Functional and evolutionary analysis of the mouse Muc-1 gene.

Andrew Paul Spicer

August 1993

A thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy of the University of London.

Molecular Epithelial
Cell Biology Laboratory,
Imperial Cancer Research Fund,
P.O. Box 123,
Lincoln's Inn Fields,
London WC2A 3PX
UK.

Department of Biology,
Medawar Building,
University College London,
Gower Street,
London WC1E 6BT
UK.

For Sharon

ABSTRACT

The mouse homologue of the human tumour-associated mucin, MUC1, was cloned and full-length sequence was determined. This mucin (previously called polymorphic epithelial mucin) is expressed by the majority of simple secretory epithelial cells in both the mouse and human and is also overexpressed in a large percentage of carcinomas. The mouse gene, Muc-1, encodes an integral membrane protein with 44% of its coding capacity made up of serine, threonine and proline, a composition typical of a highly O-glycosylated protein. The Muc-1 core protein consists of an amino-terminal signal sequence, a repetitive domain encoding 16 repeats of 20-21 amino acids, and unique sequence containing membrane-spanning and cytoplasmic domains. Although overall homology with the human MUC1 protein is only 53%, the transmembrane and cytoplasmic domains exhibit homologies of 90% and 87%, respectively. This level of sequence conservation would suggest that these regions may be functionally important. Interestingly, the mouse homologue, unlike its human counterpart does not exhibit a variable number tandem repeat (VNTR) polymorphism. However, this type of polymorphism was found to be present in all other mammalian groups analysed. Data is presented, including sequence obtained for the Muc-1 gene from a large number of species, to suggest how this gene has evolved and to explain possible reasons why the mouse Muc-1 gene does not exhibit minisatellite characteristics.

Numerous functions have been suggested for this molecule, yet it still remains unclear what role this protein plays in the tissues and tumours in which it is expressed. In an effort to learn more of the function of the mouse Muc-1 gene, the gene was specifically mutated in embryonic stem (ES) cells. Targeting vectors derived through genomic clones from two strains of mice were utilised and their relative targeting efficiencies are discussed. Several mouse cell lines were created carrying a disruption in the Muc-1 gene. These cell lines were injected into nude mice to create tumours and also injected into blastocysts, in order to generate mice carrying the Muc-1 mutation. These mouse lines will provide a crucial tool in the analysis of the function of this molecule *in vivo*.

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Sandy Gendler, in whose lab I have been privileged to work these past four years. Sandy, you have been a wonderful research supervisor and friend over these past four years and have always given me continual encouragement and support. Your enthusiasm for the subject has been limitless and infectious.

There are numerous other people that I need to thank both in the UK and in the US. My friends at ICRF in London, especially Trevor Duhig and Nigel Peat for their friendship and our numerous coffee break, lunch and 'other' interesting discussions. Between the two of them they taught me a lot of molecular biology, and a lot of 'life' in general. Discussions with them were always nothing but colourful! I would also like to thank numerous other ICRF staff, notably Lucy Pemberton and Vania Braga. Without them, ICRF lab life wouldn't have been the same. At University College I would like to thank my internal research supervisor, Dr. Bill Richardson, for his helpful discussions of my research and for his help through submitting this thesis, and Peter King for his kind donation of the wild mouse tissue samples.

In the US I would like to especially thank Suresh Savarirayan, Bob Cichon, Anita Jennings and Dawn Taylor at the Mayo Clinic. I thank Suresh for expertly carrying out all the ES animal work and teaching me the fine art of blastocyst microinjection and manipulation. Bob I thank for the routine maintenance and breeding of the mice described herein. Without Bob and Suresh, the germline mice described within would not have been possible. I thank Anita for her great stories, including the now infamous "tomato-cake story", and for her outstanding histological preparations. I acknowledge Dawn Taylor and Kelly Ann Ross in the Visual Communications department for their help in preparing the photographic plates. I acknowledge Dr. Stuart Patton for his gift of the milk fat globule proteins from the species studied herein, Dr. Pat DeCoursey of the University of South Carolina for her gift of the squirrel tissue samples and thanks to Peg Davis (University of Arizona) for her magnificent guts (guinea-pig). In addition, I would also like to thank Dr. Stephen F. Kingsmore of Duke University for his excellent work on chromosomal localisation of the mouse Muc-1 gene described herein.

This acknowledgements list would not be complete without mention of my Mum and Dad and my brothers and sister who continually encouraged me throughout my PhD, and of my grandparents who gave me a wonderful place to stay while in London, and who also gave me many hilarious evenings throughout my three year stay with them. Lastly, I acknowledge my fianceé, Sharon. Without her help with moving and adjusting to life in the US, and without her continual support, love, encouragement and the sometimes very necessary 'distractions' she provided, life this year would have been very tough. This thesis is dedicated to her.

TABLE OF CONTENTS

Figures

Figures

ABBREVIATIONS

The human mucin gene will be referred to as **MUC1** throughout, whereas the homologous mouse mucin gene will be referred to as **Muc-1**. This is to conform to the designated gene locus names as decided at the 1st International Workshop on Carcinoma Associated Mucins, San Francisco 1990. As this thesis is centred around the mouse mucin gene, Muc-1, for simplicity the homologous mucin genes from the other species discussed will be referred to as '**Species Name Muc-1, i.e bovine Muc-1 or rabbit Muc-1**'. In addition, when two species homologues are being referred to they will be referred to collectively as Muc-1 genes.

bps	base pairs of DNA
β-gal	β-galactosidase gene of *Escherichia coli*
BSA	bovine serum albumin
CAT	chloramphenicol acetyl transferase
cDNA	copy deoxyribonucleic acid
CTAB	cetyltrimethylammonium bromide
DAB	diaminobenzidine tetrahydrochloride
DMEM	Dulbecco's modified Eagles medium
DMSO	dimethyl sulphoxide
DNA	deoxyribonucleic acid
ES cell	embryonic stem cell
EDTA	ethylene diamino-tetra-acetic acid
FCS/FBS	foetal calf serum/foetal bovine serum
GanC	gancyclovir
G418	geneticin
HSV-tk	Herpes simplex virus 1 thymidine kinase gene
ICM	inner cell mass
i.p.	intra-peritoneally
IPTG	isopropyl-β-D-galactopyranoside
kbps	kilobase pairs of DNA
LacZ	bacterial β-galactosidase gene
LIF	leukaemia inhibitory factor
μl/s	microlitre/s
ml/s	mililitre/s
Mb/s	megabase pair/s of DNA
mRNA	messenger ribonucleic acid
μM	micromolar

nM	nanomolar
mM	milimolar
M	molar
Myr	million years
nm/s	nanometer/s
neo	neomycin phosphotransferase
OD	optical density
PBS	phosphate buffered saline
pBS-SKII+	pBluescriptSKII+ cloning vector
pBS-KSII+	pBluescriptKSII+ cloning vector
p.c.	post-coitum
pfus	plaque forming units
PCR	polymerase chain reaction
PEM	polymorphic epithelial mucin
PGKneo	neomycin phosphotransferase gene driven by the promoter of the mouse phosphoglycerate kinase-1 gene
pmole	picomole
PNS	positive negative selection
RFLP	restriction fragment length polymorphism
RFLV	restriction fragment length variant
RNA	ribonucleic acid
rpm	revolutions per minute
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
SSC	saline sodium citrate
SDS	sodium dodecyl sulphate
s.c.	sub-cutaneous/ly
TENS	Tris-HCl, EDTA, sodium hydroxide, SDS
T25	tissue culture flask with growth area of $25cm^2$
T75	tissue culture flask with growth area of $75cm^2$
T175	tissue culture flask with growth area of $175cm^2$
VNTR	variable number tandem repeat
(v/v)	volume to volume
(w/v)	weight to volume
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside

CHAPTER ONE:

GENERAL INTRODUCTION

Lumenal cells of the secretory epithelial organs of the body are in contact with a variety of external environmental elements. As such, the apical or lumenal surface of these cells is covered by a protective secretion which serves as a selective physical barrier between the extracellular milieu and the plasma membrane and cell interior. This secretion is often referred to as mucous and is made up of a number of protein constituents, the most prominent of which are the mucins. Mucins, or mucin-type glycoproteins, are characterised as large extended molecules with a high percentage (50-90%) of their molecular mass made up of carbohydrate, which is attached via O-glycosidic linkage through N-acetylgalactosamine to serine and/or threonine.

Mucin-type glycoproteins can be sub-divided into the classical secretory, or soluble mucins, and the membrane-bound mucin-like glycoproteins. The secretory, or soluble mucins, constitute the viscous mucous of the lungs, tracheo-bronchial tract, gut and reproductive tract, and typically form extremely large complexed oligomers through linkage of protein monomers via disulphide bonds. These proteins are secreted from the cell, and although they remain at the apical surface of the epithelial cells in the form of a mucous-gel, they are not membrane bound. The membrane-bound mucin-like glycoproteins, however, are intimately associated with the plasma membrane through a hydrophobic membrane-spanning domain and have not been observed to form oligomeric complexes. It does, however, appear that these membrane-bound mucin-like glycoproteins can also be released from the cell membrane by some sort of proteolytic cleavage event (Fig. 1.11).

Attention has been focused on one particular membrane-bound mucin-like glycoprotein, identified initially as a component of the

human milk-fat-globule membrane (Ceriani, 1977). Attention was focused on this molecule primarily due to the fact that antisera raised against elements of the human milk-fat-globule membrane were also observed to react with proteins expressed by a number of human mammary carcinoma cell lines in addition to tissue sections of human mammary tumours (Arklie, 1981; Taylor-Papadimitriou, 1981).

This mucin has been variously called epithelial membrane antigen (EMA) (Heyderman, 1979), Ca1 antigen (Ashall, 1982), PAS-O (Schimizu, 1982), DU-PAN-2 antigen (Metzgar, 1982), peanut-reactive urinary mucin (PUM) (Karlsson, 1983), non-penetrating glycoprotein (NPGP) (Ceriani, 1983), DF3 antigen (Kufe, 1984), MAM-6 (Hilkens, 1984), NCRC11 (Price, 1985); epitectin (Bramwell, 1986), polymorphic epithelial mucin (PEM) (Gendler, 1988), H23 antigen (Keydar, 1989), and episialin (Ligtenberg, 1990) as a direct result of the large number of groups worldwide that have been involved in the investigation of this molecule. However, although it has had a large number of designated names, it was found to be the product of a single gene locus, designated MUC1 in the human. As such, this protein will herein be referred to as the MUC1 protein.

1.1 The human MUC1 gene and protein

The derivation of antibodies directed to the deglycosylated mucin of the human milk-fat-globule membrane and similarly to the deglycosylated mucin of the human pancreatic carcinoma cell line, HPAF, allowed the isolation of cDNA clones for the human MUC1 gene through the screening of cDNA expression libraries constructed from mammary and pancreatic carcinoma cell line mRNAs (Gendler, 1987a; Gendler, 1988; Siddiqui, 1988; Gendler, 1990a; Ligtenberg, 1990; Lan, 1990; Wreschner, 1990). Alignment of DNA sequences derived from the mucin clones revealed that the respective cDNAs were the product of the same gene locus.

Prior to the isolation of DNA clones for this gene, it had been shown that the peanut-reactive urinary mucin (PUM), isolated from urine, kidney and lung, electrophoresed on SDS-polyacrylamide gels, and detected using lectin staining, exhibited a genetic polymorphism characterised by the existence in most individuals of two co-migrating

protein bands that showed person-to-person variation (Karlsson, 1983). The patterns of protein bands observed were consistent with the simple Mendelian inheritance of co-dominant alleles at a single gene locus, but it was not known whether this polymorphism was the result of variation of the gene encoding the protein core or the glycan part of the molecule being analysed. A variety of tumour-binding antibodies, raised against the human milk-fat-globule membrane, were demonstrated to detect proteins on Western blots with a pattern of polymorphic bands reminiscent of the PUM polymorphism. Subsequently, it was deduced that the epitopes being recognised by these antibodies were carried on the same molecules as the lectin-binding determinants (Swallow, 1986).

Analysis of the DNA sequence obtained from the respective MUC1 clones revealed the presence of multiple copies of a highly GC-rich 60 base pair tandem repeat encoding a 20 amino acid repeat motif rich in serine, threonine and proline (Gendler, 1987a; Gendler, 1988; Siddiqui, 1988; Ligtenberg, 1990; Lan, 1990; Wreschner, 1990). Southern blots, utilising probes containing this repetitive sequence, yielded a result that was directly comparable to the pattern observed on protein gels for the PUM protein (Swallow, 1987a). It became apparent that the observed genetic polymorphism was, therefore, the result of varying numbers of a 60 base pair tandem repeat within the coding domain. This variable number tandem repeat (VNTR) polymorphism could be demonstrated at the DNA (by Southern blot), the RNA (by northern blot) and protein (by Western blot) levels (Gendler, 1987b). A screen of 69 random individuals of Northern European origin identified 30 different allele lengths, and the variation in number of repeats per allele has been observed to range from a low of between 20 to 30, up to a high of over 100 repeat units (Gendler, 1990a). This variability in numbers of repeats per allele makes the MUC1 locus one of the few expressed VNTR sequences thus far identified in the human genome. Repeat length differences appear to be generated by both intra and interallelic unequal recombination events operating within the repeat array, and this topic will be discussed in detail later.

The sequence of the repeat unit encodes a motif rich in serine, threonine and proline, and it is this region of the protein that is extensively O-glycosylated. It was suggested by Gendler, 1990a, that large variations in the number of repeat units per allele might imply that the

length of the repetitive array is not critical for function, but that the repeat array functions primarily as a scaffold for the attachment of O-linked carbohydrate. As early as 1973, it was postulated that the core protein of the bovine submaxillary mucin was comprised of repeating amino acids (Pigman, 1973). Through the isolation of DNA clones for the genes encoding the human MUC2 through MUC6 mucin genes (Gum, 1989; Gum, 1990; Porchet, 1991; Van Cong, 1990; Aubert, 1991 and Toribara, 1993), the porcine (Timpte, 1988; Eckhardt, 1991) and bovine (Bhargava, 1990) submaxillary mucin genes, three *Xenopus* integumentary mucin genes (Hoffmann, 1988; Probst, 1990 and Probst, 1992; Hauser, 1992), two rat intestinal mucin genes (Gum, 1991; Xu, 1992), and a gene coding for the apo-polysialoglycoprotein of rainbow trout eggs (Sorimachi, 1988), it has become apparent that the presence of a repetitive sequence domain coding for an expressed sequence rich in serine and/or threonine and proline is a general characteristic of mucin genes. Like the human MUC1 gene, the repetitive domain of several of the other mucin genes also displays the characteristic VNTR polymorphism (Sorimachi, 1988; Timpte, 1988; Griffiths, 1990; Hauser, 1990; Eckhardt, 1991; Porchet, 1991; Toribara, 1991; Toribara, 1993). However, it should be pointed out that although all the mucin genes thus far characterised possess a repetitive domain rich in serine and/or threonine and proline, apart from similarity between homologues, the respective repeat units bear no homology to each other and also differ in the number of amino acid residues per repeat (Table 1).

