sequences at the locus. All genomic DNAs and total RNAs used in Southern and Northern analyses were confirmed as being intact by ethidium bromide staining and UV visualisation. Probe fragments used were also believed to be intact, as argued in critique I (section 2.5.1) .

In the case of Northern blots presented in the analysis of exon-trapping products not every blot was controlled for sample loading. Where blots did not have loading controls no quantitative statements can be made and as such, estimations of relative quantities have been presented only for those that were normalised. The patterns of hybridisation observed with the pooled exon-trapped products (Figures 22 and 23(a) & (b) were compared to the pattern of hybridisation observed when the same or similar Northern blots were probed with exon-trapped products from the COS-7 cells transfected with only the vector DNA. This control was performed to differentiate between any signal observed as being the result of hybridisation of incorporated exon sequences or vector derived sequences. The intact exon-trapping vector itself was not used as a probe control on a Northern blot but the protocols followed in the exon-trapping would not incorporate vector specific products that were not included in the control mentioned above.

All of the exons identified were used as probes on Southern blots of cloned genomic DNA to demonstrate that they were indeed derived from the cloned cellular DNA derived from the locus (Figure 21). However, although these probes hybridised to blots of cloned DNAs the specific activity of these exons as probes was not enough to reveal hybridisation signals to genomic DNA on Southern blots and total RNA on Northern blots. For this reason the pattern of expression at the level of RNA could not be addressed for the individual exons by Northern analysis.

In view of the technical problem of low specific activity of these very small radio-labelled exons, the alternative approach of Reverse Transcribed-PCR (RT-PCR) was performed to investigate whether or not the exons were expressed. RT-PCR is not quantitative and therefore in the text no reference is made from these experiments to the levels of expression. However, what is important is that RT-PCR was used to demonstrate that several exons are expressed at the level of RNA during embryonic development at the time points investigated. Not only were two exons (exons p52 and p57) shown to be expressed but they were also shown to be spliced together (Figure 44). Furthermore, this splicing is disrupted in the transgenically positive females (Figure 46). Figure 46 suffers from inadequate controlling as no hybridisation signal should have been observed in the 'R' tracks as these are RNA only controls. The presence of hybridisation suggests either (i), track overspill or (ii), exon contamination of samples. However, it cannot be genomic DNA contamination of RNA since in the genome the products (p52 & p57) are separated by several kilobases of DNA. Therefore, although the signal is low in the 'R' tracks, this experiment would need to be repeated to confirm the tentative conclusions regarding exon p52-exon p57 expression.

Exon-trapping was used to isolate several very small exons from the cloned DNA. The trapped exons were derived from those fragments that were shown to contain highly repetitive sequences. One of the exons identified was found to consist of purely repetitive sequence of the LINE-1 class (see section 2.8.1). It is not surprising, in view of the repetitive nature of the DNA present at this locus and the small size of the identified exons, that the Zoo blot (section 2.4) analysis did not reveal any conserved transcribed sequences.

Unfortunately, a full length cDNA was not cloned in the course of this work. However, several different approaches to clone a cDNA molecule were attempted. These attempts were hampered by the inability to use the exons as probes due to their small size and consequent low specific activity as probes. For this reason exon sequences were not useful for screening cDNA libraries.

RT-PCR had been used to demonstrate that several exons are expressed at the level of RNA during mouse embryonic development (Figure 44). Therefore RT-PCR afforded an alternative approach to the screening of cDNA libraries, namely Rapid Amplification of cDNA Ends (RACE). This was attempted to clone the 5' end of the transcript but proved unsuccessful. Unfortunately, RT-PCR only reveals whether or not a sequence is expressed and does not permit determination of the size of the transcript. Therefore it is quite possible that the 5' end of the transcript was at such a distance as to prevent efficient copying into DNA (reverse transcription) or subsequent amplification (PCR).

Subsequent to the analysis presented here, 3' RACE has been used to identify 1.5kb and a 0.8kb 3' products from mouse embryonic day 13.5 total RNA. These products, when used as probes on Northern blots of embryonic RNA, hybridise to transcripts at approximate sizes of 10kb and 5kb. These transcripts exhibit a pattern of expression completely different from each other within mouse embryonic development. These 3' RACE results indicate that the PCR extension times used for the 5' RACE in this study were probably insufficient to allow for the distance between the priming site and the 5' end of the transcripts.

Although the work presented in this thesis did not produce a complete cDNA (which subsequently has been isolated and is being characterised) the original aims of the Ph.D. project were achieved:

(i) Initial characterisation of a locus disrupted in a transgenic mouse mutant displaying sex-linked cleft palate.

(ii) The identification and isolation of transcribed sequences from the wild type locus.

(iii) Evidence of mutational disruption of the gene at the locus in the transgenic mutant (splicing of the transcript identified at this locus is altered in the transgenic mouse mutant).

The work presented in this thesis has revealed highly repetitive sequence at this locus containing gene sequences, with a complex splicing pattern, linking several very small exons.

2.15 SUMMARY

32.8kb of the wild type mouse locus whose disruption in a transgenic mouse line resulted in sex-linked cleft palate has been isolated. This locus has been characterised at the level of expression and conservation of sequence across species. This characterisation revealed that the DNA sequences at this locus are replete with repeated sequences. The method that proved successful in isolating potential transcripts was that of exon-trapping. To date five individual exons have been isolated from DNA at this locus (Table 1) and mapped onto the isolated DNA from the wild type locus (Figure 47). The exons range in size from 81 base pairs (p52) to 446 base pairs (p50) and four exons were found to contain ORFs that span the exon (p51, p52, p56 and p57). Searches of sequence databases have revealed that two of the four exons found to contain complete ORFs (p51 and p56) do not share any sequence similarity at all with any sequence present in the database. One exon (p52) shares a high level of similarity with Ca^{2+}/calmodulin-dependent protein kinase II (CaM kinase II), while the largest exon (p50, which doesn't contain a spanning ORF) is composed solely of repetitive sequence belonging to the LINE-1 class of repetitive elements.

Using RT-PCR two of these exons (p52 and p57) have been shown to be expressed at the level of RNA during development and to be spliced together. The predicted amino acid sequence of this RT-PCR product is very similar to that of CaM kinase II. RT-PCR analysis between these two exons in the transgenically positive line that exhibited sex-linked cleft palate resulted in several larger RT-PCR products that hybridise to LMP-1 sequences. This suggests that in the mutant the splicing of the RT-PCR product is defective. Visualisation of secondary PCR products from exon-trapping of cloned DNA suggests that other exons remain to be identified from this locus. The cloning of these exons and the RT-PCR products, along with the continued characterisation of this locus will facilitate the isolation of genes present.

Table 1: Summary of exon characterisation

Exon	Size(bp)	ORF	Homology	Expression (by RT-PCR)	Exon Linkage
p50	446	-	LINE-1	?	?
p51	205	+	none	?	?
p52	81	+	CaM Kinase	+	to p57
p56	99	+	none	?	?
p57	101	+	CaM Kinase (partial)	+	to p52

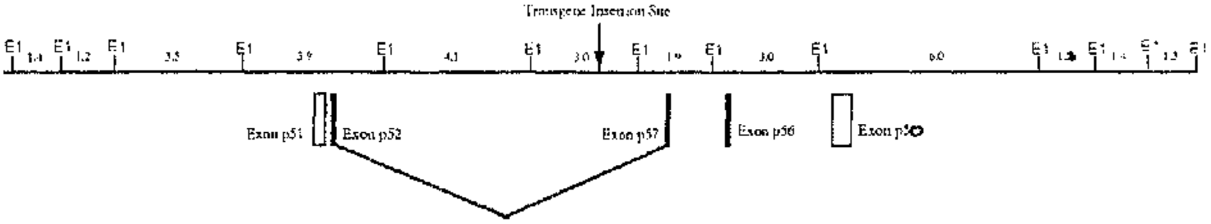


Figure 47: Schematic representation of exons trapped from isolated DNA. Exons are represented by the rectangles. The solid line between exon p52 and exon p57 indicates that these two exons are spliced together as shown by RT-PCR.

Scale
1Kb

3 DISCUSSION

3.1 Repetitive DNA sequence at the locus

Using fragments of the isolated DNA as probes on Southern blots from this locus revealed that the DNA contained extensive repetitive sequence. When the isolated DNA was used as probes on Northern blots of embryonic and post-natal RNAs these repetitive sequences were found to hybridise to transcripts. The unique sequences that were found during the Southern analysis did not hybridise to transcripts. These results suggested that only the repetitive sequences hybridised to transcripts. The lack of hybridisation observed with the unique sequences could result from one of several possibilities, the most obvious being that there were no expressed sequences present in the unique fragments. The results of the exon-trapping identified that exons were only trapped from repeat containing fragments, supporting the idea that the unique fragments do not contain expressed sequence. However, exon-trapping can only be used to identify internal exons, that is exons that have both splice donor and acceptor sites. Moreover, there is the possibility that several exons contain *EcoR I* sites, which will have been missed in the protocol used. Therefore, not all of the internal exons as well as the missing 5' and 3' exons could be isolated, as a result it remains feasible that unique stretches do harbour exons that are too small to have been detected by the Northern approach.

The analysis of conservation of sequence from this locus revealed several fragments were conserved between different mouse species e.g., the 1.3kb *EcoR I-Not I* from Phage 7, while others were conserved into human e.g., the 4.1kb *EcoR I* fragment from Phage 4. The presence of repetitive sequences complicated this analysis of conservation of sequence and so, quashing was used to overcome this problem. The quashing of repetitive sequence revealed that the Phage 4 4.1kb *EcoR I* fragment contained sequence that was conserved into rat, hamster, marmoset and human DNA samples at medium stringency of washing. Other fragments exhibited no unique sequence conservation across species whenever the repetitive sequences were quashed, e.g. the 6kb and 3kb *EcoR I*, however no unique signal was observed in the mouse DNA samples, an indication that the entire fragments are composed of highly repetitive elements. The results of the analysis of conservation of sequence revealed that there were several regions that were conserved at a low stringency, however the majority of the sequences from this locus did not contain unique sequence that was conserved in humans.

The patterns of expression observed with repetitive fragments may not reflect the actual expression of the sequences contained within the fragment. Rather these sequences could hybridise to expressed sequences also containing the repeats that are not derived from this locus. Employing the exon-trapping methodology an exon (exon p50) was isolated that was composed solely of repetitive sequence, related to LINE-1 elements. It was trapped from the 6kb *EcoR I* fragment of Phage 1 and Phage 7, so it is probable that the pattern of expression seen by Northern analysis with this fragment reflected the expression of related sequences in many other transcripts. Sequencing of this fragment and subsequent database searches revealed that the sequences found in exon p50 were widespread in the mouse genome, so the sequences of the 6kb *EcoR I* fragment probably hybridise to other expressed sequences not from this locus. Exon p50 also hybridised to the 4.1kb *EcoR I* fragment from Phage 4, though no exons were trapped from this fragment suggesting that the expression pattern observed with this fragment is a result of hybridisation to repeats expressed from other loci.

Sequence analyses of some of the isolated genomic DNA fragments and of the trapped exons revealed the presence of highly repetitive DNA, the database analyses showed that a large proportion of this sequence belonged to a group of elements known as SINE (composed of short, interspersed Alu and Alu-like repeats) and LINE (long interspersed repeated elements). These two families of repeated sequences constitute the majority of middle-repetitive DNA (Charlesworth *et al.*, 1994). The SINE and LINE sequences along with processed pseudogenes coding mRNAs and small nuclear RNA pseudogenes share common properties suggesting that all three are generated from RNA intermediates. Therefore, they have been classified as retrotransposons if they possess the ability to code for functions to facilitate the retrotransposition process or, retroposons when they do not encode reverse transcriptase. The main features of the LINE-1 retrotransposons are detailed in the introduction.

LINE-1 sequences are 6-7kb in size and are present in some mammalian genomes at 10^4-10^5 copies. They comprise 1-8% of the genome. In primates there is a single major LINE family called L1H and in mice another called L1M. L1H and L1M are 60-70% homologous, except at the 3' end (Holmes *et al.*, 1994). These retrotransposable sequences have variable (A)-rich tails at their 3' end and are flanked by short repeats. Their potential ability to multiply and re-insert into the genome means that LINE sequences may be a powerful form of mutation in the genome. The exact process by which LINE sequences have replicated is not fully understood. But the retrotransposons appear to have a preference for (A)-rich sequences to integrate into. The extent of these repetitive sequences at this locus may have been factorial in the problems encountered in cloning this locus. It was far beyond the scope of this project to investigate the nature or reason for unstable recombinant phage λ clones, (it was difficult enough to isolate stable clones). But, the presence of highly repetitive sequences would allow the bacteria that the phage were propagated in to homologously recombine the sequences and loop out sequences, complicating the cloning process.

It is interesting to speculate about the presence of the LINE-1 elements at the *DXRib 1* locus, whose disruption resulted in the mutant phenotype. There have been several documented cases of gene mutation as a result of LINE-1 insertion (reviewed in Amariglio and Rechavi, 1993). To date there have been three recorded LINE-1 insertions into the dystrophin gene of humans (Holmes *et al.*, 1994). One of these insertion events results in muscular dystrophy with a phenotype intermediate between that of Duchenne and Becker muscular dystrophies. Analysis of the affected kindred revealed that a truncated *LRE* (LINE Retrotransposable Element), 2kb in size, had inserted into exon 48 of the dystrophin gene and resulted in the mutant phenotype (Holmes *et al.*, 1994). The definitive precursor of the *LRE* has been isolated and is one of a growing number of active *LRE*s that have been isolated and analysed. The study of these isolated *LRE*s will help increase the understanding of the mechanisms used by this family of retrotransposable elements to transpose.

LINE-1 insertional mutants have been recorded in the mouse. One example is the well characterised disorder,the spontaneous mouse mutation, *spastic* (*spa*). At about two weeks of age, *spa* homozygotes develop a severe neuromuscular disease characterised by an exaggerated startle response, rapid tremor, myoclonus, rigidity, abnormal gait and impaired righting reflex. Kingsmore and his colleagues have shown that defects in the gene encoding the β-subunit of the glycine receptor (*Glrb*) underline *spa*. *Glrb* maps to the same region of mouse chromosome 3 as *spa*, and *Glrb* mRNA is markedly reduced throughout brains of *spa* mice, most likely as a result of an insertional mutation of a 7.1kb LINE-1 element within intron 6 of *Glrb* (Kingsmore *et al.*, 1994).

Several questions arise from the presence of these LINE-1 sequences at this locus. Do the LINE-1 sequences form a complete element and if so, is it capable of transposition? To answer these questions is beyond the scope of this project. However, we know that the LINE-1 sequences are not involved in the mutant phenotype observed at this locus, as these LINE-1 sequences are present at this locus in the wild type mouse. The search of the BLAST database revealed the wide dispersal of the LINE-1 sequences throughout the genomes of mammals. Not only was the LINE-1 sequence associated with the pseudogenes and small ribosomal RNAs mentioned earlier, they were also associated with the 5' and 3' of many genes (see sequence analysis in section 2.5 and section 2.8.1)

It is difficult at this stage, without a complete cDNA of the gene found at this locus, to know whether exon p50, the 450 base pair exon is part of a gene transcript from this locus or if the LINE-1 sequences harbour a common exon that is trappable but forms no part of the gene at this locus. The 450 base pairs of the exon p50 have been shown to contain several ORFs, though a nucleic acid motif search of the sequence showed the presence of stop codons in all reading frames and no splice sites, suggesting that it cannot be functionally translated. However, a recent letter to the editor of Cell (Margalit *et al.*, 1994) remarked on the presence of two Alu elements in the coding region of *RMSA-1* (regulator of mitotic spindle assembly 1) indicating that it is possible for transposon derived sequences to be part of protein encoding regions of genes.

The first of these two Alu sequences, including its poly $[A]^+$ tail, resides completely within the *RMSA-1* gene. While the second Alu is flanked by direct repeats of exactly 9 base pairs, a hallmark of transposition, there is no reminiscence of such direct repeats associated with the first Alu element. This would indicate that the first Alu element had lost its direct repeats as a result of sequence divergence, or that the first Alu element is not a transposition product. This observation has important implications for the evolution of genes and transposable elements alike and throws up an intriguing paradox. 40% of the *RMSA-1* gene is composed of Alu elements, yet this essential gene is conserved into mouse and chicken (Yeo *et al.*, 1994). The mouse genome is known to lack a dimeric Alu element and chicken does not have any Alu-related sequences at all. So the gene product seems old but the Alu elements are new. Only sequence analysis of the mouse and chicken homologs of *RMSA-1* will reveal the answers, or provide a few suprises.

Although LINE-1 sequences were found in the sequence information obtained from several fragments of cloned DNA from the *DXRIB 1* locus they alone do not account for all the repetitive sequences, other repetitive elements are present. One of these sequences is a mini-satellite repeat $(GA)_n$ that occurs in the 1.3kb *EcoR I* + *Not I* DNA fragment from this locus. The use of this fragment as a probe on a

Zoo blot showed that the repetitive sequences are present at low copy numbers in the mouse genome. However this repetitive element appears to be present at a much higher copy number in primate genomes. When this sequence was compared to sequences present in the BLAST database, the only matches were found in primate sequences.

3.2 Characterisation of exons trapped from this locus

When the total secondary PCR products from the exon-trapping procedure were used as probes on Northern blots of embryonic and post-natal RNAs an interesting pattern of expression was found. Three transcripts were detected, one at about 5kb, one at 0.8kb and one at 0.5kb. The 0.8kb and 0.5kb transcripts were detected in the body sample of embryonic day 12.5 total RNA and increased in the level of expression during development. These transcripts were also detected in embryonic head RNA samples from embryonic day 13.5 onwards. Post natally the 0.8kb transcript was detected in bone, spleen and stomach total RNA samples from day 13.5 post natal mice, indicating that this transcript is also expressed post natally in these tissues. The expression of the 0.5kb transcript post natally was not addressed.