The human mucin genes, MUC1 through MUC6, have been localised to chromosome 1q21 (Swallow, 1987b; Middleton-Price, 1988), 11p15.5 (Griffiths, 1990), 7q22 (Gum, 1990), 3q29 (Porchet, 1991), 11p15 (Van Cong, 1990) and 11p15.4-11p15.5 (Toribara, 1993), respectively. Although the genes for MUC2, 5 and 6 are located on the same chromosome, and have been mapped to lie in the same area, they show differing expression profiles and differ both in the length and sequence of their respective repeat units. It, therefore, remains to be seen whether or not these three genes could have arisen from one ancestral mucin gene through a series of duplications, and formed a multigene family in which the members subsequently diverged to take up differing patterns of expression. However, the MUC1, MUC3 and MUC4 genes are perhaps less likely to have arisen through this type of duplication process as they

Until sequence is determined for the 5' and/or 3' region flanking the central repetitive portions of the MUC3, MUC4, MUC5 and MUC6 genes it will be difficult to analyse the possible common evolutionary of the mammalian mucin genes.

share no similarity in sequence, repeat unit length, gene structure or chromosome position.

From the deduced repeat sequence of the mucin genes thus far characterised it appears that a large repetitive domain comprised of a high percentage of serine and/or threonine interspersed by amino acids with small side chains such as proline, glycine and alanine represents an efficient scaffold for O-linked glycosidic attachment, as this type of sequence structure has been independently evolved and maintained in a large number of mucin genes in several different species. O-glycosidic bonds are formed by the linkage of N-acetylgalactosamine and the hydroxyl group of serine or threonine, and represent the first step in O-linked oligosaccharide synthesis. The action of the UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase is thought to catalyse the following reactions (McGuire, 1967):

UDP-GalNAc + serine/threonine-polypeptide

→ GalNAc-serine/threonine-polypeptide + UDP

A UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase has recently been purified and characterised from porcine submaxillary glands (Wang, 1992). The porcine submaxillary enzyme was found to specifically incorporate N-acetylgalactosamine from UDP-GalNAc into glycosidic linkages only with the hydroxyl groups of threonine in peptide substrates. Moreover, neither porcine, bovine, nor ovine submaxillary gland homogenates were able to incorporate N-acetylgalactosamine into glycosidic linkages with serine residues in peptide substrates (Wang, 1992). These studies demonstrated that different transferases catalyse the the formation of O-glycosidic linkages on serine and threonine. The authors also examined the relative extent of glycosylation of threonines present in a variety of peptide substrates. Although, from this analysis, it is still not clear what the consensus amino acid sequence is for the O-glycosylation of threonine, it was demonstrated that some threonyl hydroxyl groups were more readily glycosylated than others, and some possible glycosylated products were simply not formed.

Mucin glycoproteins represent a class of proteins with possibly the highest level of O-linked carbohydrate observed (Jentoft, 1990). Therefore,

although the consensus amino acid sequence for O-linked attachment is as yet not determined, from a study of the sequences present in the repeat sequences of the various mucin proteins it can be specified that O-linked attachment to serines and threonines requires the presence of single or multiple serine or threonine residues in combination with helix-disrupting residues such as prolines, and amino acids with short side chains such as glycine and alanine.

A determination of the full-length cDNA sequence of the human MUC1 gene revealed that the sequence encoded a protein with four distinct domains; an amino-terminal signal peptide domain, the large domain made up of variable numbers of the 20 amino acid repeat motif flanked on either side by degenerate copies of the same repeat unit, and unique sequence which included a carboxy-terminal hydrophobic membrane-spanning domain of 31 amino acids, and a cytoplasmic domain of 69 amino acids (Gendler, 1990a; Ligtenberg, 1990; Lan, 1990; Wreschner, 1990). Genomic sequence of the gene revealed that the coding sequence was represented on seven exons spanning between 4 and 7 kilobase pairs of DNA, depending upon the number of repeat units present (Lancaster, 1990; Wreschner, 1990). The entire repeat domain was found to be located entirely within the second exon. The deduced amino acid sequence encoded a core protein with a predicted molecular mass ranging from 120 kilodaltons up to 225 kilodaltons (Gendler, 1990a). The size observed on SDS-protein gels, for the mature human mammary mucin, is in the 300 to > 400 kilodalton range and this would imply that as much as 50% or more of the weight of the mature glycosylated protein is made up of carbohydrate. From an analysis of the fully glycosylated MUC1 product isolated from a variety of tissues it appears that the MUC1 protein present in the pancreas has a molecular mass closer to 1000 kilodaltons, much larger than the MUC1 protein identified in the milk-fat-globule. This apparent organ-specific size variation of the mucin may be simply a reflection of the spectrum of glycosyltransferases present within a particular organ or, alternatively, may possibly be a reflection of a carbohydrate-associated function of the MUC1 protein in each organ.

Within each tandem repeat sequence there were found two doublets of serine and threonine in addition to a single threonine residue. Potentially, therefore, each repeat unit within the protein has five sites for O-linked carbohydrate attachment. An average size mucin

molecule made up of 40 repeat units could, therefore, potentially carry as many as 200 O-linked carbohydrate side chains. In addition to the numerous sites for O-linked glycosylation, five potential sites for N-linked attachment of carbohydrate were also found, located between the repetitive and membrane-spanning domains. Although the MUC1 protein product has been demonstrated to carry N-linked sugar chains (Hilkens, 1984), it is not clear which of these sites are glycosylated.

The analysis of the sequence of the respective cDNA clones for the human MUC1 gene resulted in the identification of two sites within the gene at which an alternative splicing mechanism may operate (Ligtenberg, 1990; Wreschner, 1990). However, the alternative splice site identified by Ligtenberg, 1990, is the only one of these two sites that has been verified. At the amino-terminus of the molecule, it has been shown that an alternative splice variant, located at the three-prime end of the first intron, results in the addition of nine extra amino acids onto the signal peptide. It appears that this splice site recognition is based on a single A/G nucleotide difference in exon 2, eight nucleotides downstream of the second splice acceptor site. The presence of an A at this position appears to result in the shorter of the two alternative splice products, whereas a G specifies the splicing in of the additional nine amino acid residues (Ligtenberg, 1991). It is yet to be discovered what if any functional significance this extra sequence might have, but it has been observed that in a number of cell lines and tissues expressing MUC1 the ratio of the two alternative transcripts does vary. The authors propose that, in an expressing cell line or tissue possessing alleles of two different lengths due to differing numbers of tandem repeats, the larger allele is in each case generally associated with the longer splice variant, whereas the small allele is generally found linked to the shorter splice variant. Their data show that in samples possessing only the so-called "small allele", the longer splice variant is nearly always undetectable. This observation would suggest that the majority of individuals homozygous for the so-called "small allele" are incapable of expressing the longer splice variant of the MUC1 message. Similarly, those individuals homozygous for the "large allele" will almost invariantly express only the longer splice variant. It is, therefore, hard to imagine what possible differences in function, if any, the two respective splice variants may specify.

A second variant splice site, which would generate a transcript that incorporates the entire second intron into the mRNA, resulting in the generation of a premature termination signal and thus a transcript lacking the membrane-spanning and cytoplasmic determinants, was proposed by Wreschner, 1990. They suggested that this alternative splice variant represents the transcript of the so-called secreted form of the mucin. However, this alternative splice variant has remained undetected by all other investigators, and it is now generally accepted that this transcript and clone most probably arose through an incompletely spliced messenger RNA molecule. The MUC1 protein is, however, observed to be secreted from the cells that express it, but this secretion is most likely the result of a proteolytic cleavage event which releases the extracellular portion of the protein from its membrane bound and cytoplasmic domains (Ligtenberg, 1992a; Boshell, 1992).

The post-translational modifications of the MUC1 protein represent one of the most intriguing areas of the biology of this molecule. As mentioned previously, one of the primary reasons for the initial focus of interest in this molecule was due to the observation that several antibodies directed to native and deglycosylated human milk-fat-globule proteins were found to react strongly with proteins present in sections of human mammary tumours. Indeed, some antibodies, such as SM3, reacted very strongly with the tumour-associated MUC1 product, yet showed little or no reactivity with the MUC1 product of the corresponding normal tissue (Burchell, 1987; Girling, 1989). This type of tumour-specific reactivity of certain antibodies directed to the MUC1 protein product has now been observed through several independently derived monoclonal antibodies (Devine, 1990; Taylor-Papadimitriou, 1991). Studies have shown that the tumour-specific reactivity of these antibodies is a reflection of the glycosylation status of the MUC1 core protein in these tissues. This is not surprising as many of the respective antibodies were raised to the deglycosylated mucin or milk-fat-globule. Epitope mapping revealed that the sequence, PDTRP, located within the tandem repeat, was the epitope of the tumour-specific monoclonal antibody, SM3 (Burchell, 1989). Indeed, a large number of the derived tumour-reactive antibodies were found to have epitopes either of this same sequence or of a slightly overlapping variant of it (compiled in Taylor-Papadimitriou, 1991 and Taylor-Papadimitriou, 1992), and it has subsequently been shown, through secondary and tertiary structure

predictions and antigenicity predictions, that this region of the tandem repeat forms an antigenic β-turn (Fontenot, manuscript submitted; Taylor-Papadimitriou, 1992). Flanking the SM3 epitope are the two respective doublets of serine and threonine, i.e. ...V T S A P D T R P A P G S T A.... In most normal secretory epithelia the presence of branched O-linked side-chains (Hanisch, 1989), attached to these sites, is enough to mask the protein core and epitope. The MUC1 protein product of carcinomas and carcinoma cell lines has been shown to have less complex, shorter carbohydrate side-chains which result in the exposure of novel core protein epitopes (Hull, 1989). Thus, antibodies raised to the deglycosylated mucin recognise novel exposed epitopes on the carcinoma-associated MUC1 product but are unable to recognise the product of the normal tissue due to masking by oligosaccharides. The expression of novel, tumour-associated mucin epitopes by greater than 90% of human breast, pancreatic, ovarian, colon and lung carcinomas has led to the exciting possibility for the use of the MUC1 protein as a target in the specific active immunotherapy of these cancers (Hareuveni, 1990, Lalani, 1991).

Breast carcinoma cells that are found to express the MUC1 protein typically express levels higher than the level observed in the corresponding normal tissue (Zaretsky, 1990). In addition colon carcinomas also express this molecule at high levels (Boland, 1982; Irimura, 1989; Hoff, 1989; Irimura, 1991). It is, perhaps, the high levels of expression of the MUC1 protein in these cells that results in the addition of shorter carbohydrate side chains to the tumour-associated MUC1 protein. One can envisage that an over-expression of mucin without a parallel over-expression of the enzymes required for O-glycosylation may lead to under-glycosylated mucin molecules reaching the cell surface (Carraway, 1992). Alternatively, the shorter side chains added to the tumour-associated MUC1 protein may be a reflection of a general or specific disruption in the activity of certain glycosyltransferase enzymes present within the Golgi apparatus of these cells.

A second post-translational modification of the MUC1 protein that has been of great interest is its secretion. This molecule can be detected in the lumenal secretions of the epithelial cells that express it, and the

molecule can also be detected circulating in the serum of patients with breast and pancreatic carcinoma (Burchell, 1984; Metzgar, 1984; Hayes, 1985; Hilkens, 1986; Tsarfaty, 1988). As mentioned previously, Wreschner, 1990, reported the isolation of a cDNA clone for the so-called secreted form of the human MUC1 protein. However, as described earlier, this clone has now been generally accepted to have been generated from an incompletely spliced messenger RNA species. It has been assumed that the most likely mechanism for the secretion of MUC1 is some kind of proteolytic cleavage event occurring between the repetitive and membrane spanning domain (Hilkens, 1988). Ligtenberg, 1992a, demonstrated that this type of proteolytic cleavage-associated secretion of the molecule does indeed occur. The area of the proteolytic cleavage event was found to reside within 71 and 53 amino acids upstream of the transmembrane domain. In the region to which the cleavage event has been mapped, two phenylalanine-arginine doublets are present, putative substrates for a family of serine proteases, the kallikreins (Fiedler, 1987). The authors demonstrated that the cleavage site may be cleaved while the protein is still within the endoplasmic reticulum and, thus, prior to the protein's passage to the apical surface of the cell. This implies, however, that the extracellular and membrane-spanning/cytoplasmic portions (the two cleavage products) remain associated, as antibodies derived to sequences within the cytoplasmic portion of the MUC1 protein efficiently immunoprecipitate the entire molecule (Ligtenberg, 1992a; Pemberton, 1992). The possible mechanism for this stable association is not known, but it is known that a disulphide bond cannot be involved, as the only cysteine residues present within the protein sequence are located within the transmembrane domain.

To date the cleavage of the MUC1 molecule and the subsequent stable association of the cleavage products prior to transport to the cell surface, remain points of contention. One example in which a membrane-bound mucin-like glycoprotein has been shown to be comprised of two cleavage products that remain stably associated at the cell surface is in the rat mammary carcinoma-associated mucin glycoproteins, ASGP-1 and ASGP-2. In this instance, however, it has been shown that this stable association is due to the formation of disulphide bonds between the numerous cysteine residues present within each sub-unit (Sheng, 1990 and Sheng, 1992).

Cloning of regions encoding the the non-repetitive portions of the respective secreted mucins has been extraordinarily difficult, and this is presumably due to the fact that these genes are characterised by extremely large mRNAs, in the range of 11-14 kilobase pairs, comprised primarily of multiple copies of the respective tandem repeat. In general, clones for the mucin genes have been obtained through screening cDNA expression libraries with monoclonal antibodies or polyclonal serum raised to the deglycosylated mucin of the particular tissue or cell line being studied (Gendler, 1987a; Gum, 1989; Lan, 1990; Gum, 1991; Toribara, 1993). Due to the repetitive nature of the mucins, the majority of antibodies raised to the deglycosylated mucins react with epitopes present within the repeats and, therefore, the use of antibodies such as these, in library screening, has yielded clones containing multiple copies of the tandem repeat. To date, the human MUC1 gene and its mouse homologue, Muc-1, remain the only mucin genes to be completely sequenced.

Cloning of the mouse homologue (designated Muc-1) of the MUC1 gene (Spicer, 1991; Vos, 1991), and the derivation of a MUC1 species cross-reactive antiserum (Pemberton, 1992), has allowed a detailed investigation of the pattern of expression of the mouse Muc-1 gene during mouse embryogenesis and in the adult mouse (Braga, 1992; Pemberton, 1992). The pattern of expression observed in the adult mouse reflects precisely the expression pattern previously reported for the human MUC1 gene and for the human MUC1 gene in transgenic mice (Zotter, 1988; Peat, 1992). The pattern of expression of Muc-1 during mouse embryogenesis was investigated using northern analysis of mRNA, RT-PCR (reverse transcriptase-polymerase chain reaction), and immunohistochemistry, utilising the polyclonal antiserum, CT1, raised to a peptide composed of the last 17 amino acids of the human MUC1 protein sequence (Braga, 1992; Pemberton, 1992). By immunohistochemistry, the Muc-1 protein was first detectable in mouse embryonic stomach, pancreas and lung at gestational day 12 (vaginal plug = day 1). In each case the protein was detected on the apical surface of the lumenal epithelial cells. Although the protein was detectable in several different organs of the developing mouse, its expression was observed not to be induced systemically, but according to the particular onset of epithelial polarisation and branching morphogenesis in each individual organ (Fig 1.12). Muc-1 expression was observed to correlate well with the epithelial differentiation status of the stomach, pancreas,

lung, trachea, kidney and salivary glands and was detectable prior to the onset of glandular activity of these organs. The authors propose that this pattern of expression may indicate an important function for the Muc-1 molecule in the early development of these organs.

Interest in the human MUC1 protein was initiated by the fact that it was detected as being highly expressed by human carcinomas. The molecule has been estimated to be overexpressed by at least an order of magnitude in carcinomatous tissue versus the normal tissue. Perhaps the highest levels of expression of this molecule are observed in the lumenal cells of the lactating mammary gland and in mammary carcinomas. The temporally and spatially regulated expression in the developing embryo, the overexpression in human carcinomas, and the up-regulation of expression at lactation have generated a great deal of interest in the promoter and regulatory elements of the MUC1 gene. Transgenic mouse lines carrying the human MUC1 gene flanked by as little as 1.6 kilobase pairs of 5' regulatory sequence, were observed to express MUC1 in the correct tissue-specific manner, and this suggested that most, if not all, of the elements required to define the precise expression pattern of the MUC1 gene were present within the 1.6 kilobase pairs of 5' flanking sequence (Peat, 1992).

Subsequently, two groups have analysed the regulatory sequences of the human MUC1 gene in a more detailed fashion. Both groups have utilised deletion fragments of the human MUC1 promoter region coupled to the reporter plasmid for the CAT (chloramphenicol acetyl transferase) gene expressed in cell lines that had been previously shown to express the MUC1 gene. The two respective groups have identified two different regions, within 550 base pairs of the transcription start site of the gene, that appear to be involved in the control of expression of the MUC1 gene. It should be pointed out that upstream of the MUC1 gene, within 3 kilobase pairs, the polyadenylation sequence for the thrombospondin-3 (Thbs-3) gene is located (Vos, 1992). This protein shows a very different pattern of expression to that observed for its immediate neighbour the MUC1 gene and, therefore, the promoters of the two genes are presumed to be independent. The close proximity of an adjacent gene displaying a different expression profile suggests that the promoter of the MUC1 gene is relatively compact. Kovarik, 1993, identified a sequence, E-MUC1 (5' CACCTGTCACCTG 3'), located immediately upstream of the tata box,

within 100 base pairs of the transcription start site, which appears to be binding specific transcription regulating factors responsible for determining the tissue-specific expression of the MUC1 gene. This binding site seems to be binding one or more factors which operate in concert with Sp1 protein, binding to an adjacent Sp1 site. Specific mutation of the Sp1 or E-MUC1 sequences in a CAT reporter construct containing an SV40 enhancer (pEnCAT) resulted in a moderate increase in transcription of the CAT reporter gene in a cell line (HT1080 human fibrosarcoma) in which the MUC1 gene is not normally expressed. Mutation of both the Sp1 and E-MUC1 sites resulted in a marked increase of expression of the reporter gene in this cell line. The same constructs, analysed in the breast adenocarcinoma cell line, ZR-75-1, a cell line that has been shown to express the MUC1 gene at high levels, indicated that mutation in the E-MUC1 site alone resulted in no significant difference from wild type in transcription levels of the CAT reporter. Mutation of the Sp1 site alone resulted in a moderate decrease in the rate of transcription. In the ZR-75-1 cell line, the double mutant construct gave the same level of transcription as that observed for the Sp1 single mutant. This observed level of expression was also similar to the level observed in the HT1080 cell line when transfected with the double mutant construct. These results indicate that the E-MUC1 site is binding factors involved in the repression of MUC1 transcription in cells that do not normally express MUC1. Gel-shift assays, employing an E-MUC1 oligonucleotide, generated what appeared to be three shifted DNA-protein complexes from nuclear extracts of the expressing cell line, ZR-75-1, and only a single complex from extracts of the non-expressing cell line, HT1080.

Taken together, these results may suggest that non-expressing cells possess a single factor that binds immediately adjacent to the tata box, repressing transcription of the MUC1 gene. Expressing cells, on the other hand, may possess additional protein factors that could possibly bind to the E-MUC1 recognition site and the repressor protein, forming a multi-protein DNA complex which relieves the repressor activity, thus allowing the initiation of transcription. It is interesting to note that the sequence of the E-MUC1 binding site includes a direct repeat of 5' AGGTGA 3' separated by a single nucleotide. This sequence has been identified as a binding site for members of the retinoic acid receptor

family (Umesono, 1991; Cooney, 1992), some of which (e.g. Coup2) are able to act as negative regulators of transcription.

Recently, a second region has been identified, within 550 base pairs of the MUC1 transcription start, that also appears to be involved in the regulation of transcription of the MUC1 gene (Abe, 1993). Employing a similar stategy to that of Kovarik, this group utilised the breast adenocarcinoma cell line, MCF-7, in combination with CAT reporter constructs to identify a specific sequence between -507 and -483 that appears to specifically bind a 45 kilodalton (kDa) protein present only in MCF-7 cells. The authors demonstrated that this sequence acts as an orientation independent enhancer in reporter constructs. It is interesting to note that the sequence to which the 45 kDa protein binds, 5'-GGG AAG TGG TGG GGG GAG GGA-3', overlaps with a sequence that shares homology with the binding site for a milk protein binding factor (MPBF), a factor that has been shown to be abundantly present in both lactating sheep and mouse mammary gland (Watson, 1991). Potential binding sites for the MPBF have been identified in a large number of milk protein genes and therefore it has been speculated that this protein may be a mammary gland-specific transcription factor with an essential role in milk protein synthesis. In this light, the fact that the MUC1 gene is expressed at its highest levels in both the lactating mammary gland and carcinomas is an intriguing observation. It will be of great interest to determine whether or not the 45 kDa protein identified by Abe, 1993, is a protein factor generally present in tissues that express MUC1, i.e. both normal and carcinoma, whether the factor is present only in cells of the mammary gland that express MUC1, or whether this factor is carcinoma-specific and is responsible for the general elevation of expression of MUC1 observed in carcinomas.

In addition to the possibility of the 45 kDa protein identified by Abe, 1993, being responsible for the up-regulation of expression of the MUC1 gene in carcinomas, a soluble factor capable of stimulating the production of MUC1 by human colon carcinoma cells has recently been identified (Irimura, 1990). This protein, designated mucomodulin, has an apparent molecular mass of 20 kilodaltons and is expressed by fibroblastic cells. Mucomodulin was observed to stimulate the production of MUC1 in greater than 70% of colon carcinoma cell lines tested. At this time it is not known if mucomodulin represents a factor that interacts directly

with the upstream sequences of the MUC1 gene or whether it operates via an intermediate.

Recently, factors influencing the expression of the mouse Muc-1 gene in mouse mammary epithelial cell lines and in the differentiating mammary gland during pregnancy and lactation have been identified (Parry, 1992). These studies indicated that expression of Muc-1 builds from a minimal level in the virgin mammary gland and increases dramatically in the late stages of pregnancy. Expression levels reached a maximum at between 18 days of gestation and the first couple of days of lactation. CID-9 mammary epithelial cells expressed Muc-1 at the highest observed levels when cultured in the presence of insulin with prolactin and hydocortisone and when cultured on a basement-membrane-like extracellular matrix. A 5 to 10 fold increase in expression levels of Muc-1 was observed when cells were cultured on EHS matrix as opposed to tissue culture plastic. A role for the presence of an extracellular matrix in the regulation of gene expression has also been demonstrated for a number of other milk proteins, including the caseins (Li, 1987; Barcellos-Hoff, 1989). The culture of mouse mammary epithelial cells on EHS matrix resulted in them becoming hormone-responsive. Parry, 1987, demonstrated that permitting hormones to interact with proteins present at the basal surface of the cells is crucial for the secretory differentiation of mammary epithelial cells.

It appears, therefore, that the regulation of expression of the MUC1 gene is controlled by a combination of several factors, including tissue-specific promoters, enhancers, repressors and hormone responsive factors, and by the presence of an extracellular matrix-like substratum. It is hoped that through the continued investigation of different aspects of the regulation of expression of the MUC1 gene, an overall picture will develop of how the various factors interact with each other and of the role each plays in the regulation of expression of this gene.

In the preceeding pages, an overview has been presented of the MUC1 gene, its regulation and its protein product. Although, as can be seen, an enormous amount of research has been carried out into the biology of this molecule, one of the biggest questions remains, what is the function of this molecule in both the normal and tumour-associated situations? Numerous functions have been proposed for this molecule,

and the following couple of pages will attempt to collate some of the functions that have been attributed to MUC1.

The protein product of the MUC1 gene is present on the apical cell surface of the majority of simple secretory epithelial cells. As described previously, the MUC1 protein is extensively glycosylated through O-glycosidic linkage to serine and threonine, with as much as 50 to 90% of its molecular mass made up of oligosaccharide side chains. According to Jentoft, 1990, an extensively O-glycosylated polypeptide of 28 amino acids is approximately 7 nanometers (nm) in length. This would imply that the human MUC1 protein may extend as much as 200-500 nm above the cell surface, far above all the other membrane-associated proteins. Indeed, electron microscopic studies of the human MUC1 protein indicated that the molecule is present as a single strand with no discernible tertiary structure and had an average length of 270 nm (Bramwell, 1986). Ligtenberg, 1992b, demonstrated that transfected melanoma cell lines expressing high levels of MUC1 do not aggregate as efficiently as their control cells. Similarly, in mouse L-cells, super-transfection of cells previously transfected with an expression vector for E-cadherin, the homotypic cell adhesion molecule, with vectors expressing MUC1, resulted in the blocking of E-cadherin promoted cell adhesion and in many cases also resulted in the loss of adherence of the cells to various matrix components (Wesseling, 1992). The anti-adhesion properties observed for MUC1 are most probably a reflection of its large size. Most cell surface proteins remain within the boundaries of the glycocalyx which is approximately 10 nm thick, whereas the MUC1 protein may extend up to 500 nm above the cell surface (Fig. 1.11). Presumably, a high concentration of the MUC1 protein at the cell surface of the transfectants masks or blocks cell-adhesion molecules. The presence, on the molecule, of a high concentration of terminal sialic acid which is both bulky and carries a strong negative charge would presumably also contribute to the anti-adhesion properties of the MUC1 protein. Treatment of MUC1 transfectants with neuraminidase, an enzyme which removes terminal sialic acid residues, resulted in only the partial restoration of cell aggregation ability (Ligtenberg, 1992b). MUC1 may, therefore, function as an anti-adhesion molecule; high expression of MUC1 on the cell surface could have the same effect as the down modulation of cell-adhesion proteins such as E-cadherin. One can

envisage that this type of property could play a role in both normal and cancerous tissue where MUC1 is expressed.

The detection of Muc-1 in the mouse embryo at the apical surface of the developing lumens during epithelial organogenesis suggests that the mucin may be playing an important role in the development of these organs. In particular, it can be imagined that a concentration of an anti-adhesion protein, such as Muc-1, on the apical surface of an aggregation of differentiating epithelial cells may have the effect of repelling adjacent cells. It can be envisaged that one of the ways cells expressing Muc-1 in a polarised manner could escape from the repulsion effects of neighbouring cells would be through the formation of a lumen into which the extracellular portion of the proteins would project. The human MUC1 protein has been shown to be associated with elements of the actin cytoskeletal network, presumably through interaction of its 69 amino acid cytoplasmic tail (Parry, 1990). These studies also demonstrated that the MUC1 protein is found exclusively on the apical cell surface even in the absence of tight junctions. Thus, it is possible that early in the development of the respective organs where the mucin is expressed, before the formation of lumens, the mucin is restricted to the apical cell surface where it may play an important role in the subsequent formation of lumens and glandular differentiation.

Overexpression of MUC1 has been associated in particular with malignant metastatic carcinoma (Irimura, 1991). In many carcinoma cells, polarisation of the epithelial cells is lost and the MUC1 protein can be detected on all cell surfaces, including those facing the stroma and adjacent cells. Under these circumstances, the anti-adhesive property of MUC1 may have the effect of destabilising cell-cell and cell-substratum interactions, thus promoting the disaggregation of a tumour site, leading to tumour spread and metastasis. Electron microscopic studies of sections of human breast carcinoma have revealed that in regions where the mucin protein is particularly abundant the adjacent membranes make no direct contact (Hilkens, 1984). One of the steps thought to be involved in the invasive spread of tumours is the lowering of intercellular adhesion through the down-modulation of E-cadherin (Behrens, 1989; Vleminckx, 1991). The increase in expression of MUC1 in non-polarised epithelial cells may have the same effect as the functional down-modulation of cell adhesion molecules.