Probing of a Northern blot of embryonic RNA with total secondary PCR products trapped from Phage 4 detected another transcript at day 13.5 in total body RNA that did not appear to be expressed in any of the other RNA samples tested. The transcript was about 5kb in size and was not detected in a total head sample at day 13.5 of embryonic development. This suggests that exons present at this locus either belong to alternatively spliced transcripts or to separate transcripts. This 5kb transcript was not detected with exons derived from Phage 1 and Phage 7 suggesting that the exons that hybridise to the 5kb transcript are located only in Phage 4 DNA, although the absence of the 5kb transcript on blots probed with total secondary products from Phage 1 and Phage 7 may reflect the quality of the RNA samples on the Northern blots probed or the conditions of hybridisation used.

To date, five individual exons have been isolated from the total PCR products of exon-trapping. Southern analysis revealed the presence of the exons in the cloned DNA they were trapped from. Using the exons as probes on genomic Southerns and Northerns awaits optimisation of conditions. RT-PCR analysis demonstrated that exon p52 and exon p57 are expressed during mouse embryogenesis.

3.2.1 Putative gene(s) present at this locus

Sequence analysis and database searches have been carried out in order to identify a potential gene that the isolated exons belong to. These database searches with the nucleic acid sequence of the exons and putative amino acid sequence have revealed that some of the exon sequences are novel and that the other exons show interesting similarities to sequences already present in the database.

The sequence of exon p50 is composed of repetitive sequence as shown by Southern analysis and matches found with the database search. The repetitve sequence has been identified as belonging to the LINE family of repetitive sequences, detailed above. No ORF was found to span the exon suggesting that it may belong to the 5' or 3' untranslated sequences of a gene or alternatively be unrelated to the gene(s) at this locus as discussed above.

Another possibility is that exon p50 is a pseudoexon. Pseudoexons belong to a rare form of pseudogene (Thompson *et al.*, 1995). Most pseudogenes were processed mRNAs that have been reverse transcribed and inserted back into the genome. Pseudoexons belong to a class of pseudogene that contain functional exons and introns, i.e. exons that are capable of being functionally spliced. Thompson and his colleagues report that these pseudoexons are capable of being trapped using the exon-trapping protocol. For people investigating large regions of DNA the presence of pseudoexons in genomic DNA is another complication in the isolation and characterisation of genes. Several groups have cloned pseudoexons using various methods from a region on chromosome 5 that spans the Spinal Muscular Atrophy

(SMA) critical region. Exons have been identified that are 90% identical at the nucleotide level to exons of β-glucuronidase, the functional gene for which is on chromosome 7. Two variant forms of the gene encoding promelanin-concentrating hormone (the functional gene being located on chromosome 12) have also been isolated and a pseudogene of a novel member of the cadherin gene family (Thompson *et al.*, 1995). However, the pseudoexons found to date contain stop codons and so cannot be translated allowing them to be identified as 'pseudo'.

Exon p51 shows only a low degree of similarity with a limited number of sequences in the database in the BLAST database, while exon p56 shows no similarity to any sequences in the database. The database searches using these exons putative amino acid sequence revealed that although many amino acid sequences shared short regions of homology there were no sequences present that had a high level of similarity over the majority of the sequence of the exons. The results of these database searches would indicate that these exons belong to a gene that has not been identified before and is unique. However these exons are very small and therefore represent only a small part of a gene's sequence. This may mean that the sequence of these exons represents a region of a gene that is not very highly conserved and would therefore not be identified by a database search.

Sequence comparisons carried out using exon p52 and the RT-PCR product of exon p52-p57 have been most informative regarding the possible nature of the gene(s) at this locus. Exon p52 was found to contain two ORFs that span the entire exon in different plus reading frames. Database searches using the nucleotide sequence and deduced amino acid sequence of the putative ORF1 of exon p52 revealed similarities to the nucleotide sequence and amino acid sequence of Ca^{2+}/calmodulin-dependent protein kinase II isoforms (CaM kinase II). A database search using the sequence of exon p57 alone produced no matches at the nucleotide level and the deduced amino acid sequence of putative ORFs revealed only short regions of homology with other proteins. RT-PCR analysis of embryonic RNA has shown that not only are exon p52 and exon p57 expressed, they are spliced together in the same transcript. When this RT-PCR product was sequenced an ORF was found that spans both of these exons. Translation of this ORF, using the MacVector^{TM} program, resulted in an amino acid sequence that is the same as the deduced amino acid sequences of the ORF of exon p52 and the ORF of exon p57, joined together. This suggests that in this transcript at least, ORF3 of exon p52 and the ORF of exon p57 form the *bona fide* polypeptide.

The amino acid sequence of the other complete ORF (ORF2) of exon p52 was found to have a low level of similarity to the HMG-box sequence of *Sry* from mouse. The 28 amino acids of ORF2 exhibited a low level of identity 34% to a region of the SRY protein of mouse (Gubbay *et al.*, 1990). When similar amino acid residues are included the level of similarity increases to 69%. A comparison to the HMG-box sequences from several HMG-box containing proteins did not identify any higher level of sequence similarity. The HMG-box motif is a 79 amino acid sequence that contains many basic residues and is a DNA binding motif. It has been found in many different proteins with known or suspected DNA-binding properties, including several transcription factors (Ner, 1992). The importance of the HMG-box has been emphasized by the finding that the HMG-box alone confers DNA-binding activity. In the case of *Sry*, the binding of DNA induces a substantial bend in the DNA which may be necessary for the formation of transcriptional complexes (Goodfellow and Lovell-Badge, 1993). However, with the splicing of exon p52 to exon p57, the ORF reaches a stop codon in exon p57 and is not identified by an ORF search of the joined sequences. It is not possible to state whether the ORF2 is part of an authentic polypeptide expressed from a different transcript or merely a coincidence of the sequence contained within exon p52.

When the extended sequence of the RT-PCR product spanning both exons p52 and exon p57 was used for database analysis, CaM kinase II isoforms were identified as being highly similar at both the nucleotide level and amino acid level. The region of similarity spans the splice site between the two exons. This is interesting as a search of the BLASTP database with the putative amino acid sequence of exon p57 failed to find any similaritites with Ca^{2+}/calmodulin-dependent protein kinase sequences present in the database. The lack of matches between the sequence of exon p57 and those in the database is probably a result of the parameters of the database search. The CaM kinase II sequence exhibiting the highest level of identity (47%) was that of the mouse γ subunit of brain-specific Ca^{2+}/calmodulin-dependent protein kinase II. Analysis of the amino acid sequence of the rat γ subunit of brain-specific Ca^{2+}/calmodulin-dependent protein kinase II, revealed that the region of highest similarity with exon p52-exon p57 is the region constituting the calmodulin binding domain. This region is the most highly conserved in the family of multifunctional Ca^{2+}/calmodulin-dependent protein kinases, CaM kinase, (Tobimatsu *et al.*, 1988). This analysis suggests that the potential polypeptide of exon p52-exon p57 is a member of this family of proteins.

3.3 MULTIFUNCTIONAL CaM KINASE FAMILY

3.3.1 Isoforms of Ca^{2+}/calmodulin-dependent protein kinase

CaM kinase II is a ubiquitous enzyme mediating the diverse effects of hormones and neurotransmitters that utilize Ca^{2+} as a second messenger. CaM kinase II is present in most tissues as an oligomer, composed 6-12 subunits, depending on the isoform and tissue. Differences between the cloned mammalian isoforms mostly lie in a variable region which lies between the calmodulin binding domain and the association domain and consist mostly of 11-39-amino acid insertions and deletions. Thus far, five isoforms (α, β, β', γ, and δ) have been cloned from rat brain. α-CaM kinase cDNA encodes a protein of 478 amino acids (54kDa) from a 5.0kb message. The β-CaM kinase cDNA is translated to a 542-amino-acid protein (60kDa) encoded by a 4.8kb message. The α and β sequences are closely related, with 91% identity in the amino terminal half of the protein, and 76% identity in the carboxy-terminal half. Alternative splicing generates both the β mRNA and a distinct β' isoform (Hanson and Schulman, 1992). Two other isoforms designated γ and δ, have been cloned from rat brain cDNA libraries (Tobimatsu and Fujisawa, 1989; Tobimatsu *et al.*, 1988). Northern blots demonstrate a broad tissue distribution for both the δ and γ subunits, with γ detectable in all tissues examined and δ detectable in all except liver. Each of these is a distinct gene product with approximately 85% homology in amino acids and 75% homology in nucleotides to the α subunit. The γ cDNA encodes a protein of 527 amino acids (59kDa), and the δ one of 533 amino acids (60kDa). It should be noted that antibodies to CaM kinase II have detected subunits, that do not correspond in size to any of the cloned isoforms, in both neuronal and non-neuronal tissues. Therefore, it is likely that additional isoforms that are either distinct gene products or alternatively spliced variants of the known isoforms exist.

Expression of the α and β isoforms of CaM kinase appears to be mostly confined to the brain. The β isozyme of type II CaM kinase is expressed specifically in the brain of the rat, it appears to be concentrated in the forebrain postsynaptic densities. It acquires a Ca^{2+}-independent activity when it phosphorylates itself, suggesting that it may be involved in producing long lasting changes in response to transient calcium signals (Miller and Kennedy, 1986). The rat γ and δ isoforms appear to be more widespread as RNA blot analysis shows transcription in many rat tissues (Tobimatsu and Fujisawa, 1989). All isoforms share a highly conserved catalytic domain at the amino-terminal portion of the molecule, an autoinhibitory sequence overlapping with a calmodulin binding region, followed by an association domain which is important in holoenzyme formation (Figure 48).

Nghiem and colleagues have reported the cloning of two new isoforms of CaM kinase, γ_B and γ_C, from non-neuronal human tissue, the sequence of their calmodulin binding domains is shown in Table 2 (Nghiem *et al.*, 1993). Outside of the variable region all three variants of the γ CaM kinase II are nearly 100% identical. Human γ_B CaM kinase II differs from the rat brain γ_A by the insertion of a novel 23-amino acid segment in the variable region of γ_B, and the deletion of two segments of 21 and 11 amino acids. The human γ_C clone differs from human γ_B in that it does not include the 23-amino acid insert (Nghiem *et al.,* 1993). The

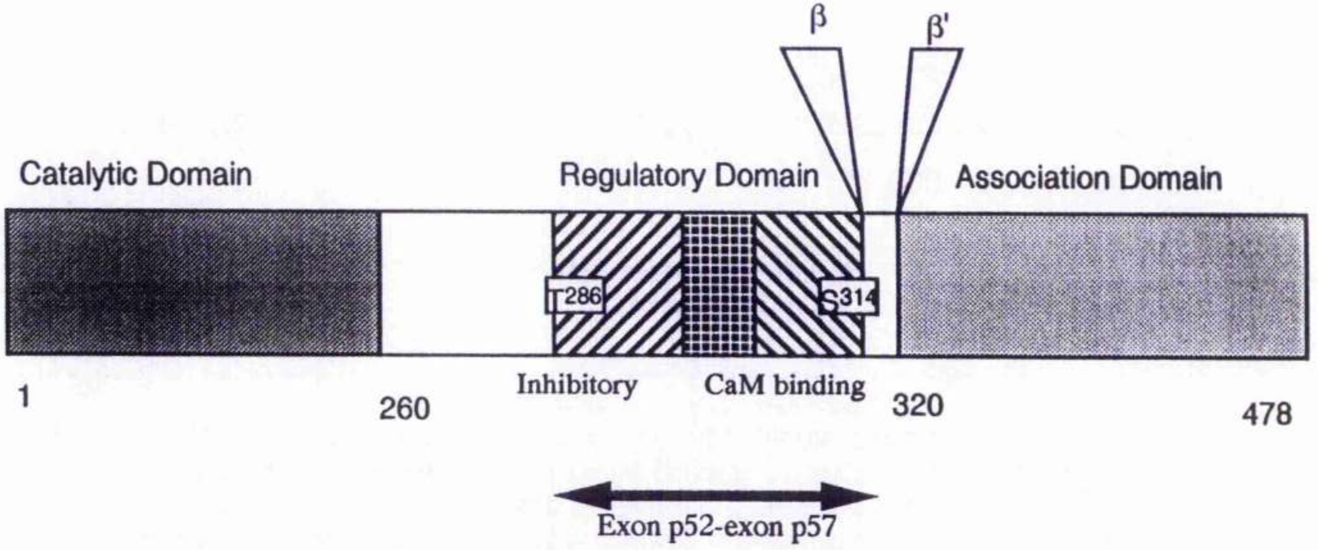


Figure 48: Domain structure of the multifunctional CaM kinase α-subunit

Putative phosphorylation sites are shown flanking the inhibitory/calmodulin-binding domains and the site of variable segments in the β-and β'- subunits is shown also. The region of sequence that is similar to that of exon p52-exon p57 is indicated by the solid black arrow below the α-subunit.

comparison of amino acid sequences from CaM kinase γ_B and γ_C with that of exon p52-exon p57 reveal that they share a 47% level of identity across the 44 amino acid residues shown in Table 2. Comparison of the exon p52-exon p57 sequence to that of CaM kinase II-β reveals a 46% level of identity. However, the sequences of CaM kinase II-γ_B/γ_C and CaM kinase II-β are more similar at 75% identity, to each other than either are to the sequence of exon p52-exon p57. These sequence comparisons demonstrate that the amino acid sequence of exon p52-exon p57 is related to the amino acid sequences of the members of the CaM kinase family and may be a unique new isoform or a related homologue.

Other members of the CaM kinase family have been described. CaM kinase I has been isolated from bovine brain and found to be a 42 kDa monomeric enzyme with a kinase activity similar to that of CaM kinase II (Nairn and Greengard, 1987). It differs from CaM kinase II in the site of phosphorylation on its substrate. CaM kinase IV is a 474 amino acid protein and its kinase domains maintain the conserved features of protein kinase catalytic domains, but is only 32% identical overall to CaM kinase II (Hanson and Schulman, 1992). One gene was found to encode both calspermin and CaM kinase IV utilising two transcriptional start sites to generate the different messages (Ohmstede *et al.*, 1991). Calspermin message is only detected in the testis, while CaM kinase IV has been detected in brain, testis, spleen and thymus. The kinase activity of CaM kinase IV differs from that of CaM kinase II in that it phosphorylates substrates not phosphorylated by CaM kinase II, e.g. Rap-1b (a Ras related GTP-binding protein) (Sahyoun *et al.*, 1991). It may be that CaM kinase IV is another multifunctional CaM kinase given the different substrates to those of CaM kinase II. To address this requires the identification of other substrates and functions of CaM kinase IV *in vivo* .

3.3.2 Gene structure and domain structure of CaM kinase

The β subunit of brain-specific Ca^{2+}/calmodulin-dependent kinase II was the first member of the CaM kinase II family to be cloned from mouse,*Camk2* (Karls *et al.*, 1992). Sequence analysis of the coding region showed 97% homology at the nucleotide and close to 100% homology at the amino acid level with the published sequence of the rat β CaM kinase II gene. This high level of homology extended into the 3' untranslated region of the gene (Karls *et al.*, 1992). The genetic structure of *Camk2* appears rather complex with the 16 exons, ranging in size from 38bp to 233bp, that comprise the 3' untranslated region and four-fifths of the coding region spread over 28kb of cellular DNA. The β' isoform appears to be generated by alternative splicing of the transcript of the β gene, resulting in the deletion of exon VI. The 5' end had not been isolated for presentation in this paper (Karls *et al.*, 1992).

The majority of the isoforms identified to date have been isolated from various cDNA libraries, resulting in a lack of information of the gene structure for each isoform. Thus, it is not clear as to whether the exon/intron structure is maintained or not between isoforms. However what is clear is that the domain structure of the proteins encoded by these isoforms is similar.

The deduced amino acid of the exon p52-exon p57 ORF shows homology to the regulatory domain of type II CaM kinases. This is composed of the primary phosphorylation sites, the inhibitory subdomain which maintains the kinase in an inactive conformation and a calmodulin-binding subdomain that is unique to the Ca2+/calmodulin-dependent protein kinases. The core of the calmodulin-binding subdomain is the amino acid sequence RRKLKGAILTTMLA. Synthetic peptides corresponding to this region of the kinase bind calmodulin in a Ca^{2+}-dependent manner thereby blocking activation of calmodulin-dependent kinases. In the β CaM kinase II, as in many other CaM-binding proteins, this domain consists of a basic α helix which, in a helical wheel projection, forms a basic and a non-polar surface (reviewed in O'Neill and DeGrado, 1990).

The bipartite character of this core binding domain is emphasised by the fact that it is split between two exons in such a way that the basic part (amino acid residues R-297RKLK-301) is in one exon and the non-polar part (amino acid residues G-302AILTTMLA-310) is in another exon, a splicing event brings these two sub-domains together. As described earlier the sequence of the binding domain is found in the deduced amino acid sequence of the RT-PCR product and demonstrates an identical split of residues between exon p52 and exon p57. There are differences though, while the basic part of the domain is present fully (RRKLK) in exon p52, the non-polar part of the domain does not appear to be as well conserved (**GAVLAAV**SS) in exon p57 (substituted residues are underlined). However the amino acid substitutions that have occurred retain the non-polar character of the amino acid sequence (non-polar residues are shown in bold type). A deduced secondary structure for the amino acid sequence of these two exons reveals that a putative α helix would be formed, corresponding to that formed by the bipartite binding domain of type II CaM kinases (Schulman and Lou, 1989).