In addition to the protein playing a possible role in the initial invasive spread of carcinomas, the presence of large amounts of the protein on the surface of metastatic carcinoma cells may also effectively shield them from immune surveillance. This is thought most likely to be the result of masking of cell-surface antigens involved in immune recognition. Tumour-reactive cytotoxic T cells (CTLs) can be isolated from tumour-draining lymph nodes of patients with breast and pancreatic cancer. These CTLs have been found to recognise the MUC1 protein as their target, in a non-MHC restricted manner (Barnd, 1989). The T cell epitopes have been identified as being carried on the tandem repeats (Jerome, 1991). It is thought that numerous repetitive epitopes allow the MUC1 protein to crosslink the T cell receptors directly, thus activating CTLs independently of antigen processing and presentation in the context of self MHC molecules. The previously mentioned tumour-specific antibody, SM3, was able to significantly inhibit lysis of tumour cells by cytotoxic T cells. More significantly, normal breast epithelial cells expressing MUC1 on their cell surfaces were not lysed by these CTLs nor could they act as "cold target inhibitors" of lysis of tumour cells (Jerome, 1991). This suggests that the CTLs are recognising tumour-associated MUC1 only. In this respect, it may seem that a high level of expression of MUC1 by tumours could be disadvantageous to the tumour cells, but this is not necessarily so. In patients with breast and pancreatic carcinoma, the MUC1 protein can be detected in the circulation (Burchell, 1984; Metzgar, 1984; Hayes, 1985; Hilkens, 1986; Tsarfaty, 1988), and free MUC1 protein, produced by tumours, has been demonstrated to inhibit the CTL lysis of target cells (Barnd, 1989). High levels of circulating MUC1 might, therefore, block the specific T cell activity and thus aid the cells in escaping from immune surveillance. Thus, the cancer-associated functions of the MUC1 protein are two-fold; firstly, overexpression of MUC1 in non-polarised carcinoma cells may be associated with the lowering of cell-adhesion and the initial tumour spread and, subsequently, high expression of MUC1 by tumour cells may effectively shield them from immune surveillance and aid in their spread.

A variety of other functions have been proposed for the MUC1 protein, including possible roles for the milk mucin in the neonate gastro-intestinal tract. Milk mucin has been shown to inhibit the adhesion of S-fimbriated *Escherichia coli* to epithelial cells (Schroten, 1992). S-fimbriae are a common property of *E. coli* strains causing sepsis

and meningitis in neonates (Korhonen, 1985). Conversely, milk mucin has been demonstrated to enhance the growth of the normal, constitutively present gut flora in the intestine of newborn infants. In addition to numerous possible functions attributed through MUC1 in milk, MUC1 present on the endometrial lining may even play a role in contributing to the receptiveness of the female reproductive tract to embryo implantation (Smith, 1989; Graham, 1990; Hoadley, 1990; Kimber, 1990). It can be seen, therefore, that numerous functions have been attributed to the MUC1 protein.

Through the advent of transgenic mouse technology and the subsequent development of pluripotent mouse embryonic stem (ES) cells it has become possible to create designed mutations in specific genes within the mouse genome by gene targeting. The derivation of mouse ES cells carrying such a mutation can subsequently lead to the generation of mice both heterozygous and eventually homozygous for the desired mutation. This technology is thus a powerful way of analysing the possible function of a particular protein in both mouse embryo development and in the adult mouse. The generation of mice carrying a mutation in the mouse Muc-1 gene could lead to a better understanding of the role of this molecule in the tissues in which it is expressed. The following section describes some of the parameters involved in gene targeting in embryonic stem cells.

1.2 Gene targeting, by homologous recombination, in mouse embryonic stem cells

Within the last decade, science has seen transgenic mouse technology become a sophisticated assay system to study a wide range of diverse biological problems. Transgenic mouse technology has provided a powerful tool for studying the effects of ectopic gene expression within the intact animal. However, it has only been relatively recently, through the isolation of pluripotent mouse embryonic stem cells (Evans, 1981; Martin, 1981; Robertson, 1986; Gossler, 1986) and the development of gene targeting procedures for 'knocking-out' genes (Smithies, 1985; Thomas, 1987, Mansour, 1988), that it has become possible to specifically modify any endogenous gene of interest.

Mouse embryonic stem cells are permanent, pluripotent, euploid cell lines isolated from the inner cell mass of 3.5 day mouse blastocysts (Evans, 1981; Martin, 1981). These cells, when maintained under the correct culture conditions, retain the ability, when injected back into a host blastocyst, of contributing to the formation of all tissues of the normal animal, including its germ cells (Robertson, 1986; Gossler, 1986). Such an animal is described as a chimaera due to its development from two different origins. The development of permanent mouse cell lines which could be utilised as a route into the mouse germline was a necessary prerequisite for the development of the gene targeting strategy (Fig 1.21). The advantage of the cell line approach to gene targeting lies in the ability to screen potentially millions of transformed ES cells in culture. The desired clones can thus be identified prior to injection into the host blastocyst and the subsequent derivation of chimaeric mice.

Gene targeting relies on the selection of particular ES clones carrying a specific mutation in the gene of interest. From these cells, carrying the designed mutation, germline chimaeras can be derived which can be utilised to establish a permanent mouse line heterozygous for the mutation. Animals heterozygous for the mutation can then be interbred in order to create mice homozygous for the mutation in the specific gene being analysed (Fig. 1.21). Mice lacking a particular protein, or possessing a designed mutation in that protein may, as a result, either i) exhibit specific phenotypic changes that can be attributed to the loss of function of the protein being analysed, ii) exhibit no phenotypic change or a cryptic phenotype which may be masked by compensation of function by other similar proteins, or iii) fail to develop properly due to the mutation being an embryonic lethal. This technology, therefore, represents a means by which the function of particular proteins can be studied *in vivo* and also by which mouse models of specific human inherited genetic diseases can be generated. Such mouse models will be invaluable in testing novel therapeutic strategies, including gene-therapy based approaches.

ES cell lines currently in use in labs involved in gene targeting based research are derived from mouse blastocysts of the strain 129 (chinchilla). As such, host blastocysts utilised to generate chimaeras are usually either C57Bl (black coat colour) or Balb/c (white coat colour), which give rise to agouti/black and agouti/white coat colour chimaeras,

respectively. Coat colour chimerism represents a quick method for the identification of mice in which there has been ES cell developmental contribution. In addition to all ES cell lines being derived from 129 blastocysts, the ES cell lines in use have the male chromosome constitution (XY) (Bradley, 1984; Doetschman, 1985; Robertson et al., 1986; Hooper, 1987; McMahon, 1990). The advantages in the use of a male (XY) ES cell line over a female (XX) ES cell line are several fold. The XY constitution appears to be more stable in prolonged culture when compared to the female karyotype. Also, it is only the XY cells that give rise to functional sperm. The generation of male chimaeras which can potentially be back-crossed to give rise to many more offspring than female chimaeras is preferable. Thirdly, the microinjection of male ES cells into female blastocysts can, if the male ES cells contribute to the formation of the germline, result in the phenomenon of sex-reversal, the resultant mouse being male. In such animals, since only XY cells can form functional sperm, the entire germline will be derived from the ES cells. However, it has been demonstrated that these sex-reversed chimaeras are generally less fertile than chimaeras which formed from male blastocysts, and may even be sterile (Patek, 1991).

The maintainence of the pluripotentiality of embryonic stem cells depends upon the correct culture conditions. In general, mouse ES cells can be maintained in their pluripotent state by culture on a layer of mitotically inactivated fibroblast cells, described as feeder cells. Feeder cells synthesise factors that are released into the medium that specifically inhibit the differentiation of embryonic stem cells (Smith, 1987). As ES cells have the potential to give rise to all tissue types of the developing mouse embryo, they are capable of differentiating *in vitro* to form a wide range of cell types. Once differentiated, however, they lose the ability to contribute to chimaera formation. Recently, a specific factor has been identified, leukaemia inhibitory factor (LIF) (Williams, 1988), which, when added above a threshold concentration to the culture medium, helps prevent the differentiation of ES cells. Indeed, a number of germline chimaeras have been obtained through ES cells cultured for part or all of their culture life, prior to blastocyst microinjection, in medium with LIF only (no feeders) (Pease, 1990; Mortensen, 1991; Fung-Leung, 1991; Grusby, 1991; Miller, 1992).

The *in vitro* differentiation of mouse embryonic stem cells proceeds through a series of well-defined stages (Doetschman, 1985). ES cells, when cultured in bacteriogical dishes in suspension, and in the absence of a feeder layer and/or LIF, differentiate within three to four days to form structures described as simple embryoid bodies. These structures are characterised as a ball of cells surrounded by a ring of endoderm with a defined basement membrane (Reichert's membrane). Simple embryoid bodies will, if cultured further, differentiate a layer of ectoderm below the endodermal layer and become fluid-filled. These structures are referred to as cystic embryoid bodies. Embryoid bodies can be transferred to gelatinised tissue culture plates where they are able to attach. Soon after attachment, cells of a large variety of types differentiate and migrate out of the embryoid body. Differentiation is chaotic at this point, although certain cell types typically differentiate and appear earlier than others. Neuronal cells, skeletal, smooth and cardiac muscle, epithelial sheets, adipocytes and other cells can all be observed in a single dish. This *in vitro* differentiation process can be used to analyse the effect of targeted disruption on the function of a protein *in vitro* (Baribault, 1991; Mortensen, 1991) and also to verify the correct expression pattern of genetically created fusion proteins such as those in which the bacterial β-galactosidase (LacZ) gene is fused to the gene being targeted (Mansour, 1990).

The differentiation potential of ES cells can also be analysed through the sub-cutaneous injection of ES cells into athymic nude mice. In this instance, cells differentiate within two to three weeks to form teratocarcinomas made up of numerous tissue types. Sections of teratocarcinomas reveal tissues such as skeletal muscle, simple, stratified, keratinising and pigmented epithelium, neural tissue, cartilage and bone as well as numerous other tissues. Again, this type of system can be used to study the effect of targeted disruption of a gene on the function of the protein in question. Although this type of analysis has not been widely used to date, it could be effectively utilised to study the effect of a gene knockout on the function of certain types of proteins, in particular those thought to be involved in processes of tissue formation.

It is only comparatively recently that numerous groups have started reporting the germline transmission of targeted mutations. However, the process of obtaining germline transmission is still not

trivial and is certainly a time-consuming process. Therefore, a more rapid method for functional analysis, such as the nude mouse assay described above, may provide investigators with an initial answer as to the possible biological function of a protein. This type of analysis can be carried out in conjunction with the generation of mice lacking the protein being studied. It should be pointed out, though, that unless the gene being studied is X-linked, the nude mouse tumour assay of protein function requires the generation of ES cell lines in which both copies of the gene of interest are mutated (so-called double knockouts). This is generally carried out through two successive rounds of gene targeting, which require the construction of two different targeting vectors (te Riele, 1990; Mortensen, 1991).

Gene targeting by homologous recombination requires a knowledge of the exon-intron boundaries of the gene being studied. This allows the construction of specific targeting vectors containing fragments of the gene of interest with one or more selectable marker genes. Targeting vectors can be one of two types, described as either replacement or insertion vectors. Replacement (omega type) vectors typically consist of a selectable neomycin phosphotransferase (neo gene) cassette driven by a strong promoter such as tk (herpes simplex virus thymidine kinase) or Pgk-1 (mouse phosphoglycerate kinase-1) (McBurney, 1991) inserted into one of the exons of the gene being analysed. Thus, the neo cassette is generally flanked on both sides by fragments of the gene being studied. These fragments are referred to as the arms of homology, there being 5' and 3' arms of homology, respectively. Replacement vectors are described as such due to the fact that homologous recombination between the vector and the endogenous gene, through reciprocal crossover events on both arms of homology, results in the replacement of the corresponding sequence of the gene being targeted (Fig. 1.22). The insertion of the neo cassette into the coding domain of a gene is expected to create a disruption in the gene, such that a functional gene product is no longer synthesised. Such an allele is referred to as a 'null' allele.

Insertion vectors (O type) rely on the insertion of the entire vector into the target site through a single reciprocal crossing over event (Fig. 1.22). Recombination between the insertion vector and its target creates a duplication of part of the gene, and this is expected to disrupt gene function. However, it is possible that through splicing events, a

functional product could be generated from a gene that has been targeted by an insertion vector. Insertion vectors are thought to give a higher targeting frequency when compared to the equivalent replacement vector, and this is presumably due to the fact that these vectors require only a single recombination event as opposed to the double event required by replacement vectors (Hasty, 1991a). However, an independent study has indicated that the targeting frequencies observed for replacement and insertion vectors, containing the same respective sequences, are equivalent (Deng, 1992). To date, of the two types of vectors, the replacement type vector system remains the vector of choice as this system is designed to create null alleles.

In order to mutate a particular gene, the construct DNA must first be introduced into the ES cells where it may recombine with the endogenous gene. There are numerous methods for introducing DNA into ES cells, including calcium-phosphate precipitation (Gossler, 1986), retroviral infection (Robertson, 1986; Kuehn, 1987) and microinjection (Zimmer, 1989). However, the method that has been used almost exclusively to date is electroporation. Although electroporation is a relatively inefficient way of obtaining transformants it does have advantages. Firstly, it can be applied to a large number of cells simultaneously, for instance we typically electroporated 4×10^7 cells per electroporation. Secondly, the conditions of electroporation are such that they result primarily in single site, single copy vector insertions (Thomas, 1987). This fact is important when one is attempting to create a specific mutation at only one chosen site within the genome. Both replacement and insertion vectors are linearised prior to their electroporation into the cell. The stable integration of the vector DNA into the ES cell genome is screened for by selection with the cytotoxic drug, G418, which selects for cells expressing neomycin phosphotransferase.

The insertion of a selectable neo cassette into a gene not only disrupts the gene but also creates new sites for restriction endonucleases within the gene. The generation of novel restriction fragment length polymorphisms in a targeted allele can be utilised to detect the desired targeting event. Detection is carried out through restriction endonuclease digestion and Southern blotting of DNA from individual or pools of G418 resistant clones. The resultant blots are screened for the targeting

event through the utilisation of a DNA probe which recognises sequences flanking the arms of homology. The use of a flanking probe is critical for detection purposes as this type of probe will only hybridise to endogenous sequences present in the chromosome and will not hybridise to introduced vector sequences. Typically, if a clone has been specifically targeted, the presence of two hybridising bands of similar intensity will be observed. These bands represent the restriction fragment derived from the parental, wild-type allele, and the novel restriction fragment derived from the targeted allele, respectively (Fig. 1.23). Once clones possessing the desired targeting event have been identified, it is important that they be extensively analysed with a variety of DNA probes to determine that the structure of the targeted allele is as expected, with no rearrangements, and that there have been no additional insertions of the vector DNA elsewhere in the ES genome. Insertions elsewhere in the genome could potentially inactivate a second gene, thus making eventual phenotype analysis difficult.

The advent of polymerase chain reaction (PCR) technology has permitted the development of alternative methods for rapidly screening large numbers of antibiotic resistant ES clones (Kim, 1988). In order to use PCR for screening, the targeting vector is designed in such a way that a PCR primer present within the vector (usually within the neo gene) and a primer specific for endogenous chromosomal sequences not present in the vector, will only amplify a detectable PCR product if the vector integrates by homologous recombination. Homologous recombination will juxtapose the sequences containing the two primers, whereas random integration will not. Thus, a PCR product of the correct size should only be amplified from cells containing the correct targeting event (Fig. 1.23). The use of PCR as a screening strategy places a restriction on the size of one of the arms of homology of the vector, as the two respective primers should be separated by no more than 2.0 kilobase pairs.

PCR is an extremely sensitive method for screening and has been shown to be capable of detecting a single targeted ES cell in a population of 10^5-10^6 wild-type cells (Kim, 1988). However, in a typical targeting experiment a cell population consists of correctly targeted cells in a background of cells in which the vector has integrated at random. As such, the PCR detection of the targeting event becomes increasingly more

susceptible to artifacts. This type of artifact is presumed to occur if the polymerase fails to extend the neo primer through the vector sequence and into the neighbouring random DNA. The resultant partially extended molecule can serve as a primer on DNA from the wild type allele and thus artificially link the neo sequences to the flanking DNA, amplifying a product that is identical in length to that expected from a targeted allele.

PCR screening has an advantage over screening by Southern blots, in that it can be used to screen pools of colonies. For instance, where ten genomic DNA preparations may be required to screen ten clones by Southern blotting, the same ten clones can be screened by a single PCR reaction. However, although PCR screening represents a method by which large numbers of colonies can be rapidly screened, conditions need to be rigorously optimised to ensure that all targeted clones that may be present will be identified, and that non-targeted clones do not generate artifactual bands. The optimisation of PCR screening conditions is not trivial and, as the identity of PCR positive targeted clones also needs to be confirmed by Southern blotting, many groups have used Southern blotting with a flanking probe as the method of choice for the initial screening of ES colonies (Johnson, 1989; Mansour, 1990; Grusby, 1991; Mombaerts, 1991; Li, 1992). The use of Southern blotting as a screening procedure also means that the arms of homology of a targeting vector can be designed to be greater than 2.0 kilobase pairs in length.

One of the main difficulties initially encountered in gene targeting in ES cells was the high frequency of insertion of the targeting vector DNA at random sites in the genome as opposed to integrations at the homologous target site. Random non-homologous integration of the targeting vector was typically observed to be several orders of magnitude more frequent than homologous recombination. However, as more has been learned regarding the parameters involved in gene targeting, and their relative importance in terms of targeting frequency, targeting frequencies have risen sharply, with groups reporting frequencies of homologous recombination as high as 78% (te Riele, 1992).

Initially, one of the methods that was employed in an effort to enrich for cells in which homologous recombination had occurred was the Positive-Negative-Selection (PNS) system described first by Mansour,

1988. This system is based upon positive selection for an integration event, and negative selection against a random integration event. In this way, cells in which the targeting vector DNA has been integrated through homologous recombination are enriched. The gene that has been widely used for the purpose of negative selection is the thymidine kinase (tk) gene of herpes simplex virus. PNS targeting vectors typically consist of the two arms of homology of the gene being targeted, a neomycin cassette inserted between the two arms, and one or two copies of the HSV-tk gene flanking the 5' and/or 3' arms of homology (i.e outside of the region of homology) (Fig 1.24). During homologous recombination, the heterologous flanking sequences, including the HSV-tk gene, should be lost, whereas during random integration the HSV-tk sequences should integrate along with the rest of the vector. The presence of an active HSV-tk gene can be selected against using the nucleotide analogue gancyclovir (GanC). Cells expressing HSV-tk are able to metabolise GanC, which upon phosphorylation is integrated into the DNA leading to the termination of DNA synthesis and concomitantly to cell death. The endogenous mammalian tk gene has a lower affinity for GanC than HSV-tk, and thus, if the optimal concentration of GanC is used, cells that have integrated the HSV-tk gene through random non-homologous integration will be selected against.

The use of this type of system appeared to be a way around the high levels of random integration that were being encountered, and enrichment factors achieved using PNS vectors and double selection with G418 and GanC were initially reported to be in the range of 1000-fold (Mansour, 1988; Johnson, 1989; Thomas, 1990). However, the continued use of this type of system by numerous groups has shown that a more typical enrichment factor obtained using the HSV-tk gene and GanC is in the range of 2 to 5-fold (Mombaerts, 1991; Rudnicki, 1992; Li, 1992; Lee, 1992), although higher enrichment factors are occasionally reported. It appears that the main reason for the disappointing enrichment factors observed with this system, is the loss of activity of the tk gene. Loss of activity of tk is thought to occur by deletion of the ends of the targeting vector prior to its integration into the chromosome (Mansour, 1988). To partially circumvent this problem, targeting vectors have been designed in which there are two copies of the negatively selectable marker gene, one at each end of the two respective arms of homology.

Other genes that have been used as negatively selectable markers are the diptheria toxin A chain (DT-A) (Yagi, 1990), and in the slime-mold, *Dictyostelium discoideum*, a gene encoding a lethal tRNA stop codon suppressor has also been been utilised (Morrison, in press). The expression of both of these genes, in cells in which the vector has integrated randomly, is directly toxic to the cell and does not require the addition of any other agents to the culture medium.

In the past, genomic clones for the majority of mouse genes have been isolated from Balb/c genomic libraries as these libraries were commercially available. Thus, early targeting vectors were often constructed from Balb/c-derived genomic DNA. Recently, it has been shown that the use of targeting vectors made with DNA that is isogenic with the ES cells, i.e. 129-derived genomic DNA, increases gene targeting frequencies by a factor of 10 to 20-fold (te Riele, 1992). It appears that when Balb/c derived constructs are utilised, base sequence mismatches between the donor (Balb/c) and the recipient (129) target DNA are enough to significantly reduce the targeting frequency. The authors reported the detection of numerous strain specific differences between the retinoblastoma (Rb) gene sequences of Balb/c and 129 mice. These differences included base-pair substitutions, small deletions of up to 6 base pairs, and the presence of a polymorphic CA repeat in the Balb/c derived sequence that was absent in the 129 derived sequence. Indeed, the longest stretch of perfect homology detected between the two sequences was only 278 nucleotides.

The fact that base pair mismatches between donor and recipient DNA is one of the factors responsible for a low frequency of homologous recombination in ES cells is not surprising, as it has previously been demonstrated that base pair mismatches have a strong effect on genetic recombination in bacteria (Shen, 1986). In bacteria, the exchange of information between homologous regions of DNA is edited by mismatch-repair proteins, whose function is to abort or inactivate heteroduplex intermediates containing excessive numbers of mismatched base-pairs (Rayssiguier, 1989). In bacteria, a reduction in homology from 100% to 92% was observed to result in a decrease in recombination frequency of up to 45-fold (Shen, 1986). Similarly, it has been shown that the efficiency of intrachromosomal recombination in mammalian cells also decreases as a function of percentage homology

between two sequences (Bollag, 1989) and can be sensitive to even a single mismatch within a 1 kilobase pair interval (Letsou, 1987).

In addition to increasing targeting frequency through the use of isogenic-derived constructs, there are several other ways in which it has been shown that targeting frequencies can be increased. The most direct way of increasing targeting frequency is through an increase in the length of homology used in the targeting vector. An increase in extent of homology results in an exponential increase in targeting frequency (Deng, 1992). However, it appears that the mammalian homologous recombination machinery saturates at approximately 14 kilobase pairs of homologous sequence (Deng, 1992).

Genes that are known to be expressed in ES cells can be targeted at high frequency utilising constructs containing neomycin genes without a promoter, so-called promoterless neo genes. In these constructs the neomycin phosphotransferase gene could confer resistance to G418 only if expression of the neo gene was activated by juxtaposition to a random active cellular transcriptional promoter by random non-homologous integration, or by transcription initiated from the specifically targeted endogenous gene after integration into the locus by homologous recombination (Charron, 1990; Schwartzberg, 1990; Stanton, 1990; Bernelot Moens, 1992). This type of targeting vector has been shown to give frequencies of homologous recombination as high as 25% (Charron, 1990). It has also been shown that the use of this type of targeting vector is not neccessarily restricted to genes that are highly expressed in ES cells. For instance, expression of the the Hox-1.3 gene can only be detected by reverse-transcriptase PCR in ES cells. Nevertheless, the use of promoterless neo Hox-1.3 targeting vectors resulted in a significant enrichment in homologous recombination (Jeanotte, 1991) in these cells.

The gene targeting strategy has reached a level whereby it is now feasible to create mutations in any cloned gene. Indeed, numerous papers have been published over the past four years, reporting the creation of mice carrying mutations in Hox genes (Chisaka, 1991; Lufkin, 1991; Chisaka, 1992; Le Mouellic, 1992), in genes encoding growth factors (Lee, 1992), in proto-oncogenes (Thomas, 1990; McMahon, 1990; Soriano, 1991; Tybulewicz, 1991; Bernelot Moens, 1992; Stanton, 1992; Schwartzberg, 1991; Stein, 1992; Appleby, 1992), tumour suppressor genes (Matzuk, 1992;

Donehower, 1992), genes encoding extracellular matrix proteins (Saga, 1992), genes for transcription factors (Rudnicki, 1992; Braun, 1992), genes that are associated with specific human genetic diseases (Snouwaert, 1992; Tybulewicz, 1992) and numerous others.

The technology is now entering a new phase where it will now be possible to interbreed mice carrying a targeted mutation in one particular gene with mice carrying a targeted mutation in another gene. These kind of experiments may help to deduce functional relationships between specific proteins involved in complex pathways. The creation of mouse models for genetic diseases such as cystic fibrosis (Snouwaert, 1992), Gaucher's disease (Tybulewicz, 1992) and osteopetrosis (Soriano, 1991) through gene targeting will be crucial in designing novel therapeutic strategies for these diseases. The development of techniques for introducing subtle mutations into specific genes by the "In-Out" or "Hit and Run" gene targeting methods (Valancius, 1991; Hasty, 1991b) will make the creation of precise models of specific human genetic diseases a reality. These techniques allow the incorporation into a gene of a mutation as small as a single base transition or transversion, or the deletion of a single nucleotide. Thus, it is now possible to create mouse models that are directly representative of a genetic disease as it is encountered in humans.

As advancing technology makes the feasibility of mutating any cloned gene more of a reality, we should become wary of a situation in which genes are knocked out without a first consideration of the biology of the system being studied. It is hoped, therefore, that future gene-knockout experiments will be designed in such a way that the generation of mice mutant for a particular protein is not viewed as the ultimate goal of the experiment, but rather the starting point for further investigation. In conclusion, ES gene targeting technology has reached the point where we are no longer restricted by what can be accomplished, but rather by a lack of understanding of the basic biology of what has already been accomplished.

1.3 Minisatellites and mucin genes

The human genome is scattered with hypervariable minisatellite sequences: regions of DNA consisting of tandem repeats of a short base sequence that can show extensive variation in the number of repeats, leading to multi-allelic variation and high degrees of heterozygosity (Jeffreys, 1985a). Human minisatellites typically consist of multiple copies of tandem-repetitive sequences that share a common 10-15 base-pair 'core' sequence similar to the generalised recombination signal, chi (χ) of *E. coli*.

Due to their repetitive nature, many minisatellites are highly polymorphic, as a result of allelic variation in repeat copy number. Such loci are described as variable number tandem repeat loci or VNTRs. The vast majority of VNTR sequences that have been identified are non-expressed sequences. This fact is important in terms of selection pressures that may operate on tandem-repetitive arrays. If a sequence does not directly code for protein, then fluctuations, either up or down, in repeat unit numbers will presumably not be subject to strong selection pressure (Gray, 1991). The VNTR regions of mucin genes represent a small sub-set of VNTR sequences that actually code for an expressed protein product. Presumably, as these sequences code directly for an expressed product, fluctuations in repeat number will be subject to selection pressure. Mucin genes, therefore, represent a unique group of genes in which the evolution of expressed VNTR domains can be studied. The relative contribution of both neutral random mutations and directly selectable mutations can be considered. This section will consider the characteristics of minisatellite sequences, their evolution and their possible function. The mucin VNTR domains will then be considered as a unique class of VNTR sequences. In addition, we will consider the processes that may have been integral in leading to minisatellite-like polymorphism in the majority of mucin genes.

Several lines of evidence have suggested that minisatellite sequences may play a role in recombination. The repeat sequence of a sub-set of minisatellites share a common 10-15 base-pair sequence that shows high homology to the generalised recombination signal, chi (χ), of the bacterium *Escherichia coli* (Jeffreys, 1985a). In addition, minisatellites show preferential localisation to the proterminal regions of human

autosomes and the pairing region of the human sex-bivalent (Royle, 1988). Chiasmata are also observed to be preferentially localised to these regions, suggesting that minisatellite sequences may play a role in the pairing and recombination process during meiosis. Indeed, *in situ* hybridisation of a human minisatellite core sequence to human meiotic metaphase I preparations showed clustering of grains principally at or around chiasmata (Chandley, 1988). In addition, a sequence sharing high homology to the minisatellite core sequence has been identified at the MHC meiotic recombination hot-spot in the mouse (Steinmetz, 1987). Thus, taking all these lines of evidence into consideration it has been proposed that minisatellite sequences may represent hot-spots or even the sites for homologous recombination during meiosis.