An inhibitory subdomain may extend the N-terminal portion of the calmodulin-binding domain to include a threonine residue (Thr286) which is autophosphorylated upon calmodulin binding to CaM kinase. Synthetic peptides corresponding to this region inhibit activity of the autonomous kinase *in vitro*. It is also interesting that the two threonine/serine residues which match the substrate consensus sequences for the kinase that are found in both the α-, β- and γ- subunits, Arg-X-Y-Thr286 and Arg-X-Y-Ser314, flank the inhibitory and calmodulin-binding subdomains. These two threonine/serine residues also occur in the deduced amino acid sequence of the exon p52-exon p57 ORF and are located in a similar position relative to the calmodulin-binding domain. However in the amino acid sequence of exon p52-exon p57 the substrate consensus sequence is not present. Part of the deduced amino acid sequence of exon p52-exon p57 ORF is given below, the calmodulin binding domain is shown in bold type, while the threonine/serine residues are shown in bold type and underlined.

(LPET-286VEQLRKFNA**RRKLKGAVLAAVS**SHKFNS-314FYD)

For a comparison of amino acid sequences of the calmodulin-binding domains of CaM kinase II and those of several known protein kinases, with the

Table 2

Calmodulin-binding domains of several enzymes

IV	TGK----AANFVHMDTAQKKLQEFNARRKLKAAVKAVVASSRLGSASSSHTS
II-α	SHR--STVASCMHRQETVDCLKKFNARRKLKGAILTTMLATRNFSGGRSGGN
II-β	CQR--STVASMMHRQETVECLKKFNARRKLKGAILTTMLATRNFSVGRQTTA
-M	QKD----TKNMEAKKLSKDRMKKYMARRKWQKTGHAVRAIGRLSSMAMISGM
-K	NNLAEKAKRCNRRLKSQILLKKYLMKRRWKKNFI-AVSAANRFKKSSSGALM
PHK-γ	QEY--------------VVEEVRHFSPRGKFKVICLTVVASVKIYYQYRRVKP
$\gamma_{B/C}$	-------VASMMHRQETVECLRKFNARRKLKGAILTTMLVSRNFSAAKSL--
p52-p57	-------YAYKIHLPETVEQLRKFNARRKLKGAVLAAVSSHKFNSFYGDPP-

Comparison of the amino acid sequence of the putative calmodulin-binding domain of exon p52-exon-p57 with those of several known CaM kinases and protein kinases. Single-letter amino acid code is used and the amino acids that are found in exon p52-exon p57 are in bold type. Predicted calmodulin-binding domains are underlined. Alignment is by visual inspection and effort was made to minimize gaps. Enzymes listed: IV, CaM kinase IV (mouse); II-α and II-β, -α and -β subunits of CaM kinase II (rat), -M, myosin light-chain kinase from smooth muscle (chicken); -K, myosin light-chain kinase from skeletal muscle (rat); PHK-γ, phosphorylase kinase γ subunit (mouse) (Sikela *et al.*, 1989)); $\gamma_{B/C}$, γ_B and γ_C isoforms of CaM kinase II (human) (Nghiem *et al.*, 1993).

deduced amino acid sequence of exon p52-exon p57 see Table 2. The comparison reveals that the sequence of exon p52-exon p57 shows a high level of similarity with mouse CaM kinase II-β and -γ. However, the complete amino acid sequence of exon p52-exon p57 shows only a 46% level of identity to the CaM kinase II-β partial sequence. As the β subunit of CaM kinase II is brain specific in mouse and rat we can suggest that exon p52-exon p57 is part of a gene that may be a new CaM kinase isoform. The exon p52-exon p57 sequence was a product RT-PCR analysis of total RNA derived from the body of day 15.5 embryonic mice, this shows that the isoform found by this analysis is not brain specific.

3.3.3 Functions of CaM kinase

CaM kinase II was first specifically described in synaptosomal (nerve terminal) membranes (Schulman and Greengard, 1978), and was subsequently shown to be present in membranes from a variety of tissues (Kennedy and Greengard, 1991). This kinase phosphorylated the vesicle-associated protein synapsin I as well as several other proteins, two or three of which were later found to be autophosphorylated subunits of the enzyme itself. CaM kinase II was purified from rat forebrain cytosol using a number of substrates, including myosin light chains, synapsin, tryptophan hydroxylase, tubulin, microtubule-associated protein-2 (MAP-2) and casein. All of the isolated kinases share the broad substrate specificity characteristic of CaM kinase II and are now recognised as being the same enzyme (Schulman and Greengard, 1978).

CaM kinase II prepared from rat forebrain typically contains subunits of 50-55kDa (α- subunits) and 58-60kDa (β- subunits) in a ratio of 3-4:1. Four Ca^{2+} ions bind to a molecule of calmodulin and this then binds to CaM kinase, this in turn stimulates autophosphorylation of both subunits of the CaM kinase. The CaM kinase II holoenzyme isolated from rat forebrain has an apparent molecular weight of between 460,000 and 650,000 and is therefore likely to be a decamer or dodecamer. As detailed above, seven isoforms of CaM kinase II have been isolated to date. Thus far all mammalian isoforms identified utilise the same protein domain structure and differ from each other primarily in the size and nature of the inserts at the end of the calmodulin binding domain.

Mechanistic studies of multifunctional CaM kinase and its activation by Ca^{2+}/calmodulin have provided evidence for a model in which CaM kinase is kept relatively inactive by the presence of an autoinhibitory domain (Figure 48). Binding of Ca^{2+}/calmodulin eliminates this inhibitory constraint and allows the kinase to phosphorylate its substrates and also itself. This autophosphorylation significantly slows dissociation of calmodulin thereby trapping calmodulin even when Ca^{2+} levels are below those initially needed for activation. Once calmodulin dissociates, CaM kinase remains Ca^{2+}-independent until the kinase is dephosphorylated. These effects of autophosphorylation enable the potentiation of brief transient Ca^{2+} fluxes.

The primary role of CaM kinase is as a kinase that regulates cellular processes by phosphorylating critical proteins in the cell. A growing list of putative substrates is being formed on the basis of *in vitro* assays in which proteins are found to be phosphorylated by the kinase. This list can be used to identify the consensus amino acid sequence around the phosphorylation sites recognised by CaM which permit identification of other possible substrates. Table 3 provides a list of the proteins that have been found to be good substrates of CaM kinase in *vitro* and *in vivo*. Consistent with being a multifunctional kinase, the enzyme phosphorylates a diverse group of proteins that includes enzymes, cytoskeletal proteins, ion channels and transcriptional factors. Other proteins have been identified as substrates of CaM kinase II *in vitro* but remain to be shown as being substrates *in vivo* and are not described here (Hanson and Schulman, 1992).

Distinct functions have not been demonstrated for all the isoforms identified to date, but their variety and differential expression suggest that they have become specialised for roles they play in multiple tissues. A role for CaM kinase, which has been identified in some nonneuronal tissues, includes regulation of a chloride-specific ion channel in human tissues. Cystic fibrosis (CF), a common lethal genetic

Table 3: CaM kinase substrates*

Protein (size kDa)	Function
Acetyl-CoA carboxylase (260)	Fatty acid synthesis
Alzheimers's disease amyloid precursor peptide(228)	
Calcineurin (58-61)	Phosphatase
Caldesmon (87)	Regulates actin-myosin binding
Glycogen synthase (85-90)	Carbohydrate metabolism
Myosin light chain, smooth muscle (20)	Myosin subunit
Myosin light chain kinase, smooth muscle (130)	Initiates muscle contraction
Phenylalanine hydroxylase (51)	Tyrosine synthesis
Phospholamban (11)	Ca^{2+} uptake in SR
Pyruvate kinase (61)	Carbohydrate metabolism
Ryanodine receptor (565)	Ca^{2+} release
Synapsin (80-86)	Neurotransmitter release
Tyrosine hydroxylase (59)	Catecholamine synthesis

*Proteins listed have been shown to be substrates of CaM kinase II *in vivo.* The pathway in which they are involved is indicated in the right hand column. For a complete list of potential substrates *in vitro* and *in vivo*, see Hanson and Schulman, 1991.

disease in Caucasians, leads to defective regulation of chloride ion transport which compromises the function of secretory epithelia of the lung and gut resulting in death. A major pathway for activation of chloride conductances, via cAMP-dependent protein kinase, is blocked by mutations in the CF gene, the cystic fibrosis transmembrane conductance regulator. A parallel pathway utilizing Ca^{2+} as a second messenger remains functional in CF and has recently been shown to be mediated by CaM kinase (Wagner *et al.*, 1991). CF airway epithelial cell lines which were deficient in chloride channel activation by the cAMP pathway were activated by Ca^{2+} ionophore or by injection of activated CaM kinase. Because this pathway to activation of chloride channels bypasses the defective signalling in cystic fibrosis, activation of human epithelial cell CaM kinase by increasing intracellular Ca^{2+} may provide an alternative approach of therapy for this disease.

The mouse CaM kinase II-β gene, *Camk-2,* was isolated from a mutant mouse, CAT40, which harboured a recessive lethal mutation caused by a transgene insertion event on chromosome 11 (Karls *et al.*, 1992). Embryos homozygous for the transgene insertion died shortly after implantation. However, it is highly unlikely that a lack of *Camk-2* expression is responsible for the death of homozygous embryos. *Camk-2* codes for a protein kinase that is found only in the brain. During normal development, transcripts are found for the first time in the head of day 12.5-embryos, much later than the time of death of homozygous embryos. A large deletion was found to accompany the insertion event, with 80kb of cellular DNA 3' to *Camk-2* deleted. To date no gene(s) have been located in this deleted region as attempts to clone the region have proved unsuccessful, due to the DNA inserts being extremely unstable in cosmid and phage vectors. It seems probable that gene(s) vital for embryonic development are located on this region and it is their loss, not Camk-2, that results in the death of embryos shortly after implantation. Thus far, there have been no reports of transgenic knockout mice of any CaM kinase isoforms and what phenotype such knockout mice would exhibit.

3.4 Possible causes of phenotype observed in the transgenic mutant

Exons that contain an ORF which encodes an amino acid sequence very similar to the calmodulin-binding domain found in proteins belonging to the multifunctional CaM kinase family, have been located at the locus whose disruption resulted in sex-linked cleft palate. This family of genes encodes at least seven different isoforms that have a widespread pattern of expression and a role in many different signalling pathways in the cell. Using RT-PCR to analyse the expression of exon p52 and exon p57, which span the insertion site in the mutant, revealed that these two exons are expressed and spliced together in the same transcript. When the expression of these exons was analysed in transgenically positive RNA it was found that the RT-PCR product was altered. The 177 base pair RT-PCR product of exon p52-exon p57 was still present as expected in hemizygous individuals, but a 1kb, a 350bp and a 250bp product were also detected using probes to exon p57 in the transgenically positive spleen RNA sample. Probing of the same blot of RT-PCR products from transgenically positive and negative spleen RNA with the LMP-1 transgene has revealed that the 350bp and the 250bp products hybridise to LMP-1 sequences and these are only found in the transgenically positive samples. These findings suggest that the splicing of exon p52-exon p57 is effected in the transgenic mutant. The sequence of these novel RT-PCR products has yet to be determined, sequencing of this RT-PCR product should demonstrate what other sequences are present in this product. The presence of the 177 base pair exon p52-exon p57 RT-PCR product in the transgenically positive sample reflects the fact that there is still a normal X chromosome present in the transgenically positive spleen. As X chromosome inactivation is random then transcripts can be derived from both the transgenic and the normal chromosome.

The transgene has three LMP-1 encoding exons. It is present in the transgenic mutant as two copies in a head to tail tandem array. Only the central exon contains both a splice acceptor and a splice donor and it is 80bp in size. The difference in sizes between the wildtype RT-PCR product (177bp) and the novel products is about 80bp (for the 250bp product) and about 160bp (for the 350bp product). Therefore it is possible that the single 80bp internal LMP-1 exon is present in the 250bp product and two 80bp LMP-1 exons, one from each transgene copy, are present in the 350bp product. The hybridisation of the RT-PCR products to LMP-1 sequences supports this idea and the cloning and sequencing of these RT-PCR products will reveal the character of these products.

The 1kb RT-PCR product that hybridised with exon p57 sequences found in the transgenically positive sample failed to hybridise to the LMP-1 sequences. But the LMP-1 probe hybridised to a novel RT-PCR product >2kb in size that was not detectable using the exon p57 probe. As the oligonucleotide primers used for the RT-PCR are designed to exon p52 and exon p57 then all RT-PCR products should contain at least the oligo sequences. The reason why this >2kb product fails to hybridise to the entire exon p57 sequences is unclear, but may suggest that these larger products are artifacts. However, the fact that they are not amplified from the wild type RNAs would suggest that they could be genuine products.

Since the 1kb RT-PCR product failed to hybridise with LMP-1 sequences, it is possible that the transgene has affected the splicing of other exons at this locus without being incorporated in this product. If other exons are present in the genomic DNA lying between exon p52 and exon p57, these may be spliced to exon p52 and exon p57 and be amplified by the RT-PCR. The >2kb RT-PCR product that hybridises to the LMP-1 sequences, but fails to hybridise with entire exon p57 sequences may also result from altered RNA processing at this locus as a result of the transgene. It is possible that this is a product of splicing from the transgene into a cryptic acceptor site within exon p57, allowing its amplification through the oligo but with the loss of most of p57 preventing the hybridisation to these sequences. Again, the cloning and sequencing of these RT-PCR products should clarify which of these theories is true.

If the hybrid transcripts that hybridise to LMP-1, exon p52 and exon p57 sequences are translated into protein would this be a functional protein? The fact that exon p52 and exon p57 encode the calmodulin-binding domain, any sequence splicing between these two exons would result in the disruption of this domain. The translation of the hybrid transcript would result in an altered amino acid sequence as the introduction of the exon from LMP-1 which is 80 base pairs in size introduces a frameshift and a stop codon. There is the possibility that the altered splicing suggested from the hybrid RT-PCR products introduced other sequences not yet isolated, which would keep the sequence in frame. It is feasible that the amino acid sequence of the calmodulin-binding domain would be maintained but separated by an intervening stretch of amino acids. In either situation the calmodulin-binding domain is affected resulting in an altered protein.

What effect these hybrid transcripts have in the mutant mouse is unknown. The probing of the RT-PCR products with exon p52 and exon p57 revealed that the 177 base pair product observed in the wild type mouse is also present in the mutant female mouse (mutant male mice have not been analysed). This would indicate that in the transgenically positive female mouse there is a population of transcripts producing wild type and mutant proteins.

The developing palate is a complicated developmental system that is extremely sensitive to changes at the molecular level, disruption of any one molecular process may result in cleft palate (for review see introduction). The cloning of several exons belonging to a putative member of the CaM kinase family may mean that the disruption of a novel CaM kinase results in the phenotype observed. The CaM kinases identified to date exhibit a wide range of substrates, so the phenotype observed in the transgenic mutant may be a consequence from the perturbation of any one of many molecular pathways.

If the gene that exon p52-exon p57 belongs to is a member of the CaM kinase family, then the gene structure of *Camk2* with its 16 small exons, down to as small as 38bp, would help in explaining some of the problems associated with cloning gene(s) from the *DXRiB 1* locus. It is highly unlikely that using the strategy of conservation of sequence and expressional analysis that a 38bp exon would be readily detected. The gene structure of CaM kinase IV and the presence of the calspermin gene within it also offers the possibility that the three transcripts detected by exons from the *DXRib 1* locus may be derived from separate genes and are not simply alternatively spliced transcripts. This exciting possibility relies on the gene, that the exon p52-exon p57 sequence is derived from, belonging to the CaM kinase family.

However, there is the possibility that the exons identified and shown to encode sequence similar to that of CaM kinase II belong to a gene that is not a member of this kinase family but also exhibits binding of calmodulin. The lack of substrate consensus sequences at the Thr-286 and Ser-314 phosphorylation sites suggest that if these exons belong to a novel CaM kinase then it may not undergo autophosphorylation. Therefore this sequence may belong to a gene product which has a completely different function compared to that of identified CaM kinases. Several types of protein that bind calmodulin have been identified in the nucleus of cells, (i) proteins involved in actin-myosin contractile systems, e.g. myosin light chain kinase and caldesmon; (ii) proteins involved in the phosphorylation or dephosphorylation of proteins, e.g. CaM kinase II and calcineurin; (iii) transcription factors, e.g. several transcription factors belonging to the basic helix-loop-helix structural group; (iv) proteins which associate to hnRNA, e.g. hnRNP A2; and (v) proteins of unknown function (Bach *et al.*, 1994). Each of these proteins has very different functions within the nucleus but they all interact with calmodulin in the presence of Ca^{2+}. The cloning of the cDNA from this locus should help in understanding the function of the gene(s) product and its relevance to the mutant phenotype.

The phenotypes observed in the CAT40 mutant mouse probably do not reflect disruption of CaM kinase II-β expression (Karls *et al.*, 1992) and do not help in understanding the events that result from disruption of the gene at *DXRiB 1*. If the gene located at the *DXRiB 1* locus is a member of the CaM kinase II family then it is quite possible that it has a specific function, not identified in any of the other isoforms, the disruption of which impinges on palate development. An understanding of the phenotypes that result from transgene integration at *DXRiB 1* is complicated by the fact that the transgene itself encodes a protein, LMP-1, that has been shown to have pleiotropic effects. The evidence to date is that LMP-1 induces a range of changes in the cell that affects its state of differentiation, with one of the affected cell types being epithelial cells (Wilson *et al.*, 1990).

In palate development the epithelium plays a crucial role in inducing and responding to changes in the mesenchymal cells it is in contact with. Due to the speed of palatal development the epithelium responds rapidly to the inducing signals. If anything changes the differentiation state of the cells in the developing palate then it may be expected that palatal development itself may be affected. It was argued that the fact that these LMP-1 expressing lines which did not exhibit sex-linked cleft palate indicated that it is the insertion event and not transgene expression which caused the mutant phenotype. However, the transgenic line that exhibited cleft palate expressed LMP-1 at the highest level of all the transgenic lines. So the phenotype observed in the mutant line may be a result of the level of transgene expression directed by the CaM locus. This problem can only be rigorously addressed by the regeneration of the transgenic mouse for the site of insertion with a different transgene. If these transgenic mice exhibit sex-linked cleft palate then it is the disruption of the cellular locus that resulted in the mutant phenotype and not induced LMP-1 expression by the cellular locus.