A DNA probe based on the core minisatellite sequence is able to detect numerous highly variable loci simultaneously, providing an individual-specific DNA fingerprint (Jeffreys, 1985b). The most unstable VNTR loci, such as the human MS1 VNTR locus, exhibit a mutation rate to new length alleles of as high as 5.2% per gamete (Jeffreys, 1988a). The generation of new length repetitive alleles, with novel numbers of repeat units, was originally proposed to be the result of unequal crossing over between alleles at meiosis. Slippage of repeat arrays relative to one another, followed by a recombination event within the array, generates two new alleles containing different numbers of repeats (Jeffreys, 1985a) (Fig. 1.31). Another way in which novel repeat array lengths could possibly be generated is through the slippage of repeat units relative to each other during DNA replication. Replication slippage (Levinson, 1987) is thought to play a major role in the evolution of short simple sequence repeats.

The formation of new length VNTR alleles through unequal exchange between homologous chromosomes at meiosis predicts that markers flanking the tandem repeats will also be recombinant following the mutation event. However, it has subsequently been shown that this is not generally the case (Wolff, 1989) and, therefore, it is now thought that the generation of new length alleles at VNTR loci occurs primarily through mechanisms of unequal sister-chromatid exchange at meiosis and/or mitosis and/or through a process of DNA replication slippage. Although unequal exchange between VNTR alleles on homologous chromosomes has been shown not to be the major mechanism through

which new length VNTR alleles are formed, it has recently been shown that the frequency of homologous recombination at the human minisatellite locus, MS32, is 700-fold higher than the mean observed for the genome as a whole (Jeffreys, 1991). This observation suggests that although the majority of novel length minisatellite alleles are generated through processes of unequal sister-chromatid exchange and/or replication slippage, minisatellite sequences may also be general hot-spots for homologous recombination within the genome. This finding revitalises earlier speculation that minisatellites may play an active role in processes such as homologue recognition, synapsis and meiotic recombination.

Recently, two groups have independently identified specific proteins that bind to tandemly repeated minisatellite sequences (Collick, 1990; Wahls, 1991). One of these proteins, Msbp-1, has been shown to be a sequence specific DNA binding protein that binds not to double-stranded DNA, but exclusively to single-stranded DNA (Collick, 1991). Other well characterised single-stranded DNA binding proteins, such as RecA (Dressler, 1982) and single-stranded binding protein (ssb) (Chase, 1986) of *E. coli*, and gene32 of bacteriophage T4 (Chase, 1986) have been shown to play important roles in recombination processes. Minisatellite binding proteins such as Msbp-1 might function in analagous ways. For instance, Msbp-1, could promote recombination by stabilising minisatellite DNA in a single-stranded conformation, thereby promoting strand-exchange in a manner similar to RecA. However, there have also been reports of proteins that are able to bind to specific single-stranded DNA sequences that do not appear to play a role in recombination processes. These proteins include those able to bind to yeast autonomously replicating sequence (ARS) elements (Schmidt, 1991), Oxytricha telomeres (Gottschling, 1984), and $(CT)_n$ microsatellites (Yee, 1991). Therefore, until formal proof of the role of minisatellite binding proteins, such as Msbp-1, in the recombination process has been demonstrated we should be wary of jumping to premature conclusions regarding the function of these proteins. Presumably, formal proof will require the isolation and characterisation of the genes coding for Msbp-1 and other minisatellite binding proteins.

The development of PCR based approaches in the analysis of VNTR loci has allowed major advances to be made in our understanding

of the biology of minisatellite sequences. In particular, PCR based analysis has revealed that within a repeat array there can be variation in the sequence of the individual repeat units (Jeffreys, 1990 and 1991). For instance, within the human minisatellite locus MS32, repeats can differ by a single base substitution which either creates or destroys a site for the restriction endonuclease HaeIII. Using the presence or absence of the HaeIII restriction site as the criterium, a repeat array can be specifically typed; repeat units possessing the site are represented by a 1, and those in which the site is absent are designated 0. For instance, the typing of a particular MS32 allele with 25 repeat units may yield the code 0001101110011100101101111. It has been calculated that this additional level of minisatellite variation provides a system in which potentially $>10^{70}$ allelic states can be distinguished (Jeffreys, 1990 and 1991).

In addition to providing a new level of minisatellite variation which will be invaluable for DNA profiling purposes, the analysis of variant repeat units can provide important information regarding the formation of novel length minisatellite alleles. Through PCR amplification and typing of MS32 alleles from both parental and offspring derived genomic DNA, this type of analysis has allowed the identification of specific sites within parental minisatellite alleles where an unequal crossover event, occurring during gametogenesis, has resulted in the formation of a novel length allele that is transmitted to the offspring. Although the average number of repeats at the MS32 locus is 200, these analyses have indicated that the majority of mutation events are extremely clustered and map to within the first 10 repeat units of the array. This would suggest the presence of a localised mutational hot-spot in this region of the repeat array. In addition, most new mutation events have been found to result in the gain of a small number (1-3) of repeat units, which would suggest a directional bias in the mutation process.

The development of methods to specifically amplify minisatellite alleles by the polymerase chain reaction has recently been utilised in order to investigate the evolution of two minisatellite loci. Utilising PCR primers directed to sequence flanking the VNTR domain of the two human minisatellite loci, MS1 and MS32, the corresponding sequence was specifically amplified from genomic DNA from a number of primate species (Gray, 1991). In this way it was demonstrated that where the human MS32 locus is made up of multiple (ranging from 12 to more

than 600 in any one allele) copies of a 29 base pair tandem repeat unit that exhibits 97.5% heterozygosity (Armour, 1989a), the MS32 locus of all other primates investigated (except the owl monkey) is invariant and consists of only two complete repeat units. One of these repeat units is identical to the 29 base pair unit that makes up the majority of the human MS32 locus. Similarly, where the human MS1 locus is made up of multiple copies (ranging from approximately 140 to 2500) of a 9 base pair repeat unit that exhibits heterozygosity greater than 99% (Royle, 1988; Wong, 1987), the MS1 locus of the chimpanzee and orangutan was observed to be invariant and the MS1 locus of the gorilla was observed to exhibit only a small degree of length variation. However, the corresponding locus of the Old world monkeys was found to exhibit VNTR polymorphism.

These findings suggest that the minisatellite loci MS32 and MS1 have attained extreme levels of length variability in man over a comparatively short period of evolutionary history, subsequent to the divergence of man and the great apes approximately 7 million years ago (Koop, 1986). In each case, however, the repeat unit that has been amplified at the human locus, is ancient in origin. Using known values for the mutation rate and average array length at the MS32 locus, computer modelling of the possible evolutionary expansion of the MS32 locus from an ancestral state of 2 repeats, revealed that on average one lineage in 250 will attain a transient phase of high (>200 repeat units) hypervariablity. Humans represent one such lineage. These models (Gray, 1991) predict that such lineages will have done so within the last 700,000 generations, a period of time similar to that elapsed since man diverged from the great apes. These models also predict that hypervariability at such a locus will be maintained only for a brief period of evolutionary history before the locus collapses to become monomorphic once again.

In contrast to the MS32 locus, hypervariablity and high repeat unit copy number has evolved on at least two separate occasions at the MS1 locus. One event appears to have occurred during recent human evolution and a second event within a predecessor of some or all of the Old World monkeys (Gray, 1991).

These evolutionary studies have demonstrated a number of important principles. Firstly, hypervariablity at minisatellite loci can be

rapidly established and rapidly lost in a particular lineage in evolutionary history; in other words hypervariability at minisatellite loci is transient. Secondly, minisatellites are amplified from an ancient sequence.

In the following section we will consider the evolution of the mucin genes in light of what is known of the evolution of non-expressed VNTR loci. The majority of mucin genes that have been cloned and characterised, do exhibit hypervariability of their respective repetitive domains (Fig. 1.32 and Table 1). The hypervariable domain of the mucin genes is comprised of expressed sequence, in each case making up the region of the gene encoding the central highly O-glycosylated portion of the protein. As the hypervariable domain of mucins is expressed, fluctuations in repeat unit number will be translated into fluctuations in the length of the corresponding peptide sequence. It has been suggested that large variations in the number of repeat units imply that the length of the molecule is not crucial to the function of the protein, but rather that the repetitive domain acts as a scaffold for carbohydrate attachment (Gendler, 1990a). However, although it has been suggested that the length is not crucial to function, presumably there are selection pressures that operate on mucin alleles such that there is a minimum and a maximum number of repeat units that are optimal for a particular mucin gene.

The number of repeat units that have been observed at the human MUC1 locus, for instance, ranges from a low of around 20 up to a high of 125 (Gendler, 1990a). However, a screen of 69 randomly selected Northern European individuals indicated that there are two common allele lengths corresponding to 40 and 66 repeat units, respectively (Gendler, 1990a). The most common allele length observed is one containing 40 repeat units. This might suggest that there is an optimal number of repeat units corresponding to an optimal MUC1 protein size.

Consider the processes that may be involved in the evolution of a hypervariable mucin gene. Presumably, an ancestral mucin gene would contain a region coding for an O-glycosylated domain of the protein. If this protein is a functional mucin-like protein, this domain would contain several potential attachment sites for O-linked carbohydrate. Similarly, from a knowledge of the amino acid sequences of highly O-glycosylated proteins we can say that the ancestral sequence would probably contain a high proportion of the amino acids serine, threonine,

proline, glycine and alanine. Assuming the O-glycosylated portion of a mucin is important for protein function, the ancestral mucin sequence would be likely to possess a region containing several potential O-linked attachment sites made up of similar sequence. In addition, the spacing of the respective O-linked attachment sites may be important, both for function of the protein and for the efficient attachment of carbohydrate to the protein core.

Several factors may, therefore, contribute to the formation of an ancestral sequence that is already reminiscent of a repetitive structure. The partial repetitiveness of the ancestral mucin sequence may mean that the sequence is predisposed to unequal crossover/replication slippage events. Since so many of the mucin genes that have been cloned and characterised are hypervariable, this seems to be likely. The human MUC1 consensus repeat sequence, like a sub-set of the non-expressed minisatellites, contains a sequence with homology to the generalised recombination signal, chi, of bacteriophage λ (Gendler, 1991). In *E. coli* and bacteriophage λ, chi functions as a signal for homologous recombination through binding the RecBCD protein complex which locally unwinds and nicks the DNA to create a free single-strand that is able to invade a homologous duplex to form a Holliday intermediate. Jeffreys and co-workers postulated that the minisatellite core sequence may act in a similar way to chi, to generate minisatellites (Jeffreys, 1985a). The minisatellite core sequence may stimulate nicking of the DNA duplex in its vicinity. DNA repair synthesis from the site of nicking and subsequent ligation of the nicked strand results in the duplication of the core sequence. Thus, a chi-like sequence present within mucin ancestral tandem repeats may have the effect of predisposing them to minisatellite-like expansion (Gendler, 1991).

The initial unpredictable duplication event would result in the generation of a larger allele possessing a larger O-glycosylated domain. The extended protein may convey a selective advantage and, therefore, it may spread comparatively rapidly through a population. For instance, the larger protein may better protect the epithelial surface from environmental agents such as dessication or bacterial infection. It has been suggested that for such a duplicated sequence, the probability of subsequent misalignment and unequal crossing over is high and expansion to form a tandem array is favoured (Smith, 1976). As more

length mutation events occur at the locus, the length of the repeat unit will be set and a consensus repeat sequence will be formed.

Once the locus has become hypervariable, the high frequency of recombination events occurring within the repetitive domain would have the effect of maintaining the sequence of the consensus repeat unit. This has been demonstrated for the non-expressed minisatellite loci. Those that exhibit the highest levels of variability have the most conserved consensus repeats. It has been proposed that a new base substitution that arises in one repeat unit may either rapidly spread to other repeat units through frequent unequal sister chromatid exchange and/or replication slippage events or, alternatively, may be lost from the array by the same mechanism (Stephan, 1989). This process has been referred to as crossover fixation. In other words, the fate of all point mutations is either to be purged or eventually spread to all the units of a repeat array.

If the evolution of hypervariability of the mucin gene was to occur early in evolutionary history, a number of predictions can be made. Firstly, as the initial recombination events, repeat selection and the ancestral sequence act together to specify the repeat unit length, a mucin gene that has become hypervariable early in evolution will most probably be found to possess a similar repeat unit length in diverse species. Secondly, as each lineage diverges through evolution they will be expected to maintain a similar repeat unit length, but eventually accumulate specific base substitutions such that each group of organisms will possess a characteristic consensus repeat sequence. Thirdly, if the position and sequence of the potential O-glycosylation sites within the consensus repeat is important to the function of the protein then selection processes will act on the sequence such that diverse species will be expected to maintain O-glycosylation sites. Similarly, if specific amino acids are important to the conformation of the protein then these will also be expected to be conserved. And lastly, other positions within the consensus repeat will be predicted to diverge within the limits of mucin repeat sequence, such that they will be one of a restricted set of amino acids. Conversely, if a particular mucin gene undergoes independent evolution to hypervariability in separate lineages then the sequence and length of the consensus repeat units isolated from different species may

be expected to differ widely. This may occur as a result of differential repeat selection and expansion.

Examples of both types of situation described above have been found in the involucrin gene of primates. Although this gene is not a mucin gene, it does possess an expressed variable repeat region (Eckert, 1986; Simon, 1991). It has been shown that the repetitive domains of the human and lemur involucrin genes have evolved through expansion of different segments of the ancestral genes. The respective human and lemur involucrin consensus repeats are 10 and 16 amino acids in length and share little homology (Tseng, 1988). In contrast, the repetitive domains of the involucrin genes of the New World monkeys and the higher primates have been derived from expansion of the same region of the gene, have the same consensus repeat length and are homologous (Tseng, 1989; Dijan, 1989a and 1989b). Of the mucin genes that have been cloned, a comparison of the partial sequence available for the repetitive portion of the rat Muc-2 gene (Xu, 1992), with the consensus repeat sequence for the human MUC2 gene (Gum, 1989) suggests that the repetitive domains of the Muc-2 genes in these two species may have evolved independently.