3.5 Further characterisation of the locus

As this may prove to be a unique gene with a potentially important role in palatogenesis then there are an abundance of avenues of research for continued study. The immediate goal of future work will be the isolation of the cDNA in order to obtain the complete sequence. This can be attempted in one of several ways. From the results of the exon-trapping protocol it is clear that there are several exons present at this locus. Some of these exons have already been characterised and cloned. The analysis of the total exons-trapped revealed that some of the exons are part of a transcription unit. Individual exons were shown to be expressed using the technique of RT-PCR indicating that the lack of signal on Northerns, probed with individual exons, is a result of the hybridisation conditions used. RT-PCR can only detect the presence of a sequence in a transcript, it cannot reveal the size of that transcript. Although the RT-PCR revealed that exon p52 and exon p57 were parts of the same transcription unit, it could not identify if that transcript was the same as the transcription units detected with the total exons trapped from the locus.

An immediate goal of future work will be the optimisation of the conditions used to identify the transcripts from which the isolated exons are derived. Once the conditions have been optimised it should be possible to use the individual exons to isolate the cDNA for this locus from the relevant cDNA libraries. The optimisation of hybridisation conditions will allow other analyses, e.g. an analysis of the expression patterns of the locus during development.

3.5.1 Cloning of the cDNA(s) for the gene(s) at this locus

The results of the exon-trapping experiments suggest that there are other exons present at this locus that have been trapped during the exon-trapping but these have not been cloned in the subsequent steps and remain to be isolated. If these uncloned exons, that hybridise to transcripts during embryonic development, could be isolated then the process of isolating the entire cDNA would be aided. There are several ways of overcoming this problem, the first is to repeat the exon-trapping and the subcloning using the DNA from the entire locus restricted with an alternative enzyme. This may result in a similar situation where, for whatever reason, the trapped exons remain uncloned. An alternative approach would be, rather than use all the DNA fragments from this locus to use the individual *EcoR I* DNA fragments for the exon-trapping. By using only individual *EcoR I* DNA fragments for exon-trapping there should only be a few exons trapped per DNA fragment. These exons can then be used as probes on Northern blots of mouse embryonic RNA and characterised in a similar fashion as before.

Two of the transcripts identified as being expressed during mouse embryonic development, with the total exons from this locus are quite small at 800 and 500 nucleotides. The isolation and cloning of exons that belong to these transcripts offer the opportunity to isolate the cDNA in one of several ways. The most obvious method of cloning a cDNA is the construction and screening of a cDNA library. The transcripts revealed by Northern analysis are expressed from day 10.5 of mouse embryonic development, and I have already constructed two cDNA libraries, one from day 12.5 embryonic RNA and the other from day 16.5 embryonic RNA. These libraries can be readily rescreened with any exon that is expressed during development.

There is another more direct approach to cloning the cDNA for the gene(s) at this locus. As the transcripts identified are quite small then it should be possible to use anchored PCR to amplify the cDNA and subclone it. There are two possible approaches using anchored PCR either 3' RACE or 5' RACE. These methods are now standard procedures and a wide range of anchored primers are available commercially. The basis of anchored PCR is that a transcript is converted to a cDNA molecule then either the 5' or 3' end of cDNA molecule is tailed using a single nucleotide during the cDNA synthesis. Then PCR is carried out on the tailed cDNA molecules using a primer to known internal sequence of the transcript and a primer to the sequence added in tailing.

Two of the exons p52 and p57 were shown to be expressed using RT-PCR with RNA extracted from day 13.5, 14.5, 15.5 day 16.5 mouse embryos. It was possible to use RT-PCR to amplify between the exons indicating that the exons were adjacent to each other in the RNA molecule. The cDNA used for this analysis was primed using oligo dT and this indicates that the exons belong to a transcript in the poly[A]$^+$ fraction of the RNA. Therefore it should be possible to use 3' RACE with an anchored primer to the poly[A]$^+$ region of the transcript to produce a cDNA molecule which will be 3' to the exons. It will be interesting to compare the products of the 5' and 3' RACE experiments. It may reveal that the exons trapped from this locus belong to more than one gene or alternatively, the products of 5' and 3' RACE may reveal that the exons belong to the same gene but are parts of a differentially processed transcript.

The cloning of the cDNA(s) isolated from this locus will allow a number of questions to be addressed. A simple comparison of the sequence of this gene(s) to sequences present in the databases will reveal a great deal of information about the gene product. Similarities between DNA sequences and between putative protein sequences may give clues about the protein domains and functions of the gene product. The cloning of the 5' end of the gene(s) and a sequence analysis of it may indicate the sequence domains that are important in gene regulation e.g. binding domains for regulatory proteins which can be tested further. Identification of the regulatory gene sequences will permit an in depth analysis of gene expression control, both spatial and temporal.

3.5.2 Characterisation of expression of the gene(s) identified

With the identification of cDNA molecules derived from this locus the patterns of expression of the gene at the RNA level can be addressed. Again there are several ways in which to approach this analysis. The preliminary results of Northern analysis in this thesis (see section 2.6) suggest that the gene is widely expressed in different tissues during development and post-natally. Although the analysis to date has been limited there appears to be some tissue specificity in the pattern of expression already described. The analysis of the expression between transgenically positive and negative post-natal tissues, namely kidney and spleen, revealed that there was no difference in size of transcripts detected in the spleen RNA samples. However there was no transcript detected in either transgenically negative or positive kidney RNA samples. This clearly shows that although the gene is expressed in many tissues, the expression is not ubiquitous and that there is a level of specificity. The Northern analysis will be extended to many other tissues at different time points during development and post-natally to give a fuller understanding of the pattern of expression at the tissue level.

The extended Northern analysis will reveal only in which tissues the gene is expressed, which will not ascribe a function to the gene product. However if the gene is expressed in one specific cell type then this may help reveal some information about gene function. One of the methods for addressing the pattern of expression within the tissues, which cell types and even the cellular location of the transcripts, is to use the method of *in situ* hybridisation. As the preliminary characterisation shows that the level of expression of the transcripts increases during embryonic development, then it is probable that this gene is important during embryonic development. So whole section mounts of mouse embryos taken at different points during development could be effectively used for *in situ* hybridisation analysis. The sequence analysis of the cDNA should reveal which part of the sequence would be best to use as probe (i.e. least related to known sequences). Either riboprobes or oligos from the isolated cDNA could be used as probes for the *in situ* analysis.

The use of *in situ* hybridisation can help in predicting the function of the gene product, as probing whole embryo sections at different time points of embryonic development will reveal the pattern of expression across the embryo and any changes in expression that are occurring with time. This may take the form of local accumulations or depletions of transcripts at certain times, or prior to certain developmental changes suggesting a particular gene function. A good example of this is the expression of the TGFβ-3 gene expression during palatogenesis. TGFβ-3 is expressed just prior to palatal shelf fusion and only at the tips of the palatal shelves (Ferguson, personal communication). Mice homozygous for the knockout of this gene exhibited the phenotype of cleft palate. The knockout mice, and the temporal and the spatial expression patterns of this gene indicate that it has an important role in palate development.

The expression of a gene at the RNA level while revealing one level of expression does not indicate where or to what degree the protein product of that gene is expressed. To address this requires the generation of an antibody to the gene product. Knowing the sequence of the gene permits prediction of the protein sequence, this in turn allows the design of peptide antigens and the generation of a specific antiserum. With the generation of an antiserum the pattern of expression within tissues and the location of the protein within the cell can be analysed. These studies will take the form of immunohistochemical analysis of the whole mouse embryonic sections as described above. This analysis should reveal the location of the gene's protein product within the cells, complementing the *in situ* analysis described earlier and the sequence analysis of the cDNA.

The antisera may also be used to investigate any modifications that occur to the protein during development or between tissues. This could be addressed using the antisera against Western blots of protein extracts from embryonic samples and post natal tissues, thereby revealing the size of the protein product and changes in gel migration due to post-translational modifications (e.g. phosphorylation) or protein complexing.

As this gene has been isolated as a result of a transgene insertion event resulting in sex-linked cleft palate, then it may be expected to be expressed at high levels during palatogenesis. Alternatively, since other phenotypes were observed in the transgenic mutant mouse, (e.g. the spinal deformities and general skeletal defects) it may suggest that this gene may be of more importance in bone or extracellular matrix formation. This hypothesis is supported by the observation that the gene(s) expression is readily detected in embryonic body RNA samples. There is no reason to suggest that this gene does not have a role in palatogenesis but clearly it could have a wider role in development and post-natal viability. These expression studies coupled with the sequence analysis will help aid our understanding of the structure and function of the gene.

3.5.3 Regeneration of the transgenic mouse exhibiting sex-linked cleft palate

However, to fully understand the phenotype observed in the transgenic mutant will require the extension of the above analyses to the transgenic mutant. Unfortunately the transgenic line that exhibited the mutant phenotype became extinct and so this study cannot be carried out in the original line. However, as mentioned earlier the transgenic line is being recreated. This is being achieved using homologous recombination in embryonic stem cells. Several new lines are being generated. One of these will be the regeneration of the original transgenic line containing the LMP-1 transgene at the same site of transgene insertion. A second line will be generated without the LMP-1 but still disrupting the locus in order to address the contribution (if any) of the LMP to the final phenotype. The final transgenic mutant that is planned is the targeted disruption of one of the exons identified from this locus. This will specifically address the role of the suspected gene in the phenotype.

The regeneration of new transgenic mice should unravel the role of the transgene insertion in the mutant phenotype. By the time that the transgenic mouse is regenerated the wild type gene(s) will already be well characterised and there will be numerous probes and antisera available for a study of the gene's expression in the mutant line. As a potential second messenger (CaM kinase), the chain of signals involved in the morphogenic processes affected by this gene can be begun to be elucidated. The isolation of the gene(s) from this locus and the information gained from analysis of the wild type and mutated locus will allow an insight into the action of a developmentally important gene(s).

4 MATERIALS AND METHODS

4.1.1 Bacterial and bacteriophage culture

All of the strains of *Escherichia coli* and bacteriophage λ used were obtained from Stratagene. The genotytpes of the strains used are given in the Stratagene manual along with maps of the phage used. All of the basic techniques used for bacterial and bacteriophage growth, plating, purification and titration are given in Current Protocols Vols 1 & 2. *E.coli* XL1-Blue (Stratagene) was routinely used for the propogation of plasmids and grown in L-broth with out any antibiotic present. To ensure the maintenance of the F' episome in XL1-Blue bacteria, they were plated on LB agar supplemented with 25μg/ml tetracycline. The plasmid pBluescript (Stratagene) was used for the subcloning of DNA fragments. *E.coli* transformed to ampicillin resistance were grown in L-broth supplemented with 50μg/ml ampicillin or 50μg/ml carbenicillin (a more stable form of ampicillin).

E.coli strains SRB, SRB(P2), XL1-B MRA and XL1-B MRA(P2) (all obtained from Stratagene) used in bacteriophage λ related procedures, were grown in L-broth supplemented with 0.2% maltose and 10mM $MgCl_2$ (LBMM).

4.1.2 Plasmid transformation of *E.coli* XL1-Blue

E.coli XL1-Blue were made competent for DNA transformation, starting from 1ml of a saturated overnight culture used to innocculate 100ml of of prewarmed 2YT broth. The bacteria were grown to an absorbance at 600nm of 0.4 (approximately 2 hours). The culture was cooled on ice and pelleted at 4,000*g* / 5 minutes / 4°C. The bacterial pellet was gently resuspended in 50ml of ice cold 50mM $CaCl_2$ and left on ice for 20-30 minutes. The cells were then repelleted at 4,000*g* / 4°C / 5 minutes and resuspended in 2.5ml of ice cold 50mM $CaCl_2$ for at least 1 hour (cells made competent in this way remained competent for up to 24 hours as long as they were stored in an ice bath). For each transformation, 200μl of competent cells were transformed with no more than 100ng of plasmid DNA in a maximum volume of 80µl. This was incubated on ice for a further 20 minutes, then heat shocked at 42°C for 90 seconds and returned to ice for a further 2 minutes. The transformed cells were added to 800µl of pre-warmed L-broth and incubated at 37°C for 50 minutes prior to plating on L-broth, 1.1% agar, plates (containing the appropriate antibiotic).

These plates also contained the lacZ transcription inducer isopropylthiogalactoside (IPTG) and a chromogenic substrate of β-galactosidase, 5-bromo-4-chloro-3-inodyl-β-galactosidase (Xgal) which allowed for the identification of non-recombinant plasmids. The plasmid pBluescript contains the lacZ gene which allows for the colour selection of recombinant plasmids. The presence of a ligated DNA molecule in the plasmid disrupts the lacZ gene and hence recombinant plasmids produce colourless colonies. Non-recombinant plasmids produce a blue colour as the lacZ gene present in the pBluescript plasmid remains intact.

To produce competent cells for storage at -70°C (for up to 3 months) a few drops of saturated overnight culture were grown in 5ml of L-broth to an absorbance at 600nm of 0.3. The 5ml of culture was then added to 100ml of L-broth and grown to an absorbance at 600nm of 0.48. The culture was cooled on ice, pelleted at 4,000*g* / 5minutes / 4°C and gently resuspended in 40ml of ice cold TfbI [30mM KOAc, 100mM $RbCl_2$, 10mM $CaCl_2$, 50mM $MnCl_2$, 15% (v/v) glycerol; pH / 0.2M acetic acid / 5.8], incubated on ice for 5 minutes, then repelleted. This pellet was resuspended in 4ml of TfbII [10mM MOPS-acid, 75mM $CaCl_2$, 10mM $RbCl_2$, 15% (v/v) glycerol; pH / KOH / 6.5] and 200µl aliquots snap frozen on dry ice. Prior to transformation the aliquot is gently thawed and stored on ice for 10 minutes, then transformed as detailed above.

4.2 MOLECULAR CLONING AND MANIPULATION OF DNA

For the precipitation of nucleic acid, in the absence of any salt in the solution, NaOAc (pH5.5) was added to a final concentration of 0.3M. The DNA or RNA was precipitated by the addition of an equal volume of isopropanol or 2X the volume of ethanol, at -20°C or on dry ice for fast precipitation. All collected precipitates were washed with 70% ethanol to remove excess salts and DNA pellets dried under vacuum, before resuspending in the appropriate volume.

4.2.1 Small scale preparation of plasmid DNA

One method was generally used (modified from: Birnboim and Doly, 1979). From overnight, saturated bacterial cultures, 1.5ml of each was microfuged in Eppendorf tubes for 2 minutes. The pellets were resupended by vortexing in 100µl of solution I [25mM Tris-HCl pH8, 10mM EDTA-NaOH pH7.5, 50mM glucose, 1mg/ml lysozyme (freshly added)]. 200µl of solution II [200mM NaOH, 1% (w/v) SDS (freshly made)] was then added to each, mixed gently by inverting the tube several times and the samples immediately placed on ice for 5 minutes. Following lysis, the samples were neutralised by the addition of 150µl solution III [5M KOAc pH4.8] and incubated on ice for 5 minutes. The flocculated mixtures were microfuged for 5 minutes and the supernatants transferred to new Eppendorf tubes. Each sample then underwent a protein extraction, by the addition of an equal volume of 2FC, vortexing, microfuging for 10 minutes and transfer of the overlying aqueous phase to new tubes, followed by an extraction using chloroform as the organic phase solvent. The plasmids were precipitated at room temperature for 2 minutes with the addition of 700µl of ethanol. The precipitates were collected by microfuging for 15 minutes. Each pellet was washed with 70% ethanol and microfuged for 10 minutes. The precipitates were air dried at room temperature and resuspended in 50ml of TE, sufficient for 5-6 enzyme digests, with the simultaneous addition of RNase A to 50μg/ml .

4.2.2 Large scale preparation of plasmid DNA

Relatively large quantities of very pure plasmid DNA (100μg) were prepared using Qiagen Columns (following the Manufacturers instructions) as this did not necessitate ultracentrifuging and was less time consuming. When very large quantities of highly pure plasmid DNA were required then plasmid DNA was prepared by caesium chloride continuous gradient centrtifugation (described by Maniatis *et al.*, 1982). A 30ml culture of the *E.coli* carrying the plasmid of interest was grown to an OD at 600nm of 0.6 (late log phase). 25ml of this culture was then used to innocculate 500ml of L-broth in a 2L flask. The culture was incubated for 2.5 hours at 37°C to an OD at 600nm of 0.4. The bacteria were pelleted at 4,000*g* / 15 minutes / 4°C. The cells were lysed by a scaled up version of the alkaline lysis detailed above.

Briefly the bacterial pellet was resuspended in 18ml of solution I, and 40ml of freshly prepared solution II added to this. After mixing the contents by gently inverting the tube several times the solution was stored at room temperature for 10 minutes. 20ml of ice cold solution III was added, the contents mixed by shaking and the tube stored on ice for 10 minutes. The bacterial lysate was centrifuged at 5,000*g* / 15 minutes / 4°C to pellet the debris. The supernatant was carefully transfered to a 250ml bottle and the plasmid DNA precipitated with the addition of 0.6 volume of isopropanol and incubation for 10 minutes at room temperature. The plasmid DNA was collected by centrifugation at 5,000*g* / 15minutes / 4°C, washed with 70% ethanol and air dried. The pellet was resuspended in 3ml of TE (pH8).

The plasmid DNA was purified by equilibrium centrifugation in a continuous caesium chloride (CsCl)-ethidium bromide gradient. For every millilitre of DNA solution 1g of solid CsCl was added and dissolved. 0.8ml of ethidium bromide (10mg/ml in water) was added for every 10ml of the DNA / CsCl solution and immediately mixed. The density of the resulting solution should be 1.55g/ml and was checked by removing 1ml of solution and weighing it. The solution was transfered to a Beckman Quick-Seal tube suitable for centrifugation in a Beckman vertical Ti65 rotor. The remainder of the tube was filled with paraffin oil. The density gradients were then centrifuged at 45,000rpm for 16 hours.

Two bands were routinely observed in the tube after centrifugation. The upper band consisted of linear bacterial chromosomal DNA and nicked circular plasmid DNA, the lower band consisted of closed circular plasmid DNA. The lower band was collected with a hypodermic needle and the ethidium bromide extracted with 1-butanol. The DNA was pelleted by adding 3 volumes of water to the DNA solution followed by 2 volumes of ethanol (this prevents the co-precipitation of CsCl), incubation at 4°C for 15 minutes and centrifugation at 10,000*g* / 10 minutes / 4°C. The pelleted DNA was washed with 70% ethanol, resuspended in 1ml of TE (pH8), absorbance measured at 260nm and stored in aliquots (in 2 volumes of ethanol at -20°C).

4.2.3 Small scale preparation of bacteriophage λ DNA

Phage plaques required for DNA isolation, were picked in agar plugs (from overnight, bacterial lawn infections grown on BBL, 0.7% agarose) and placed (separately) into 1ml SM buffer in 1.5ml Eppendorf tubes. These were stored at 4°C overnight, with shaking to facilitate phage release. Three dilutions of each phage suspension, (1, 10^{-1}, 10^{-2} in 1ml SM buffer) were incubated with 200µl of saturated *E. coli* XL1-Blue MR culture, for 30 minutes at 37°C and then transfered to 50ml of LB medium (0.2% maltose, 10mM $MgSO_4$) for overnight growth at 37°C. The dilution from each phage which produced the best bacterial lysis was used for DNA preparation. Alternatively, a phage plate stock in SM buffer was used.

The cultures for DNA preparation were cleared of intact bacterial cells, by the addition of a few drops of chloroform and vigorous shaking for 15 minutes. 40ml of each lysate was placed in a corex tube and centrifuged at 10,000*g* / 4°C / 10 minutes. The supernatant was transfered to a fresh 40ml tube and incubated with RNase A (1µg/ml) and DNase 1 (20µg/ml) at 37°C for 30 minutes. The phage particles were pelleted by centrifugation at 45,000*g*/ 4°C / 2 hours and resuspended in 0.5ml of SM buffer. The phage were lysed by the addition of 5µl 10% SDS and 25µl 0.5M EDTA-NaOH pH8, and incubation at 68°C for 15 minutes. Phage proteins and any residual bacterial proteins were extracted by addition of an equal volume of 2FC and subsequently extracted with an equal volume of chloroform. The phage DNA was precipitated with 2X volumes of ethanol, washed with 70% ethanol several times and resuspended in 50µl TE.

Large scale preparations of bacteriophage λ FIXII (Stratagene) were attempted using the protocols described by Maniatis *et al.,* (1982) and those found in Current Protocols Vols 1 & 2 (1992).

4.2.4 Preparation of high molecular weight DNA

High molecular weight mouse DNA was prepared from adult mouse brains using a variation of the procedure described by Blattner *et al.* (1978). Mouse brains (either fresh or stored at -70°C) were placed in a 2ml Nunc screw cap tube, chopped into smaller pieces using a scalpel blade and 1ml of Tail solution added [50mM Tris (pH8.0), 100mM EDTA (pH8.0), 100mM NaCl, 1% SDS]. The brain tissue in tail solution was incubated at 55°C overnight with gentle shaking. The lysed tissue cells were protein extracted twice with 2FC and then with chloroform. The DNA was collected by precipitation with 1X volume ethanol, washed several times with 70% ethanol and pelleted by centrifugation in a microfuge 5, 000*g* / 4°C / 2 minutes. The DNA pellet was air dried and resuspended in 315µl of TE and stored at 4°C.

4.2.5 DNA quantition, restriction and electrophoresis

DNA was quantitated by determining the absorbance, of dilutions from stocks, at 260nm (1=> 50mg/ml DNA). Digestion buffers and temperatures for DNA restriction endonucleases were formulated as recommended by the manufacturers.

When cleaving small quantities of cloned DNAs, the typical total used was 20μl, using 3 units of enzyme/μg DNA. To avoid shearing of cellular DNAs, the buffers and enzymes were initially mixed in gently and the samples incubated at the required temperature (37°C for most enzymes) for 3-16 hours. Following which, more enzyme was added, vigorously mixed and incubated for another 2 hours. 20-30μl of digested DNA samples were routinely used for agarose gel electrophoresis.

300ml (maxi) or 100ml (midi) agarose gels were made in the range of 0.3-2.5% agarose (depending on the size of the fragments under analysis). The gels were run in horizontal electrophoresis chambers in either 1X TAE or 1X TBE running buffer. The running buffer chosen depended on the size of the fragments to be electrophoresed, 1X TAE was routinely used with lager fragments (> 1kb), while 1X TBE was used with smaller fragments (< 1kb). DNA was visualised under long wave (280nm) ultraviolet light after staining the gels with ethidium bromide (0.5μg/ml). Samples were loaded following the addition of 0.1X the volume of loading buffer [50mM EDTA-NaOH pH7.6, 20% (w/v) ficoll, 0.05% (w/v) bromophenol blue, 0.05% (w/v) xylene cyanoll]. The gels were run between 20-150 Volts for 1-16 hours.

4.2.6 Isolation of specific DNA fragments and ligation

Specific DNA fragments were isolated in one of two ways. [A] The restricted samples were run in agarose gels made with low melting point agarose in TBE. As this does not give very good DNA band resolution, it was only used for non-complex DNA digestion products. The DNA fragment is excised in as small a volume of agarose as is possible, melted at 70°C and the desired proportion used directly for ligation. Other DNAs for the ligation were added at 70°C and on cooling to 37°C (at which temperature the agarose has still not set), the ligation buffers and enzyme were added and mixed quickly. Contrary to interfering with the ligation reaction, the solidified agarose acts to reduce the effective reaction volume and enhances the speed and frequency of ligations.

[B] An efficient method for purifying DNA fragments was achieved with the use of DNA binding DEAE paper (NA45). A small piece of the paper was inserted into a cut in the gel, made in front of the desired DNA fragment. Another was inserted behind it, to prevent contamination with more slowly migrating DNA fragments. Continued running of the gel electrophoresed the DNA onto the paper (allowed binding, as monitorred under UV light). The paper was rinsed in TE and the DNA eluted with two 15 minute incubations in 250ml 1M NaCl at 70°C. The eluates were pooled and extracted with 2FC and chloroform and the DNA, ethanol precipitated and dissolved in TE.

Ligations of more than one DNA species were performed in small volumes (10μl for a usual total DNA content of 1μg), with larger volumes used for the simple circular closing of DNA linears. The DNA ratios generally used were1:10 for plasmid vector:insert and 2:1 for λ DNA arms:genomic DNA inserts. Restricted vector DNAs were generally treated with alkaline phosphatase (to prevent self ligation) prior to the ligation reaction with insert DNA. The samples were incubated at 12°C in ligation buffer [10mM Tris-HCl pH7.5, 10mM $MgCl_2$, 6mM KCl, 1mM dithiothreitol (DTT) and ligation additions [1mM spermidine (plus Tris-acid to pH7), 1mM ATP (plus NH_4OH to pH7), 10mM DTT 0.1mg/ml BSA], with the quantity of T4 ligase recommended by the manufacturer (Promega).

PCR products were cloned into the TA cloning vector (Invitrogen), this vector takes takes advantage of the non-template-dependent activity of the thermostable polymerases used in PCR that add a single deoxyadenosine to the 3' ends of duplex molecules. The 3' overhangs are used to insert the PCR product into the vector which contains single 3' T-overhangs at its insertion site. The protocols supplied by the manufaturer for the TA vector were followed for cloning PCR products.

4.2.7 Southern blotting of DNA and specific fragment detection

DNA in agarose gels was denatured in 500ml of denaturing buffer [1.5M NaCl, 0.5M NaOH] until the bromophenol blue marker band appeared green (approximately 45 minutes). The gels were then washed in 0.5X TAE for electroblotting for 45 minutes. The gels were electroblotted onto Nylon membranes in 0.5X TAE using a constant voltage electroblotter (Hoeffer) at 40 volts for 3-4 hours. The nylon filters were then baked at 80°C for 2 hours and then UV crosslinked for 90 seconds.

The filters were prehybridised within pyrex hybridisation tubes in a Hybaid Hybridisation Oven in 1X Church buffer [7% (w/v) SDS, 1% (w/v) BSA, 1mM EDTA, 0.25M $Na_2HPO_4{\cdot}H_2O$ (pH7.2)] at 68°C for 3-16 hours. Hybridisations were performed in 1X Church buffer (generally 8-10ml per filter) at the required temperature (50°C-68°C) for 16-20 hours. Random primed probes were used at a final concentration of 10^7 d.p.m./ml. The filters were washed twice for 20 minutes at room temperature in a low stringency wash [1X SSC, 0.1% (w/v) SDS] and then at 68°C for 1 hour in a high stringency wash [0.1X SSC, 0.1 (w/v) SDS]. Filters were exposed to Kodak XAR5 film at -70°C, using intensifying screens.

4.2.8 Preparation of ^{32}P labelled DNA probes

Hybridisation probes for all the blot analyses (nylon and nitrocellulose) were labelled by the random oligonucleotide-primed method (detailed in Current Protocols Vol 1) using the Klenow fragment of *Escherichia coli* DNA Polymerase I. Either 50μCi of (α-^{32}P)-dCTP or (α-^{32}P)-dATP (3000Ci/mmol) were routinely used to label 50ng of probe material. Specific activities of 5 X 10^8d.p.m. (per μg DNA) were routinely obtained. Unincorporated nucleotides were removed over G75 Sephadex, disposable plastic columns, using 1X STE buffer [10mM Tris.HCL, pH7.5, 10mM NaCl, 1mM EDTA]. The second column peak (representing the probe DNA) was collected and used directly after boiling for 10 minutes, or stored at -20°C.

Probes made from DNA that contained repetitive sequences resulted in cross-hybridisation of signal on the blots. To overcome this cross-hybridisation the repetitive sequences were competed out using cold DNA. This method for 'quashing' repetitive sequence was obtained from George Starks lab ICRF, London. DNA probes were generated via the random oligonucleotide-primed method, the volume of the probe was adjusted to 100μl TE, then 100-200μg of competitor DNA (same species as probe, sonicated to approximately 200-500bp, free of RNA) was added. The volume was adjusted to 176μl with TE pH8 and then boiled for 10 minutes. Immediately 24μl of 1M sodium phosphate was added and the boiled probe placed in a pre-warmed container (lead pot) in a 65°C oven for 2 hours. After the 2 hours the probe was used directly for hybridisation, as before.

4.2.9 cDNA probes for Reverse Northern analysis

The generation of cDNA probes for use in Reverse Northern analysis is basically first strand cDNA synthesis. In this protocol the first strand cDNA synthesis reaction is catalysed by SUPERSCRIPT II RNase H^- Reverse Transcriptase $(RT)^{TM}$ obtained from Gibco BRL. This enzyme has been engineered to eliminate the RNase H activity that degrades the mRNA during the first strand reaction. SUPERSCRIPT II RT^{TM} is not inhibited significantly by ribosomal and transfer RNA and can be used effectively to synthesize first strand cDNA from a total RNA preparation.

10ng of oligo dT was added to 5μg of total cellular RNA in a final volume of 5μl. This was heated at 70°C for 5 minutes and then chilled on ice. 4μl of 5X synthesis buffer [200mM Tris-HCl (pH8.4), 500mM KCl, 25mM $MgCl_2$, 1μg/μl BSA], 2μl of 0.1M DTT, 1μl of dNTPs [10mM dATP, 10mM dGTP, 10mM dTTP], 7μl of $\alpha^{32}P$ dCTP (800Ci/mMol) and 1μl (200 units) SUPERSCRIPT II RT were then added and the reaction mix incubated at 42°C for 90 minutes. This reaction was chased by the addition of 4μl of the follollowing premix [2μl 5X synthesis buffer, 2.5μl 10mM dCTP cold nucleotide, 4.5μl RNase free water, 1μl (200 units) SUPERSCRIPT II RT] and a further incubation at 45°C for 90 minutes. The reaction was terminated by incubation at 70°C for 15 minutes. To remove the RNA template 1μl (2 units) of RNase H was added and incubated at 37°C for 20 minutes. The probes were seperated over G75 Sephadex columns. As the probes are single stranded there is no need to boil the probes prior to hybridisation.

4.2.10 End labelling and hybridisation of oligonucleotide probes

Oligonucleotide primers, to the 5' and 3' ends of trapped-exons that had been designed for PCR, were radio-labelled and used as probes on Southern and Northern blots in an attempt to detect if their sequences were present in mouse genomic DNA and RNA transcripts from mouse embryos. A polynucleotide kinase (PNK) labelling reaction was set up using fresh [γ^{32}P]ATP to label the oligonucleotides. 10pmoles of oligonucleotide were incubated in 1X T4 polynucleotide kinase buffer [50mM Tris-HCl (pH7.5), 1mM $MgCl_2$, 100μM DTT] for 30 minutes at 37°C. Labelling reactions were incubated at 65°C to inactivate the kinase and unincorporated counts removed by passing the labelling reactions over a G-50 column. Filters to be probed with radio-labelled oligonucleotides were prehybridised at 65°C [prehybridisation buffer; 6X SSC, 20mM NaH_2PO_4, 0.4% SDS, 5X Denhardts and denatured, sonicated salmon sperm DNA (500μg/ml)]. The labelled oligonucleotides were hybridised overnight at 5-10°C below the T_m of the oligos in hybridisation buffer [6X SSC, 20mM NaH_2PO4, 0.4% SDS, denatured, sonicated salmon sperm DNA, boiled prior to addition to hybridisation buffer (500μg/ml)]. The filters were washed in 6X SSC, 0.1% SDS at T_m -10°C until the background had diminished. Filters were then exposed to Kodak XAR film overnight at -70°C, with an intensifying screen.

4.2.11 Screening of libraries, clone purification and subcloning

A λ Fix II Library was purchased from Stratagene. It had been constructed from size selected cellular DNA from the spleen of an adult female C57Black/6 mouse. The recombinant DNA library was titrated on *E.coli* XL1-Blue MRA (P2). Four bioassay plates were plated with 20,000 recombinants per plate (in E.coli XL1-Blue MRA (P2)). Duplicate phage DNA lifts with nitrocellulose were taken from each plate. The filters were soaked in denaturing solution [1.5M NaCl, 0.5M NaOH] for 2 minutes and then in neutralising solution [1.5M NaCl, 0.5M Tris-HCl pH 8.0] for 5 minutes, then soaked in rinse [2X SSC, 0.2M Tris-HCl pH7.5] for 30 seconds and air dried. Following baking (80°C for 2 hours under vacuum), these filters were prehybridised, hybridised and washed as described for Southern blots. Positive plaques were picked, plated and rescreened at least twice before DNA was prepared from the phage for subcloning and subsequent analysis.

DNA was prepared from positive phage and used for restriction mapping of the isolated DNA. The DNA was then digested with *EcoR I* and the *EcoR I* fragments were ligated into the *EcoR I* site of pBluescript. These plasmids were then used to prepare large quantities of the isolated DNA for subsequent analysis as described in Results section 2.

This procedure of screening was repeated for the cDNA and *Sau3A I* size selected genomic DNA libraries also used during the project. The construction of these libraries is detailed below.

4.3 ISOLATION AND CHARACTERISATION OF CELLULAR RNA

To avoid the introduction of ribonucleases, during the isolation and manipulation of RNA, all solutions and equipment which came in direct contact with the RNA, was autoclaved prior to use. All reagents were kept RNase free during all of the procedures.

4.3.1 Large scale isolation of total cellular RNA

Tissue samples were obtained fresh or from -70°C and placed in solution D [0.36ml 2X β-Mercaptoethanol 50ml guanidinum thiocyanate stock], 1ml solution D/100mg of tissue, and homogenised using a polytron. A 0.1 volume of 2M NaOAc, pH4.0, was added to the homogenised tissue and mixed vigorously. An equal volume of water saturated pure phenol was then added and mixed. To this was then added a 0.2 volume of chloroform:isoamyl alcohol (49:1). This was vortexed for 15 seconds and placed on ice for 15 minutes. RNA was seperated by centrifugation at 10,000*g* / 20 minutes / 4°C in a Beckman JA20 rotor. The aqueous phase was transfered to a fresh tube and the RNA was precipitated by the addition of an equal volume of isopropanol and incubation at -20°C for at least 1 hour. The RNA was collected by centrifugation in a Beckman JA20 rotor at 10,000*g* / 20 minutes / 4°C. The pellet was aspirated, resuspended in 300μl of solution D and transfered to an Eppendorff tube. The RNA was precipitated with 2X volumes of ethanol at -20°C for at least 1 hour, spun at 14,000*g* in a microfuge, washed twice with 75% ethanol, air dried and resuspended in an appropriate volume of TE buffer containing 0.1% SDS (w/v) before storage at -20°C. An absorbance of 1 at 260nm, for RNA in solution, is equivalent to 40μg/ml RNA.

4.3.2 Selection of polyadenylated RNA

Polyadenylated (poly[A]+) RNA was purified by chromotography on oligo dT cellulose, by the manufacturers recommended procedure. This involves the binding of the poly[A]+ fraction of total RNA to a 0.5ml (dry volume) cellulose column, in a high salt (500mM NaCl) buffer (flow through being the poly[A]- fraction), washing at mid-salt concentrations (100mM NaCl) and poly[A]+ RNA elution in a salt free buffer (all buffers containing 0.1% (w/v) SDS). The poly[A]+ and poly[A]- fractions were ethanol precipitated, quantitated and stored as given before. In general 100µg of poly[A]+ RNA was obtained per mg of total cellular RNA.