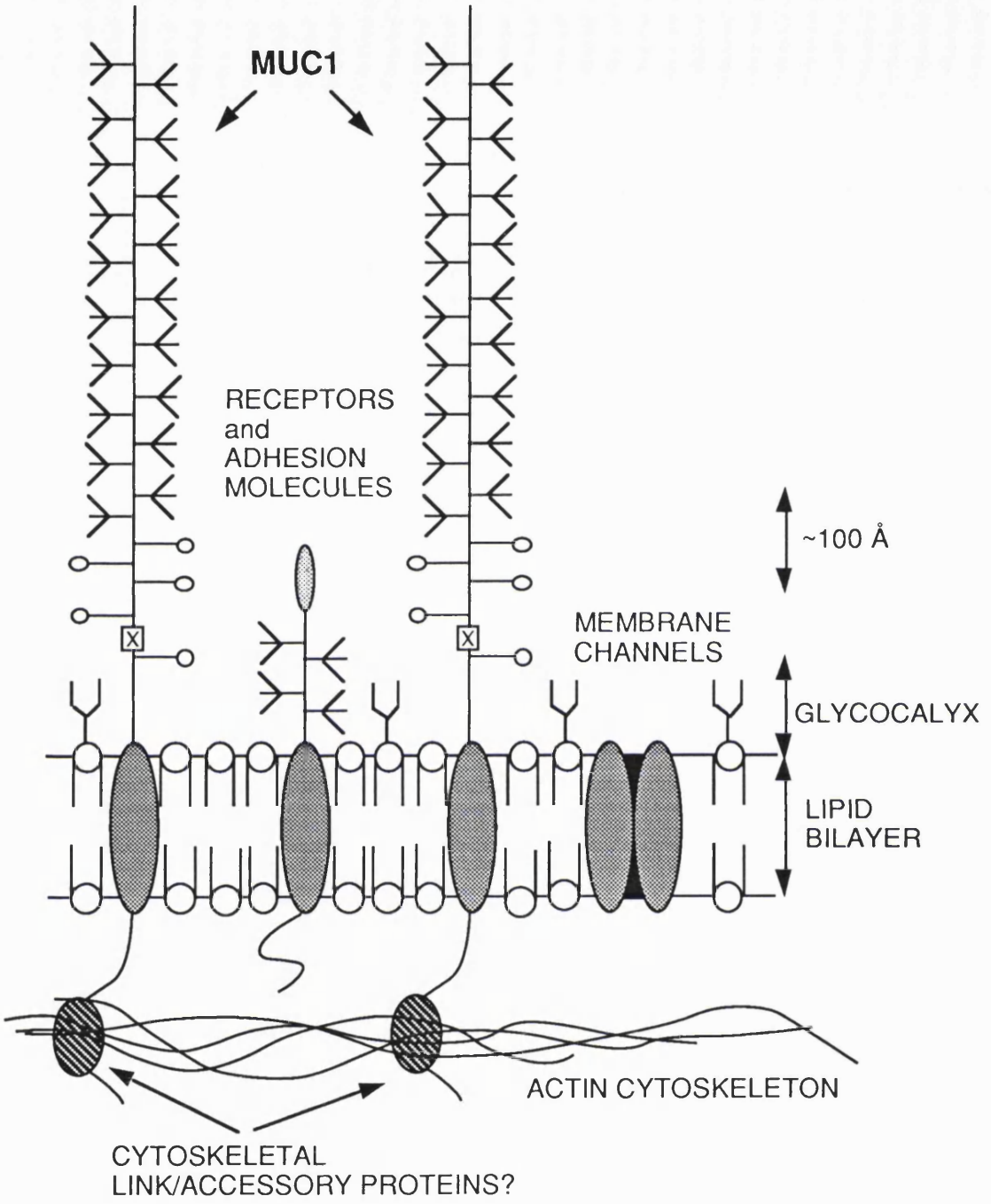
Over the last few years, numerous mucin genes have been cloned and characterised, yet it is only very recently that homologues of specific human mucin genes have been isolated (Spicer, 1991; Vos, 1991; Xu, 1992). The isolation and characterisation of mouse homologues for the respective human mucin genes will not only allow us to investigate the function of this group of glycoproteins, through for instance modulation of function experiments by gene targeting, and transgenic analysis, but will also allow us to begin an investigation into the evolution of this important class of minisatellite genes.

1.4 Aims of the project

Simple secretory epithelial cells express a variety of high molecular weight mucin glycoproteins. Attention has been focused on one particular mucin protein encoded by the human MUC1 gene due to the fact that it is overexpressed by the majority of human carcinomas. The human MUC1 gene encodes an integral membrane glycoprotein with the majority of its coding capacity made up of multiple copies of a 60 base pair tandem repeat encoding a 20 amino acid repeat motif. The gene also contains sequence encoding a 31 amino acid membrane-spanning domain and a 69 amino acid cytoplasmic domain. Variations in the number of repeat units per allele make this locus one of the few expressed hypervariable minisatellite sequences.

Numerous functions have been proposed for this molecule ranging from a role in the metastasis of carcinomas to a role in epithelial organogenesis, yet its precise function remains unclear. Recently, techniques have been developed to specifically mutate genes in mouse embryonic stem cells and subsequently to generate mice mutant for a specific gene. The initial aim of this project was, therefore, to isolate and characterise the mouse homologue of the human MUC1 gene. Once isolated, the mouse and human genes would be utilised to initiate a study of the evolution of the Muc-1 gene locus. In addition, genomic clones for the mouse Muc-1 gene would be used to specifically mutate the gene in mouse embryonic stem cells. In an effort to analyse the function of this protein, these cell lines would then be used to generate mice both heterozygous and homozygous for a targeted disruption in the Muc-1 gene.

Fig 1.11 Cartoon representation of the MUC1 membrane-associated mucin glycoprotein. The protein is restricted to the apical surface of the epithelial cells in which it is expressed. The majority of the protein is present on the outside of the cell and projects into the lumen of the epithelial glands. This portion of the protein is made up primarily of multiple copies of a 20 amino acid tandem repeat that is highly O-glycosylated. There are also five potential sites for N-linked glycosylation. The boxed cross marks the site of the potential proteolytic cleavage site described in the text. On the inside of the cell, the protein has a 69 amino acid cytoplasmic domain that may be linked to elements of the actin cytoskeleton. The large size of this membrane-associated glycoprotein may be reflected in its function, as it has been demonstrated to be capable of blocking adhesion of the homotypic adhesion protein E-cadherin.



= O-linked carbohydrate

= N-linked carbohydrate

= Glycolipid

= Potential proteolytic cleavage site

Figure 1.12 Expression of mouse Muc-1 during mouse embryonic development. Expression was spatially and temporally regulated and appeared to correlate well with the onset of epithelial differentiation and proliferation. Expression of the protein was not induced systemically, but according to the particular onset of epithelial differentiation in the respective organs (From Braga, 1992).

Muc-1 expression during mouse organogenesis

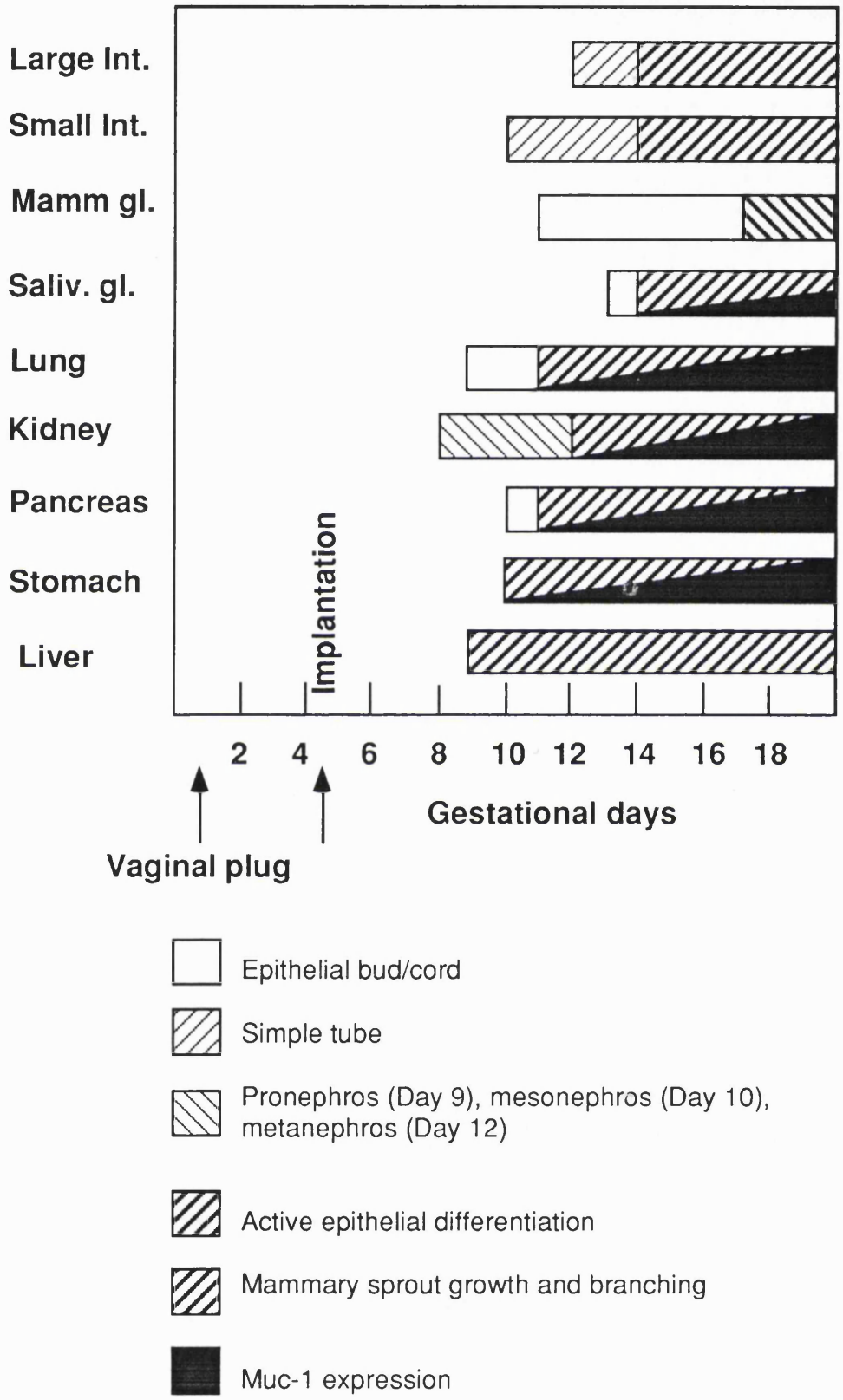


Figure 1.21 The gene targeting strategy. Mouse pluripotent embryonic stem cells are isolated and established from the inner cell mass (ICM) component of 3.5 day mouse blastocysts. ES cells are maintained in culture on a fibroblast feeder layer. A specific targeting vector construct carrying a selectable marker gene is electroporated into the ES cells and the cells are replated and selected in e.g. medium with G418. Correctly targeted cells are identified and expanded. Targeted cells can be injected into 3.5 day host blastocysts to derive coat colour chimaeras, or the differentiation of the ES cells can be assayed through sub-cutaneous injection into nude mice to form teratocarcinomas. Coat colour chimaeras are back-crossed against the host mouse strain e.g. C57BL/6 to determine whether or not there has been germline contribution of the ES cells. Germline transmission is screened for by the presence of agouti F_1 offspring. On average, 50% of agouti offspring will be heterozygous for the targeted mutation. Heterozygous mice are interbred in order to obtain homozygous mice lacking the protein being studied.

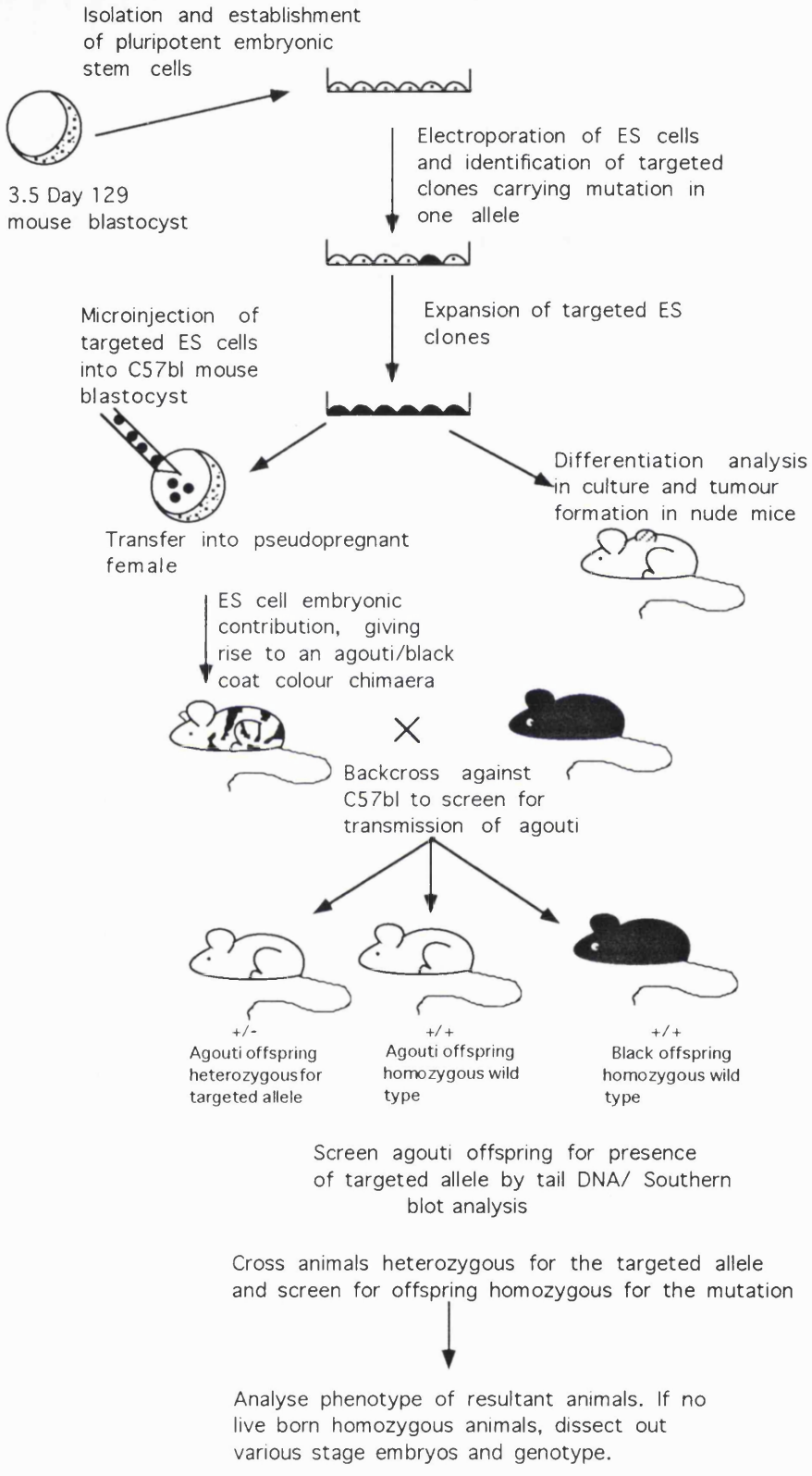

Isolation and establishment of pluripotent embryonic stem cells
3.5 Day 129 mouse blastocyst
Electroporation of ES cells and identification of targeted clones carrying mutation in one allele
Expansion of targeted ES clones
Microinjection of targeted ES cells into C57bl mouse blastocyst
Transfer into pseudopregnant female
Differentiation analysis in culture and tumour formation in nude mice
ES cell embryonic contribution, giving rise to an agouti/black coat colour chimaera
Backcross against C57bl to screen for transmission of agouti
+/-
Agouti offspring heterozygous for targeted allele
+/+
Agouti offspring homozygous wild type
+/+
Black offspring homozygous wild type
Screen agouti offspring for presence of targeted allele by tail DNA/ Southern blot analysis
Cross animals heterozygous for the targeted allele and screen for offspring homozygous for the mutation
Analyse phenotype of resultant animals. If no live born homozygous animals, dissect out various stage embryos and genotype.