4.3.3 Formaldehyde agarose gel electrophoresis of RNA, Northern blotting and detection of specific RNAs

The procedure was modified from that presented by Maniatis *et al.* (1982). For this procedure 100ml or 300ml (final volume) gels were made, containing 1% (w/v) agarose in MOPS electrophoresis buffer (MOPS-E). The agarose was melted in the buffer and cooled to 60°C, before the addition of 17.8% (v/v) formaldehyde and the addition of water to the final volume for pouring. 10µg of RNA from each sample, was ethanol precipitated and washed with 75% ethanol. This was then dried under vacuum and the RNA dissolved in 4µl of water, to this was added 16µl of loading buffer [1X MOPS-E, 50% (v/v) formamide, 17.8% (v/v) stock formaldehyde solution] with an addition of 2µl per sample of ficoll loading buffer. Each gel also included an RNA size ladder (GibcoBRL). Gels were electrophoresed in 1X MOPS-E at 50mA until the bromophenol-blue had run 2/3 of the way to the end. The gels were run in continuously circulated running buffer.

After electrophoresis the gels were rinsed three times in 500ml of 0.5X TAE (each wash lasting 15 minutes), the last wash contained 50µg of ethidium bromide (15mg/ml) to visualise the RNA size markers and the ribosomal RNA bands. The

RNA was electroblotted on to nylon membranes at 40 Volts for 3-4 hours, the filters were then baked at 80°C for 2 hours and then UV crosslinked. Prehybridisation, probe hybridisation and filter washing was performed in exactly the same way as for Southern blotting and to the same stringency.

4.4 EXON-TRAPPING PROTOCOLS

4.4.1 Subcloning of isolated DNA:

Isolated phage DNA was digested with *EcoR I*. The pSPL3 vector was also digested with *EcoR I* and calf intestinal alkaline phosphatase added to the linearized vector after the enzyme restriction reaction. The digested phage DNA and vector DNA were purified by one phenol:chloroform:isoamyl alcohol (25:24:1) extraction. The aqueous phase was transferred to a separate microcentrifuge tube and precipitated with 0.5 volumes of 7.5 M ammonium acetate and 2 volumes of absolute ethanol. This was mixed and held at -70°C for at least 30 minutes. The DNA was pelleted by centrifugation at 14,000*g* for 15 minutes at 4°C, washed with 70% ethanol and dried at 37°C for at least 10 minutes. The DNA was resuspended in TE buffer to a concentration of 0.25μg/μl.

0.5μg of *EcoR I* restriction digested phage DNA was ligated to the *EcoR I* site of the pSPL3 vector (0.25μg) by incubation with 1 unit of T4 DNA ligase in 1X ligase buffer at 14°C for at least 16 hours. This ligation mixture was used to transform XL1-Blue *E.coli* made competent by the calcium chloride method. 10% of the transformation was plated on medium containing 100μg/μl of carbenicillin. The remainder of the transformation was used to inoculate a 10X volume of selective liquid medium.

The colonies from the platings were counted and if the non-recombinant frequency was less than 10% plasmid DNA was isolated from the liquid culture. Individual colonies were screened for inserts by inoculating 3ml of liquid medium with individual colonies and preparing plasmid DNA using the alkaline lysis mini-prep procedure. The mini-prep DNA was analysed by restriction digestion with EcoR I and agarose gel electrophoresis of products. Plasmid DNA was prepared from the liquid culture using a Qiagen column.

4.4.2 Transfection of COS-7 cells

COS-7 cells are African-green-monkey kidney cells that have been transformed by a derivative of SV40 and were obtained from the European Cell Culture. These lines were maintained in D-MEM supplemented with 10% fetal bovine serum (FBS), glutamine, nonessential amino acids, penicillin, and streptomycin (for the recipe see Solutions). The recombinant plasmid DNA was used to transfect these COS-7 cells using the transfection reagent DOTAP (Boehringer Mannheim).

The COS-7 cells were passaged one day prior to transfection by placing 4 x 10^5 cells into 2ml of supplemented D-MEM in 3.5-cm, 6-well dishes. Transfections were performed the following day when the cells were 60-80% confluent. For each transfection 20μl of the transfection-reagent (1mg/ml) was diluted to 70μl with HBS [Hepes 20mmol/l, NaCl 150mmol/l (pH7.4)] in a polystyrene reaction vial. 1μg of plasmid DNA was diluted to 70μl in HBS in a separate polystyrene vial. The two solutions were combined, mixed gently and held at room temperature for 10 minutes. While this incubation was proceeding the COS-7 cells were prepared for transfection. The supplemented D-MEM was removed from the cells and 2ml of serum-free D-MEM medium added to each well. After the 10 minute incubation 0.8ml of serum-free D-MEM medium was added to the DNA/transfection-reagent mixture. The serum-free D-MEM was removed from each well and the combined DNA/transfection-reagent added to each well. The cells were incubated with the lipid-DNA complexes at 37°C in a 5% CO_2 incubator overnight (never for more than 24 hours). The DNA/transfection-reagent/medium mixture was removed by aspiration the following morning and replaced with 2ml of supplemented D-MEM.

The cells were incubated for a further 24 hours at 37°C/5% CO_2 before proceeding with RNA isolation.

4.4.3 Isolation of Total RNA

The medium was aspirated from the wells and the cells were lysed in the wells by the addition of 1ml of RNAzol^{TM} B (Biogenesis Ltd) to each well. The cell lysate was passed through a pipette several times before being transfered to a RNase free Eppendorf tube. RNA was extracted by the addition of 0.1ml of chloroform per 1ml of cell lysate, vortexed vigorously and then placed on ice for 5 minutes. The suspension was centrifuged at 12,000*g*/15 minutes/4°C, the upper aqueous phase containing the RNA was transferred to a fresh tube. An equal volume of isopropanol was added and the samples stored at 4°C for 15 minutes, the samples were then centrifuged at 12,000*g*/10 minutes/4°C. The RNA pellets were washed with 1ml of 75% ethanol vortexed vigorously and centrifuged at 7,500*g*/5 minutes/4°C. The pellets were air dried and resuspended in 50μl of RNase free water, ready for use in cDNA synthesis.

4.4.4 cDNA synthesis

The cDNA synthesis reaction is catalyzed by SUPERSCRIPT II RNase H$^-$ Reverse Transcriptase (Gibco BRL). 3μg of RNA isolated from the transfected cells was resuspended in a total volume of 12μl of RNase free water, to this was added 1μg of SA2 oligonucleotide. This mixture was incubated at 70°C for 10 minutes, chilled on ice for 1 minute and collected at the bottom of the tube by a brief centrifugation. To this mixture 4μl of 5X first strand buffer (250mM Tris-HCl (pH8.3), 375mM KCl, 15mM $MgCl_2$) was added, 2μl of 0.1M DTT and 1μl of 10mM dNTP mix (10mM dATP, dCTP, dGTP and dTTP (pH 7.0)). This was mixed gently and collected by a brief centrifugation. After an incubation at 42°C for 2 minutes, 200 units of Reverse Transcriptase was added and the solution incubated at 42°C for 30 minutes. This reaction was then transferred to a water bath at 55°C and incubated for 5 minutes before adding 2 Units of RNase H, the RNase H digestion was allowed to continue at 55°C. After a 10 minute incubation the reverse transcription reaction was collected by a brief centrifugation and placed on ice.

4.4.5 Primary PCR

The single stranded cDNA product was converted to double stranded DNA. To a thin walled PCR tube 8µl of the reverse transcription reaction was added, to this was added 4µl of 10X PCR buffer (KCL 500mM, Tris-base 100mM, $MgCL_2{\cdot}6H_2O$ 15mM (pH8.4)). 0.8µl of 10mM dNTP was added to the reaction followed by 20µM of each of the primary PCR primers SA2 and SD6, the total volume was brought to 39.5µl by the addition of 22.7 µl of water. The reaction mixture was incubated at 95°C for 5 minutes before the addition of 2.5 units of *Taq* DNA polymerase. 6 cycles of PCR amplification were then carried out, each PCR cycle consisted of 1 minute at 94°C, 1 minute at 60°C and 5 minutes at 70°C. After completion of the PCR cycles the reaction was incubated at 70°C for an additional 10 minutes and then held at 55°C. A 20µl aliquot of the reaction mixture was removed for analysis by agarose gel electrophoresis. To the remaining primary PCR reaction mixture 10 units of *Bst XI* was added and the restriction digest incubated overnight at 55°C. To ensure that digestion had gone to completion 4 units of *Bst XI* were added the following morning and the reaction incubated at 55°C for a further 2 hours.

4.4.6 Secondary PCR

5µl of the digested primary PCR reaction was used for the second PCR amplification. This PCR amplification used the secondary primers, forward primer dUSD2;

5'(CTA)$_4$GTGAACTGCACTGTGACAAGCTGCACGTGCTCTAGAGTCGACCC AGCA-3' and reverse primer dUSA4;

5'(CTA)$_4$CACCTGAGGAGTGAATTGGTCGAAATGGATCTGTCTCTGTCTCTC TCTCCACCTTCTTCTTCTATTCCTTCCTTCGGGCCTGTCGGGTCCCCTCGGG AGATCTCCAGG-3'

The secondary PCR reaction consisted of 5μl of *Bst XI* digested primary PCR product, 4.5μl of 10X PCR buffer, 1μl of 10mM dNTPs, 1μl of secondary PCR primers dUSA4 and dUSD2 (20μM) and 38μl of water. The secondary PCR mixture was incubated at 95°C for 5 minutes, after this incubation 2.5 units of Taq polymerase were added to the PCR reaction. 25 amplification cycles were then performed, each cycle comprised 1 minute at 95°C, 1 minute at 60°C and 3 minutes at 72°C. After completion of the amplification cycles the reaction was held at 72°C for 10 minutes. Products were visualised by staining with ethidium bromide after electrophoresis in 2.0% agarose gels. The secondary PCR products were radio labelled by the random priming method and used to probe Southern blots of isolated phage DNA, Southern blots of mouse genomic DNA and Northern blots of mouse embryonic total RNA

4.4.7 Cloning and identification of secondary PCR products

The primers used in the secondary PCR contain uracil residues. Treatment of the secondary PCR products with uracil DNA glycosylase (UDG) removes the uracil residues allowing the products to be subcloned into the UDG cloning vector (Gibco BRL) at high efficiency. The cloning reaction was carried out in a volume of 10μl at 37°C for 30 minutes. The reaction was composed of 100ng of secondary PCR product, 50ng of the UDG cloning vector (already linearised), 1μl of 10X PCR buffer, 1 Unit of UDG and distilled water to 10μl. After incubation at 37°C for 30 minutes, 5μl of the ligation mix was used to transform 100μl of competent *E. coli*, XL1-Blue (prepared by the Calcium Chloride method). The transformed cells were plated on LB plates containing 100μg/ml carbenicillin and incubated at 37°C overnight.

Transformants were selected the following morning and evaluated using colony PCR. A portion of each colony was placed in a tube containing the following PCR mix, 20μM each of secondary amplification primers, 1μl of 10mM dNTP mix, 5μl of 10X PCR mix, 2.5 Units of *Taq* DNA polymerase and the final volume made up to 50μl with distilled water. The PCR reactions were placed in a thermocycler and held at 94°C for 5 minutes, then 30 cycles of PCR amplification were performed, each cycle comprised 45 seconds at 94°C, 30 seconds at 55°C and 1 minute at 72°C. The reactions were then incubated at 72°C for an additional 10 minutes and then held at 4°C. The products of this PCR were visualised by ethidium bromide staining after electrophoresis in 2.0% agarose gels.

4.5 Dideoxy sequencing

In the original dideoxy sequencing procedure, primer extension was catalyzed by the Klenow fragment of *E. coli* DNA polymerase I. T7 DNA polymerase has several advantages over Klenow fragment sequencing:

(i) with it's processtivity and high rate of polymerisation, longer chain-terminated fragments can be generated very rapidily. This allows a greater length of sequence to be determined.

(ii) because of its tolerance for substrate analogues, the same set of sequencing mixes may be used with either ^{32}P or ^{35}S.

The major difference when using T7 DNA polymerase rather than Klenow fragment is that the primer-extension reactions are performed in two stages, a 'labelling' reaction and a 'termination' reaction. The two stages are required because the enzyme uses dideoxy-nucleotides very readily. To permit synthesis of long chain-terminated fragments, dideoxynucleotides are therefore excluded during the first stage, then added for the second.

Dideoxy sequencing was carried out using double-stranded template, the concentration of template was adjusted so that 32μl contained 1.5-2μg of DNA. The double-stranded template was denatured by the addition of 8ml of 2M NaOH, vortexed gently, centrifuged briefly and incubated at room temperature for 10 minutes. The denatured DNA template was precipitated by the addition of 7μl of 3M sodium acetate (pH4.8), 4μl of distilled water and 120μl of ethanol, mixed, and placed on dry ice for 15 minutes. The precipitated DNA was collected by the centrifuging for 15 minutes. The pellet was washed with ice-cold 75% ethanol, recentrifuged for 10 minutes, then dried briefly under vacuum. The DNA pellet was redissolved in 10μl of distilled water.

The concentration of the primers used for the primer extension reaction was adjusted to 2.5-5μM. The primer was annealed to the template DNA by the addition of 2μl of sequencing primer and 2μl of annealing buffer [1M Tris-HCl (pH7.6), 100mM $MgCl_2$ and 160mM DTT] to the resuspended denatured template. The template and primer mixture was mixed gently and incubated at 65°C for 5 minutes.The annealing mixture was placed at 37°C for 10 minutes and then placed at room temperature for a further 10 minutes. To the tube containing the annealed template and primer, 3μl of labelling mix [1.375mM each dCTP, dGTP and dTTP and 333.5mM NaCl], 1μl of [α-35S]dATPαS, 1μl of T7 DNA polymerase, were added and incubated at room temperature for 5 minutes.

While the labelling reaction was in progress the four sequencing mixes were warmed at 37°C for at least 1 minute. After the labelling reaction had been incubated for 5 minutes, 4.5ml of this reaction was transfered into each of the four pre-warmed sequencing mixes. The components were gently mixed and incubated at 37°C for 5 minutes. The sequencing reactions were stopped by the addition of 5ml of stop buffer [0.3% each Bromophenol Blue and Xylene Cyanol FF; 10mM EDTA (pH7.5) and 97.5% (v/v) formamide].

3μl of each sample was run on a 6% polyacrylamide, 7M urea gel that had been pre-run for 2 hours at 50mA in 1X TBE. The samples were boiled, loaded quickly and electrophoresed for the required time (2-6 hours). More than one aliquot from each sample was run on a single gel for varying lengths of time maximising the amount of sequence information. After electrophoresis the gels were dried directly on to Whatmann 3MM paper at 80°C under vacuum for 1 hour. The gels were exposed to Kodak XAR film at room temperature.

4.6 Polymerase Chain Reaction

Oligonucleotide primers were designed to the 5' and 3' ends of sequences that were to be amplified by PCR. They are listed below;

Exon p50; forward primer, 5'-GGAGCTAAAGGGATCTGC-3'
reverse primer, 5'-GAGACACGAGCTCGGGGG-3'
Exon p51; forward primer, 5'-CGAACACGAACTATAACTTTCTCG-3'
reverse primer, 5'-CTCGCCTAGCAATACGGATG-3'
Exon p52; forward primer, 5'-GGGGATCGTTATGCCTAC-3'
reverse primer, 5'-CTCCTTGCATTGAATTTCCTC-3'
Exon p56; forward primer, 5'-CTCCTAAGATGGCAGCAACTC-3'
reverse primer, 5'-CAGATGGGTTCTAGAAGCC-3'
Exon p57; forward primer, 5'-GTGCAGTACTAGCTGCTG-3'
reverse primer, 5'-GGAGGTAGGGTCTTCGGAG-3'

The conditions for the PCR were optimised by varying the Magnesium concentration in the reactions, the temperature of primer annealing, the quantity of primer required and the amount of template required. The T_m was calculated for each primer on the basis of the following equation;

$$T_m = 2^{o}\mathrm{C}\ (\mathrm{A} + \mathrm{T}) + 4^{o}\mathrm{C}\ (\mathrm{G} + \mathrm{C})$$

The basic PCR protocol involved incubation at 94°C for 5 minutes, followed by 30 cycles of a denaturation step of 94°C for 1 minute, an annealling step of 40°C for 1 minute and an extension step of 60°C for 1 minute. This was followed by an incubation at 70°C for 10 minutes. Products of the PCR reaction were analysed by electrophoresis on agarose gels (as detailed earlier).

4.7 RT-PCR

PCR was adapted to detect the presence of RNA transcripts expressed at very low levels and amplify them, allowing them to be subcloned and sequenced. The PCR was carried out on first strand cDNA synthesized from the RNA of choice. In this case day 13.5, day 15.5 and day 16.5 mouse embryonic total RNA was used as the template for reverse transcription.

The synthesis of cDNA was similar to the method given in the exon-trapping section. To ensure that there was no DNA contaminating the RNA, used as template for the cDNA synthesis, amplification grade DNase (RNase free) was used to digest any DNA present. 10 units of DNase were used for every µg of RNA to be amplified and the reaction was carried out in 1X cDNA synthesis buffer at room temperature for 15 minutes prior to cDNA synthesis. The DNase was inactivated by addition of EDTA (final conc 2mM) and incubation at 65°C. This mixture was used directly for cDNA synthesis. 0.5µg of oligo dT_{12-18} was added to 5µg of total RNA in a final volume of 13µl. This was heated at 70°C for 10 minutes and cooled on ice for 1 minute. To this was added 2µl of 10X synthesis buffer [200mM Tris-HCl (pH8.4), 500mM KCl, 25mM $MgCl_2$, 1µg/µl BSA], 1µl of 10mM dNTPs, 2µl of 0.1M DTT and finally 200 units of SUPERSCRIPT II RT^{TM}. The reaction mixture was incubated at room temperature for 10 minutes, transfered to a 42°C water bath and incubated for a further 50 minutes. The reaction was terminated by incubating the reaction at 70°C for 10 minutes. The RNA template was removed by the addition of 2 units of RNase H and an incubation of 20 minutes at 37°C. 2ml of this final reaction mix was used as cDNA template in a PCR reaction. The PCR comprised of the cycles already found to be optimal in the PCR from genomic DNA template (see PCR protocol above). For each RT-PCR a RNA only template control was included to ensure there was no contaminating DNA.

4.8 Size selection of *Sau3A I* restricted cellular DNAs and construction of partial libraries in bacteriophage λ EMBL4

10μg of cellular DNA was mixed with *Sau3A I* restriction buffer in a final volume of 150μl. 30μl was dispensed into one tube and 15μl dispensed into tubes 2-8 and the remainder into tube 9. All the tubes were placed on ice. 4 units of *Sau3A I* was added to tube 1 and mixed. 15μl of the reaction was added to tube 2. The enzyme was mixed well, with the twofold dilution series continued through to tube 8 and nothing added to tube 9. Tubes 1-8 were then incubated at 37°C for 1 hour. The reactions were stopped by heating the restriction digests to 70°C for 10 minutes. The samples were then electrophoresed on a 0.7% agarose gel in 1X TAE and DNA visualised by ethidium bromide staining and exposure to UV light. The amount of enzyme that results in the greatest intensity of fluoresence of DNA in the 9-20kb region was noted. To obtain the maximum number of molecules in this size range, half of the amount of enzyme that produces the maximum amount of fluoresence was used to produce partially digested DNA fragments for ligation into λ EMBL4 arms.

Using the exact same enzyme concentration, time, temperature, and DNA concentration, found to be optimal in producing the greatest number of DNA fragments, ranging in size from 9-20kb, 500μg of high molecular weight cellular DNA from female mouse brain was digested with *Sau3A I*. A small aliquot of this digest was analysed by electrophoresis to ensure that the size distribution of the digestion products was correct. The DNA was gently extracted twice with 2FC, then precipitated with ethanol and then washed with 70% ethanol. The DNA pellet was then resuspended in distilled water to a final concentration of 0.25μg/μl. 10μg of size selected DNA was treated with calf intestinal alkaline phosphatase to prevent smaller DNA molecules forming mulitple inserts during the ligation reaction.

λ EMBL4 arms, pre-digested with *BamH I*, were obtained from Stratagene and used to construct a library with the size selected *Sau3A I* partially digested DNA. 0.3μg of *Sau3A I* digested cellular DNAwas ligated to 1μg of λ EMBL4 / *BamH I* arms in a final volume of 5μl. The ligation was incubated at 4°C overnight.

The ligation products were then packaged using Gigapack II XL Packaging Extract (Stratagene), which preferentially size selects for extra large inserts. For example, a 20kb insert will be packaged with a 95% higher efficiency than a 14kb insert when using a Lambda replacement vector. The Ligated DNA was packaged following exactly the protocol supplied with the packaging extract and the resultant recombinant phage library were titred on XL1-Blue MRA (P2) as detailed earlier.

4.9 cDNA Library construction

Routinely cDNA synthesis was achieved using the SUPERSCRIPT II RT^{TM} (Gibco BRL), however, for the construction of a cDNA library first strand synthesis was catalyzed using the Moloney Murine Leukaemia Virus (MMLV) reverse transcriptase from Pharmacia. It was supplied with TimeSaver^{TM} cDNA Synthesis Kit. Briefly, 1μg of poly[A]$^+$ RNA derived from the total RNA of a embryonic day 16.5 mouse whole body was resuspended in a total volume of 20μl of RNase-free water. This RNA was then heated at 65°C for 10 minutes and then cooled on ice. 1μl of 200mM DTT and 0.5μg of oligo $dT_{12\text{-}18}$ were added to a first-strand reaction mix [MMLV Reverse Transcriptase, RNAguard, RNase / DNase-free BSA, dATP, dCTP, dGTP, dTTP, in aqueous buffer, (the exact composition of the first-strand reaction mix was unknown)], then the heat denatured RNA. After gentle mixing the reaction was incubated at 37°C for 1 hour. The first strand reaction was then added to a second strand reaction mix [*E.coli* RNase H and *E.coli* DNA polymerase I in aqueous buffer, (again, the exact composition of the second strand reaction mix was unknown)], this reaction was then incubated at 12°C for 30 minutes and followed by 22°C for 1 hour. The reaction was heated at 65°C for 10 minutes and then extracted with 100μl of 2FC. The upper aqueous layer was then passed over a Sepharose CL-4B column prepared in ligation buffer [66mM Tris-HCl (pH7.6), 0.1mM spermidine, 6.6mM $MgCl_2$, 10mM DTT, 150mM NaCl] (following the manufacturers instructions), and the column effluent used directly for the addition of *EcoR I / Not I* adaptors.

1μl of *EcoR I / Not I* adaptor solution, 30μl of PEG buffer, 1μl of 15mM ATP, and 2 units of T4 DNA ligase was added to the 100μl of column effluent, mixed gently and then incubated at 16°C for 1 hour. The reaction was heated at 65°C for 10 minutes to inactivate the DNA ligase and then cooled on ice. 1.5μl of 75mM ATP and 1μl of T4 polynucleotide kinase was then added, mixed and incubated at 37°C for 30 minutes. The reaction mixture was heated to 65°C for 10 minutes and then extracted with 140μl 2FC. The aqueous layer was applied to a Sepharose CL-4B column, spun at 400g for 2 minutes, the effluent was used for ligation into prepared λ Excell *EcoR I* / CIP arms (Pharmacia). A range of different volumes of column efffluent were used for test ligations. To a total volume of 20μl (column effluent plus the required amount of ligation buffer plus 2μg of prepared λ arms) 1μl of 3M sodium acetate and 50μl of cold (-20°C) ethanol was added, precipitated at -70°C for 15 minutes. The cDNA and λ arms were pelleted, washed in 70% ethanol and briefly air dried.

The pellet was resuspended in 8μl of ligation buffer, to which was added 1μl of 1 mM ATP solution and 1μl of T4 DNA ligase, this ligation mix was then incubated at 16°C overnight. The ligated DNA was packaged using Gigapack II Gold Packaging extract (Stratagene), this packaging extract gives the highest efficieny of packaging when used to package cDNA molecules (selecting against large DNA molecules thus avoiding the packaging of multiple inserts). The manufacturers protocol was followed exactly and the resultant recombinant cDNA λ library titred on NM522 *E.coli* (Pharmacia) and screened as detailed before.

SOLUTIONS

4.10 Formulation of frequently used solutions

PBS.A	1% (w/v) NaCl
	0.025% (w/v) KCl
	0.14% (w/v) Na_2HPO_4
	0.025% (w/v) KH_2PO_4
PBS	10mM $CaCl_2$
	6mM$MgCl_2$
	in PBS.A
TE	10mM Tris-Base
	1mM EDTA
	pH/HCl/7.6
SM	50mM Tris-HCl pH7.5
	10mM$MgSO_4$
	100mM NaCl
	0.01% gelatin
1X TAE	40mM Tris-acetate
	2mM EDTA
TBE	89mM Tris-Base
	89mM Boric acid
	2mM EDTA

MOPS-E	20mM MOPS-acid
	1mM EDTA
	5mM NaOAc
	pH/ NaOH / 7.0
20X SSC	3M NaCl
	0.3M Na_3citrate.2H_2O
	pH/ HCl / 7.0
2FC	49.5% (v/v) redistilled phenol
	49.5% (v/v) chloroform
	1% (v/v) isoamyl alcohol
	0.1% (w/v) 8 hydroxyquinoline
	saturated with TE:pH8
STE	10mM Tris.Cl, pH7.5
	10mM NaCl
	1mM EDTA
10X PCR BUFFER	
	500mM KCL
	100mMTris-base
	15mM $MgCl_2$.6H_2O
SAMPLE LOADING BUFFER	
	20% (w/v) Ficoll
	0.25% Bromophenol Blue
	0.25% Xylene Cyanol FF

Formulation of frequently used growth media

DMEM+	10% (v/v) FBS
	2mM L-Glutamine
	in Dulbecco's Modified Eagle Medium
L- BROTH	1% (w/v) Bacto-tryptone
	0.5% (w/v) Bacto-yeast extract
	1% (w/v) NaCl
LBMM	0.2% (w/v) maltose
	10mM MgSO4
	in L-broth
2X YT	1.6% (w/v) Bacto-tryptone
	1% (w/v) Bacto-yeast extract
	0.5% (w/v) NaCl

For the bacterial and bacteriophage plating, Bacto-agar was added to the appropriate medium at 1.1% (w/v), or for soft, top agar at 0.6% (w/v). To prevent adhesion of the overlays to nitrocellulose, in plaque-lift assays, the agar was replaced with agarose.

REFERENCES

Adams, M. D., Kerlavage, A. R., Fields, C. and Venter, J. C. 3400 expressed sequence tags identified from the Human brain. Nature Genet. **4**, 256-267 (1993).

Amariglio, N. and Rechavi, G. Insertional mutagenesis by transposable elements in the mammalian genome. Environmental and Molecular Mutagenesis **21**, 212-218 (1993).

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., Struhl, K. Current protocols in molecular biology Vols. 1 & 2. Wiley Interscience (1992).

Bach, O., Agell, N. and Carafoli, E. Calmodulin and calmodulin-binding proteins in the nucleus. Cell Calcium. **16**, 289-296 (1994).

Ballabio, A. The rise and fall of positional cloning? Nature Genet. **3**, 277-279 (1993).

Bennett, M. K., and Kennedy, M. B. Deduced primary structure of the Ca^{2+}/calmodulin-dependent protein kinase determined by molecular cloning Proc. Natl. Acad. Sci. USA. **84**, 1794-1798 (1987).

Bianchi, N. O., Bianchi, M. S., Balliet, G. and De La Chapelle A. Characterisation and sequencing of the sex-determining region-Y gene (SRY) in Akodon (Cricetidae) species with sex reversed females. Chromosoma. **102**, 389-395 (1993).

Birnboim, H. and Doly, J. A rapid alkaline extraction procedure for recombinant plasmid DNA. Nucl. Acids Res. **7**, 1513-1523 (1979).

Blattner, F.R., Blechl, A., Dennis-Thompson, K., Faber, H., Richards, J.E., Slightom, J.L., Tucker, P. and Smithies, O. Cloning human fetal γ globin and mouse α-type globin DNA: Preparation and screening of shotgun collections. Science **202**, 1279-1284 (1978).

Bowes, C., Li, T., Frankel, W.N., Danciger, M., Coffin, J.M., Applebury, M.L. and Faber, D.B. Localization of a retroviral element within the rd gene coding for the β subunit of cGMP phosphodiesterase. Proc.Natl. Acad. Sci. USA **90**, 2955-2959 (1993).

Bremer, E., Middendorf, A., Martinussen, J. and Valentin-Hansen, P. Analysis of*tsx* gene, which encodes a nucleoside specific channel-forming protein (TSX) in the outer membrane of *Escherichia coli*. Gene **96**, 59-65 (1990).

Brown, J.M., Wedden, S.E., Millburn, G.H., Robson, L.G., Hill, R.E., Davidson, D.R. and Tickle, C. Experimental analysis of the control of expression of the homeobox-gene *Msx-1* in the developing limb and face. Development, **119**, 41-48 (1993).

Burton, F. H., Loeb, D. D., Edgell, M. H. and Hutchison, C. A. III. L1 gene conversion or same site transposition. Mol. Biol. Evol. **8**, 609-619 (1991).

Capecchi, M.R. The new mouse genetics: Altering the genome by gene targeting. Trends Genet. **5**, 70-77 (1989).

Charlesworth, B., Sneigowski, P. and Stephan, W. Evolutionary dynamics of repetitive DNA in eukaryotes. Nature **371**, 215-220 (1994).

Chisaka, O. and Capecchi, M.R. Regionally restricted developmental defects resulting from targeted disruption of the homeobox gene *hox-1.5*. Nature **350**, 473-479 (1991).

Church, D.M., Stotler, C.J., Rutter, J.L., Murrel, J.R., Trofatter, J.A. and Buckler, A.J. Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. Nature Genet. **6**, 98-105 (1994).

Collins, F.S. Positional cloning: lets not call it reverse anymore. Nature Genet. **1**, 3-6 (1992).

Copeland, N.G. and Jenkins N.A. Development and applications of a molecular genetic linkage map of the mouse genome. Trends Genet. **7**, 113-118 (1991).

Dawson, C.W., Rickinson, A.B. and Young, L.S. Epstein-Barr virus latent membrane proteins inhibit human epithelial cell differentiation. Nature **344**, 777-781 (1990).

Den Dunnen, J.T. and Schoenmakers, J.G.G. Consensus sequences of the *Rattus norvegius* B1- and B2 repeats. Nucleic Acid Res. **15**, 2772-2772 (1987).

Dente, L., Ruther, U., Tripodi, M., Wagner, E.F. and Cortese, R. Expression of human α 1-acid glycoprotein gene in cultured cell and in transgenic mice. Genes Dev. 2, 259-266 (1988).

Diewert, V.M. A quantitative coronal plane evaluation of cranio-facial growth and spatial relations during secondary palate development in the rat. Archs Oral Biol. **23**, 607-629 (1978).

Dorfman, D.M., Wilson, D.B. Bruns, G.A.P. and Orkin, S.H. Human transcription factor GATA-2, evidence for the regulation of preproendothelin-1 gene expression in endothelial-cells. Journal of Biological Chemistry, **267**, 1279-1285 (1992).

Fahraeus, R., Rymo, L., Rhim, J.S. and Klein, G. Morphological transformation of human keratinocytes expressing the LMP gene of Epstein-Barr virus. Nature **345**, 447-449 (1990).

Ferguson, M.W.J. Palate development: mechanisms and malformations. Irish Journal of Medical Science, **156**, No. 11, 309-315 (1987).

Ferguson, M.W.J. Palate development. Development **103**, Supplement, 41-60 (1988).

Ferguson, M.J.W. Craniofacial malformations: towards a molecular understanding. Nature Genet. **6**, 329-330 (1994).

Ferguson, M.W.J. and Honig, L.S. Epithelial-mesenchymal interactions during vertebrate palatogenesis. In Current Topics in Developmental Biology, Vol **19** Palate Development: Normal and Abnormal, Cellular and Molecular Aspects' (ed. E.F. Zimmerman), pp. 137-164. New York: Academic Press (1984).

Fukamizu, A., Nishi, K., Cho, T., Saitoh, M., Nakayama, K., Ohkubo, H., Nakanishi, S. and Murakami, K. Structure of the rat renin gene. J. Mol. Biol. **201**, 443-450 (1988).

Goodfellow, P.N. and Lovell-Badge, R. SRY and sex determination in mammals. Ann. Rev. Genet. **27**, 71-92 (1993).

Goodfellow, P.N., Hawkins, R.J. and Sinclair, A.H. Cloning the mammalian sex-determining gene, TDF. In Genome analysis Vol. 3., Genes and Phenotypes (Ed. Davies, K.E. and Tilghman, S.M.), pp59-78, Cold Spring Harbor Laboratory Press 1991.

Gorski, S.M., Adams, K.J., Birch, P.H., Freidman, J.M. and Goodfellow, P.J. The gene responsible for X-linked cleft palate (CPX) in a British Columbian native kindred is localised between PGK1 and DXYS1. Am. J. Hum. Genet. **50**, 1129-1136 (1992).

Gubbay, R., Collingham, J., Koopman, P., Capel, B., Economou, A., Muensterberg, A., Vivian, N., Goodfellow, P. and Lovell-Badge, R. A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. Nature. **346**, 245-250 (1990).

Hanson, P.I. and Schulman, H. Neuronal Ca^{2+}/calmodulin-dependent protein kinases. Annu. Rev. Biochem. **61**, 559-601 (1992).

Hayakawa, T., Mizukami, M., Nakamura, I. and Suzuki, M. Cloning and sequencing of RNA-1 cDNA from Cucumber Mosaic Virus strain-O Gene **85**, 533-540 (1989).

Henderson, S., Rowe, M., Gregory, C., Croom-Carter, D., Wang, F., Longnecker, R., Kieff, E. and Rickinson, A. Induction of Bcl-2 expression by Epstein-Barr virus latent membrane protein 1 protects infected B cells from programmed cell death. Cell **65**, 1107-1115 (1991).

Hearn, C.M., McAleer, M.A., Love, J.M., Aitman. T.J., Cornall, R.C., Ghosh, S., Knight, A., Prins, J.B. and Todd, J.A. Further micro satellite markers for mouse genome mapping. Unpublished. Genbank accession number x58264 (Nov. 1994).

Hill, R.E., Jones, P.F., Rees, A.R., Sime, C.M., Justice, M.J., Copeland, N.G., Jenkins, N.A., Graham, E. and Davidson, D.R. A new family of mouse homeo box-containing genes: molecular structure, chromosomal location, and developmental expression of *Hox-7.1*. Gene Dev. **3**, 26-37 (1989).

Holmes, S.E., Dombroski, B.A., Krebs, C.M., Boehm, C.D. and Kazazian, H.H., Jr. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. Nature Genet. **7**, 143-148 (1994).

Ishihara, H., Engel, J.D. and Yamamoto, M. Structure and regulation of the chicken GATA-3 gene. Journal of Biological Chemistry **117**, 499-508 (1995).

Jaenisch, R., Harbers, K., Schnieke, A., Lolher, J., Chumakov, I., Jahner, D. Grotkopp, D. and Hoffmann, E. Germline integration of Maloney murine leukemia-virus at the Mov 13 locus leads to recessive lethal mutation and early embryonic death. Cell **32**, 209-219 (1983).

Jeffrey, A.J. Extended nucleotide sequences, Patent EP0370719-A 73 30-MAY-1990, Imperial Chemical Industries plc. Genbank a08893.gb_pat (Nov. 1994).

Jowett, A.K., Vainio, S., Ferguson, M.W.J., Sharpe, P.T. and Thesleff, I. Epithelial-mesenchymal interactions are required for *msx1* and *msx2* gene expression in the developing murine molar tooth. Development **117**, 461-470 (1993).

Karls, U, Mueller, U., Gilbert, D.J., Copeland, N. G., Jenkins, N. A. and Harbers, K. Structure, expression and chromosome location of the gene for the β subunit of brain-specific Ca2+/calmodulin-dependent protein kinase II identified by transgene integration in an embryonic lethal mouse mutation. Mol. Cell. Biol. 12, 3644-3652 (1992).

Kennedy, M.B. and Greengard, P. Two calcium/calmodulin-dependent protein kinases which are highly concentrated in brain, phosphorylate protein 1 at distinct sites. Proc. Natl. Acad. Sci. USA 88, 2643-2647 (1991).

Kindler, S., Schwamke, B., Schulz, B. and Garner, C. C. Complete cDNA sequence encoding rat high and low molecular weight MAP 2. Nucleic Acid Res. **18**, 2822-2822 (1990).

Kingsmore, S.F., Giros, B., Suh. D., Bieniarz, M., Caron, M.G. and Seldin, M.F. Glycine receptor β-subunit gene mutation in spastic mouse associated with LINE-1 element insertion. Nature Genet. **7**, 136-142 (1994).

Ko, L.J., Yamamoto, M., Leonard, M.W., George, K.M., Ting, P. and Engel, J.D. Murine and human lymphocyte-T GATA-3 factors mediate transcription through a cis regulatory element within the human T-cell receptor delta gene enhancer. Mol. Cell Biol. **11**, 2778-2784 (1991).

Kratochwil, K., Mark, K., Kollar, E.J., Jaenisch, R., Mooslehner, K., Schwarz, M., Haase, K., Gmachl, I. and Harbers, K. Retrovirus-induced insertional mutation in Mov 13 mice affects collagen I expression in a tissue-specific manner. Cell **57,** 807-816 (1989).

Linial, M. and Scheller, R.H. A unique neurofilament from Torpedo electric lobe, sequence, expression and localisation analysis. J. Neurochem. **54,** 762-770 (1990).

Lufkin, T., Dierich, A., LeMeur, M., Mark, M. and Chambon, P. Disruption of the *hox-1.6* homeobox gene results in defects in a region corresponding to its rostal domain of expression. Cell **66,** 1105-1119 (1991).

Lyon, M.F and Searle, A.G. Genetic variants and strains of the laboratory mouse, Second edition Oxford University Press, Oxford, (1989).

Ma, J. -X., Wang, D. -Z., Chao, L. and Chao, J. Unpublished. Molecular cloning sequence analysis and expression of the gene encoding bradykinin B2 receptor. Genbank, accession number L27595 November 1994.

MacKenzie, A., Leeming, G.L., Jowett, A.K., Ferguson, M.W.J. and Sharpe, P.T. The homeobox gene Hox 7.1 has specific regional and temporal expression patterns during early murine craniofacial embryogenesis, especially tooth development *in vivo* and *in vitro*. Development **111,** 269-285 (1991).

MacKenzie, A., Ferguson, M.W.J. and Sharpe, P.T. Expression patterns of the homeobox gene, Hox-8, in the mouse embryo suggest a role in specifying tooth initiation and shape. Development **115,** 403-420 (1992).

Mahadevan, M.S., Amemiya, C., Jansen, G., Sabourin, L., Baird, S., Neville, C.E., Wormskamp, N., Segars, B., Batzer, M., Lamerdin, J., Wieringa, B. and Koneluk, R.G. Structure and genomic sequence of the myotonic dystrophy (DM) kinase gene. Hum. Mol. Genet. **2,** 299-304 (1993).

Maniatis, T., Fritsch, E. and Sambrook, J. Molecular Cloning, a laboratory manual. publ. Cold Spring Harbor, USA. (1982).

Margalit, H., Nadir, E. and Ben-Sasson, S.A. A complete Alu element within the coding sequence of a central gene. Cell. **78,** 173-174 (1994).

Martin, J.M., Veis, D., Korsmeyer, S.J. and Sudgen, B. Latent membrane protein of Epstein-Barr Virus induces cellular phenotypes independently of expression of Bcl-2. J. Virol. **67,** 5269-5278 (1993).

Massague, J. The TGFβ family of growth and differentiation factors. Cell **49**, 437-438 (1987).

McNeish, J., Scott, W. and Potter, S. *Legless*, a mutation in pHT1 line transgenic mice. Science **241**, 837-839 (1988).

Miller, S.G. and Kennedy, M.B. Regulation of brain type II Ca^{2+}/calmodulin dependent protein kinase by autophosphorylation: a Ca^{2+}-triggered molecular switch. Cell **44**, 861-870 (1986).

Moore, G.E., Ivens, A., Chambers, J., Farrall, M., Williamson, R., Page, D.C., Bjornsson, A., Arnason, A. and Jensson, O. Linkage of an X-chromosome cleft palate gene. Nature **326**, 91-92 (1987).

Nairn, A.C. and Greengard, P. Purification and characterisation of Ca^{2+}/calmodulin-dependent protein kinase I from bovine brain. J. Biol. Chem. **262**, 7273-7281 (1987).

Ner, S.S. HMGs everywhere. Curr. Biol. **2**, 208-209 (1992).

Nghiem, P., Saati, M. S., Martens, C.L., Gardner, P. and Schulman, H. Cloning and analysis of two new isoforms of multifunctional Ca^{2+}/calmodulin-dependent protein kinase. J.Biol. Chem. **268**, 5471-5479 (1993).

Ohmstede, C.A., Bland, M.M., Merill, B.M., Sahyoun, N. Relationship of genes encoding Ca^{2+}/calmodulin-dependent protein kinase Gr and calspermin: A gene within a gene. Proc. Natl. Acad. Sci. USA. **88,** 5784-88 (1991).

O'Neill, T.K. and DeGrado, W.F. How calmodulin binds its targets: sequence independent recognition of amphiphilic α-helices. TIBS. **15**, 59-64 (1990).

Pailhouse, E., Popeseu, P., Parma, P., Bosher, J., Legault, C., Molteri, L., Fellous, M.and Cotinot, C. Screening for Y-specific DNA sequences, including SRY, in intersex pigs. Unpublished. Genbank, accession number Z23070 (November 1994).

Pohl, T.M., Mattei, M.G. and Ruther, U. Evidence for allelism of the recessive insertional mutation *add* and the dominant mouse mutation extra-toes (Xt). Development **110**, 1153-1157 (1990).

Reith, D.A. and Bernstein, A. Molecular basis of mouse developmental mutants. Gen Dev. **5**, 1115-1123 (1991).

Rijkers, T., Peetz, A. and Ruther, U. Insertional mutagenesis in transgenic mice. Transgenic Research. **3**, 203-215 (1994).

Rinchik, E.M. Chemical mutagenesis and fine structure functional analysis of the mouse genome. Trends Genet. **7**, 15-21 (1991).

Robert, B., Lyons, G., Simandl, B.K., Kuroiwa, A. and Buckingham, M. The apical ectodermal ridge regulates *Hox-7* and *Hox-8* gene expression in developing chick limb buds. Genes Dev. **5**, 2363-2374 (1991).

Rollnick, B.R. and Kaye, C.I. Mendelian inheritance of isolated nonsyndromic cleft palate. Am. J. Med. Genet. **24**, 465-473 (1986).

Rudnicki, M.A. and Jaenisch, R. Insertional mutagenesis. In Genome analysis, Vol 2 Gene expression and its control (Davies, K.E. and Tilghman, S.M.) pp. 13-39 (1991).

Sahyoun, N., McDonald, O.B., Farrel, F. and Lapetina, E.G. Phosphorylation of a Ras-related GTP-binding protein, Rap-1b, by a neuronal Ca^{2+}/calmodulin-dependent protein kinase, CaM kinase Gr. Proc. Natl. Acad. Sci. USA. **88**, 263-2647 (1991).

Sanger, F., Nicklen, S. and Coulson, A.R. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA. **74**, 5463-5476 (1977)

Sasaki, T. and Minobe, Y. Unpublished. Rice cDNA partial sequence. Genbank, accession number D15439 (November 1994).

Satokata, I and Maas, R. *Msx1* deficient mice exhibit cleft palate and abnormalities of craniofacial and tooth development. Nature Genet. **6**, 348-356 (1994).

Schulman, H. and Greengard P. Ca^{2+}-dependent protein phosphorylation system in membranes from various tissues and its activation by calmodulin (calcium-dependent regulation). Proc.Natl. Acad. Sci. USA. **75**, 5432-5436 (1978).

Schulman, H. and Lou, L.L. Multifunctional Ca^{2+}/calmodulin-dependent protein kinase; domain structure and regulation. TIBS. **14**, 62-66 (1989).

Seelig, H.P. Unpublished. Human macrogolgin. EMBL database, accession number S37536 (November 1994).

Seelig, H.P., Schranz, P., Schroeter, H., Wiemann, C., Griffiths, G. and Renz, M. Molecular genetic analyses of a 376 kilodalton golgi complex membrane protein (Giantin). Mol. Cell Biol. **14**, 2564-2576 (1994).

Shehee, W.R., Loeb, D.D., Adey, N.B., Burton, F.H., Sasavant, N.C., Cole, P., Davies, C.J., McGraw, R.A., Schichman, S.A., Severynse, D.M., Vilvia, C.F., Weyter, F.W., Wisely, G.B., Edgell, M.H. and Hutchison, C.A.. Nucleotide sequence of the BALB/c mouse beta-globin complex. J. Mol. Biol. **205**, 41-62 (1989).

Sikela, J. M., Law, M. L., Kao, F., Hartz, J. A., Wei, Q. and Hahn, W. E. Chromosomal localisation of the human gene for brain Ca^{2+}/calmodulin-dependent protein kinase type IV. Genomics **4**, 21-27 (1989).

Singh, G., Supp, D.M., Schreiner, C., McNeish, J., Merker, H.J., Copeland, N.G., Jenkins, N.A., Potter, S.S. and Scott, W. *legless* insertional mutation: morphological, molecular, and genomic characterisation. Genes Dev. **5**, 2245-2255 (1991).

Stanier, P., Forbes, S.A., Arnason, A., Bjornsson, A., Sveinbjornsdottir, E., Williamson, R. and G. Moore. The localisation of a gene causing X-linked cleft palate and ankyloglossia (CPX) in an icelandic kindred is between DXS326 and DXYS1X. Genomics **17**, 549-555 (1993).

Tanabe, O., Akira, S., Kamya, T., Wong, G. G., Hirano, T. and Kishimoto, T. Genomic structure of the murine IL-6 gene: high degree conservation of potential regulatory sequences between mouse and human. J. Immunology **141**, 3875-3881 (1988).

Tanimoto, K., Handa, S.I., Vewo, N., Murakami, K. and Fukamizu, A.. Structure and sequence analysis of human activin beta-A subunit gene. DNA Seq. **2**, 103-110 (1992).

Taylor, W.L. Calcium/calmodulin-dependent protein kinase II from *Xenopus laevis*. Unpublished. Genbank, accession number U06636 (November 1994).

The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell **72**, 971-983 (1993).

Thompson, T.G., DiDonato, C.J., Simard, L.R., Ingraham, S.E, Burghes, A.H.M., Crawford, T.O., Rochette, C., Mendell, J.R. and Wasmuth, J.J. A novel cDNA detects homozygous microdeletions in greater than 50% of type I spinal muscular atrophy patients. Nature Genet. **9**, 56-58 (1995).

Tillib, S.V., Ryskov, A.P. and Geirgiev, G.P. Nucleotide sequence of small polyadenylated B2 RNA.Nucleic Acids Res. **13**, 6423-6437 (1985).

Tobimatsu, T. and Fujisawa, H. Tissue specific expression of four types of rat calmodulin-dependent protein kinase II mRNAs. J. Biol. Chem. **264**, 17907-17912 (1989).

Tobimatsu, T., Kameshita, I. and Fujisawa, H. Molecular cloning of the cDNA encoding the third polypeptide (γ) of brain calmodulin-dependent protein kinase II. J.Biol. Chem. **263**, 16082-16086 (1988).

Trofatter, J.A., MacCollin, M.M., Rutter, J.L., Murrell, J.R., Duyao, M.P., Papru, D.M., Elderidge, R., Kley, N., Menon, A.G., Pulaski, K., Haase, V.H., Ambrose, C.M., Munroe, D., Bove, C., Haines, J.L., Martuza, R.L., MacDonald, M.E., Seizinger, B.R., Short, M.P., Buckler, A.J. and Gusella, J.F. A novel moesin-, ezrin-radixin-like gene is a candidate for the neurofibromatosis 2 tumor suppressor. Cell **72**, 791-800 (1993).

Troll, H., Winckler, T., Lascau, I., Muller, N., Saurin, W., Vernon, M. and Mutzel, R. Seperate nuclear genes encode cytosolic and mitochondrial nucleoside diphosphate kinase in *Dictyostelium discoideum*. J. Biol. Chem. **268**, 25469-25475 (1993).

Van der Hoeven, F., Schimmag, T., Volkman, A., Mattei, M.G., Kyewsi, B. and Ruther, U. Programmed cell death is affected in the mouse mutant Fused toes (Ft). Development **120**, 2601-2607 (1994).

Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J.S., Fu, Y., Kuhl, D. P. A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F., Eussen, B. E., Van Ommem, G. B., Blonden, L. A. J., Riggins, G. J., Chastain, J. L., Kunst, C. B., Galjaard, H., Caskey, C. T., Nelson, D. L., Oostra, B. A. and Warren, S. T. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell. **65**, 905-914 (1991).

Vulpe, C., Levison, B., Whitney, S., Packman, S. and Gitschier, J. Isolation of a candidate gene for Menkes disease and evidence that it encodes a copper transporting ATPase. Nature Genet. **3**, 7-13 (1993).

Wagner, E.F., Covarrubius, L., Stewart, T.A. and Mintz, B. Prenatal lethalities in mice homozygous for human growth hormone gene sequences integrated in the germ line. Cell **35**, 647-655 (1983).

Wagner, J.A., Cozens, A.L., Schulman, H., Gruenert, D.C., Stryer, L. and Gardner, P. Activation of chloride channels in normal and cystic-fibrosis airway epithelial-cells by multifunctional calcium calmodulin-dependent protein kinase. Nature **349**, 793-796 (1991).

Watanabe, H., Kimata, K., Line, S., Strong, D., Luo-yi Gao, Kozak, C.A. and Yamada, Y. Mouse cartilage matrix deficiency (cmd) caused by a 7bp deletion in the aggrecan gene. Nature Genet. **7**, 154-157 (1994).

Weber, J.L. *Homo sapiens* DNA sequence, repeat region. Unpublished. Genbank accession number 122381 (Nov 1994).

Weiher, H., Noda, T., Gray, D.A., Sharp, A.H. and Jaenisch, R. Transgenic mouse model of kidney disease: insertional inactivation of ubiquitously expressed gene leads to nephrotic syndrome. Cell **62**, 425-434 (1990).

Wilson, J.B., Ferguson, M.W.J., Jenkins, N.A., Lock, L.F., Copeland, N.G. and Levine, A.J. Transgenic mouse model of X-linked cleft palate. Cell Growth Diff. **4**, 67-76 (1993).

Wilson, J.B., Weinberg, W., Johnson, R., Yuspa, S. and Levine, A.J. Expression of the BNLF-1 oncogene of Epstein-Barr Virus in the skin of transgenic mice induces hyperplasia and aberrant expression of Keratin 6. Cell **61**, 1315-1327 (1990).

Wu, H., Bateman, J.F., Schnieke, A., Sharpe, A., Barker, D., Mascara, T., Eyre, D., Bruns, R., Krimpenfort, P., Berns, A. and Jaenisch, R. Human-mouse interspecies collagen I heterotrimer is functional during embryonic development of Mov13 mutant mouse embryos. Mol. Cell. Biol. **10**, 1452-1460 (1990).

Wu, L., Gaskins, C. J., Zhou, K., Firtel, R. A. and Devreotes, P. N. Cloning and targeted mutations of G alpha-7 and G alpha-8: Two developmentally regulated G protein alpha subunit genes in Dictyostelium discoideum. Mol. Biol. Cell. **5**, 691-702 (1994).

Yamamoto, M., Ko, L. J., Leonard, M. W., Beug, H., Orkin, S. H. and Engel, J. D. Activity and tissue-specific expression of the transcription factor NF-E1 multigene family. Genes Dev. **4**, 1650-1662 (1990).

Yeo, J., Alderuccio, F. and Toh, B. A new chromosomal protein essential for mitotic spindle assembly. Nature **367**, 288-291 (1994).

Yoshimura, M. and Oka, T. Transfection of beta-casein chimeric gene and hormonal induction of its expression in primary murine mammary epithelial cells. Proc. Natl. Acad. Sci. U.S.A. **87**, 3670-3674 (1990).

Young, J.R., Shah, J. O., Kimmel, B. E., Iams, K. P. and Williams, R.O. The 5' flanking region of *Trypanosoma brucei* variable surface glycoprotein gene. Mol. Biochem. Parasitol. **24**, 163-174 (1987).

Yuki, S., Inouye, S., Ishimaqru, S. and Saigo, K. Nucleotide sequence characterisation of a drosophilia retrotransposon,412. Eur. J. Biochem. **158**, 403-410 (1986).