Figure 1.22 Sequence replacement and insertion vectors. Black boxes numbered 1 through 7 represent the seven exons of the gene being targeted. Homologous integration of replacement vectors into the target site requires two reciprocal crossing over events on each respective arm of homology. In this way the targeting vector effectively replaces the endogenous sequence. The resultant targeted allele has part of its coding region disrupted by the insertion of a selectable marker gene, such as the neomycin resistance gene. Homologous integration of insertion vectors requires only a single crossover event in order to integrate the entire vector into the target site. Insertion of the vector results in the formation of a duplication at the target locus (modified from Deng, 1992).

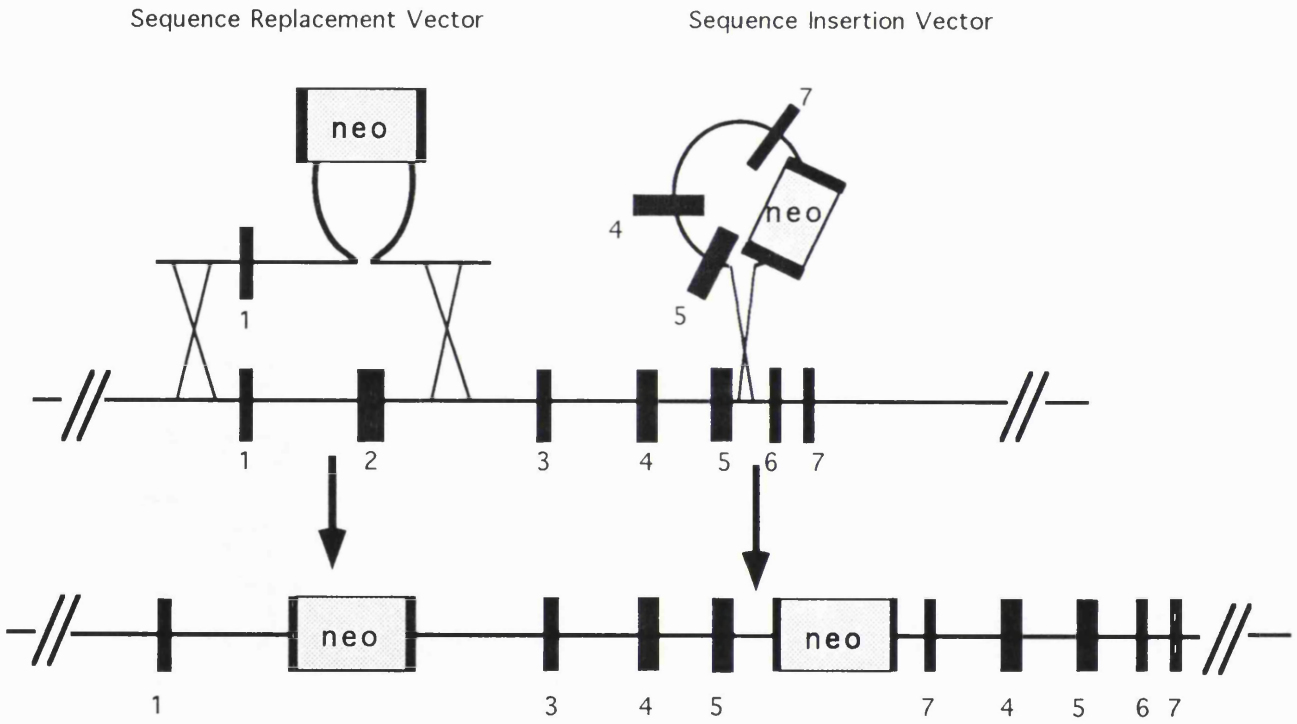
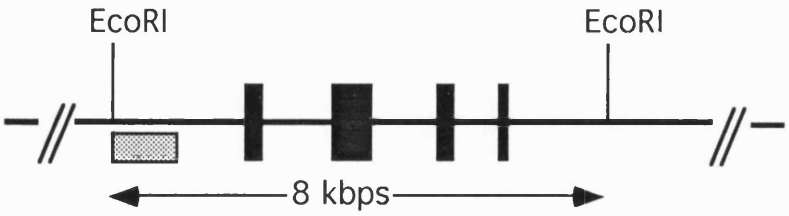

Sequence Replacement Vector
Sequence Insertion Vector
neo
neo
1
1
2
3
4
5
6
7
4
5
7
neo
neo
1
3
4
5
7
4
5
6
7

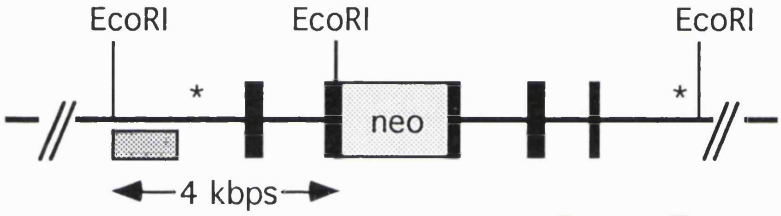
Figure 1.23 Screening for homologous recombinants: A) Screening by Southern blotting and hybridising with a flanking probe. Asterisks represent the 5' and 3' ends of the targeting vector arms of homology. Insertion of the selectable neomycin cassette has in this case resulted in the formation of a novel length EcoRI restriction fragment. A flanking probe, not present within the targeting vector, will normally hybridise to an 8 kilobase pair EcoRI fragment. However, if the gene is correctly targeted, insertion of the neomycin cassette carrying an additional EcoRI site alters the endogenous 8 kilobase pair hybridising fragment to a 4 kilobase pair hybridising fragment. Positive identification of a targeted ES cell clone, carrying one mutant and one wild type allele, is thus dependent upon the detection of hybridising fragments of 4 kilobase pairs and 8 kilobase pairs by Southern blotting (P = parental cell line DNA; T= correctly targeted cell line DNA). B) Screening by polymerase chain reaction (PCR): Homologous integration of the targeting vector DNA results in the juxtaposition of PCR primer sequences. PCR, with primers directed to sequence within the neo gene and to unique chromosomal flanking sequence, should specifically amplify a diagnostic fragment from cells carrying the correct targeting event.

A.

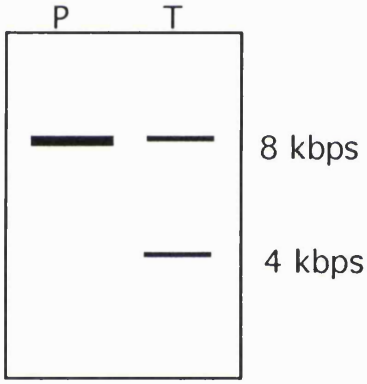
ENDOGENOUS GENE STRUCTURE



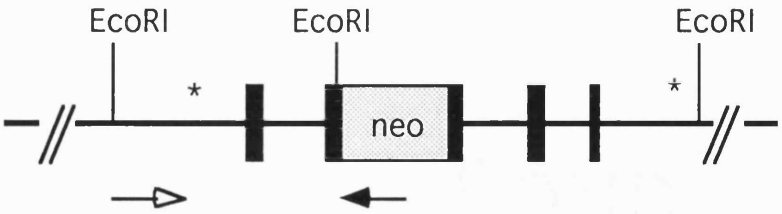
TARGETED GENE STRUCTURE



Southern blot hybridised to flanking probe



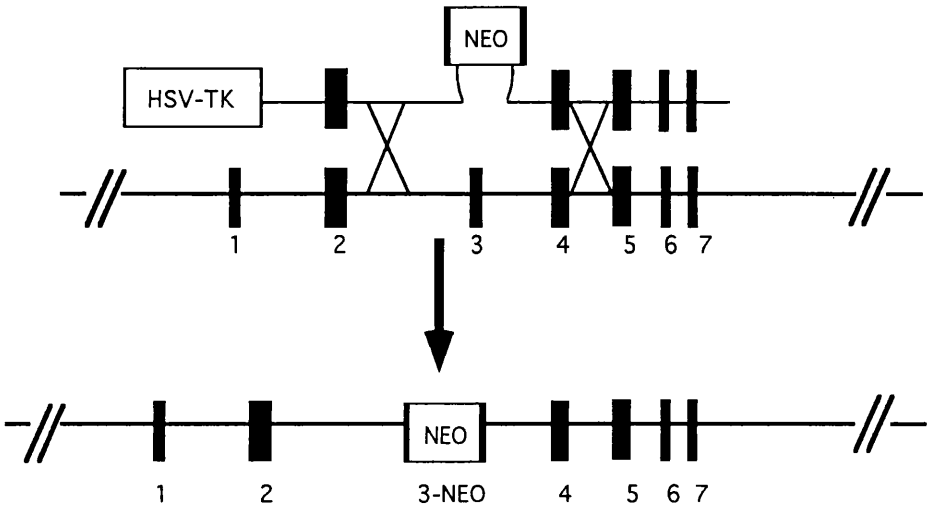
B.



PCR primers juxtaposed in targeted integration

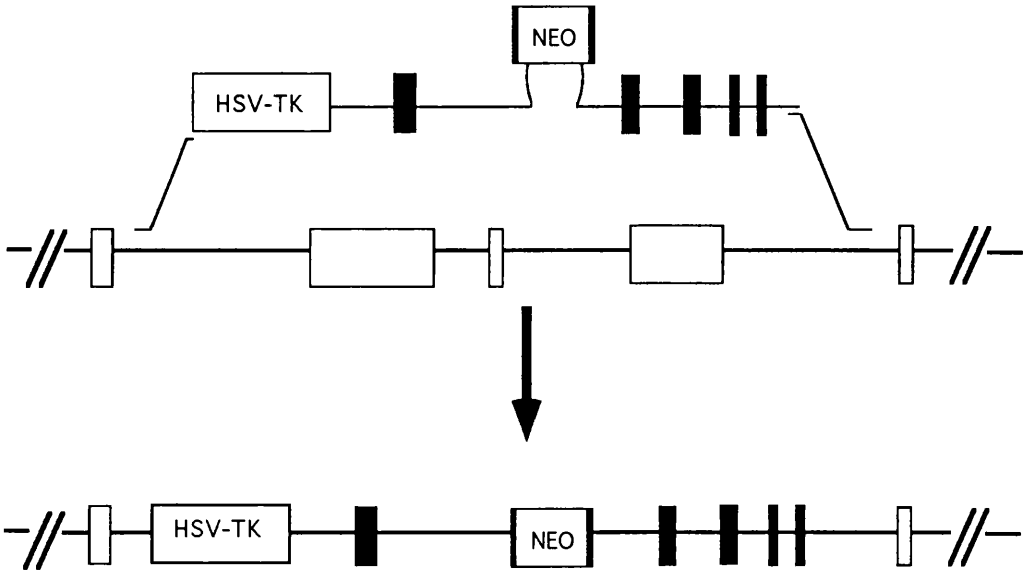
Figure 1.24 The Positive-Negative-Selection (PNS) system. The PNS targeting vector is constructed with two selectable marker genes. Filled boxes numbered 1 through 7 represent the seven exons of the gene being targeted. A selectable neo cassette is inserted within the region of homology of the gene being targeted, and a HSV-tk gene is placed outside the region of homology. Homologous integration of the targeting vector should result in the replacement of the endogenous gene with the neo-disrupted gene. Double reciprocal recombination will result in the loss of the HSV-tk gene and the gain of the neo gene. Conversely, random integration of the targeting vector at a non-homologous site within the genome should incorporate the entire targeting vector into that site. In this case, open boxes represent exons present within a non-homologous gene. If random integration occurs, both the neo gene and the HSV-tk gene will be inserted into the genome. Enrichment for homologous recombination is carried out by selection with G418, which selects for cells expressing the neomycin phosphotransferase gene (positive selection), and with gancyclovir (GanC), which selects against cells expressing the HSV-tk gene (negative selection) (modified from Mansour, 1988).

1. HOMOLOGOUS RECOMBINATION RESULTS IN INTEGRATION OF NEO AND LOSS OF HSV-TK



G418 RESISTANT + GANCYCLOVIR RESISTANT

2. RANDOM INTEGRATION RESULTS IN INTEGRATION OF ENTIRE VECTOR



G418 RESISTANT + GANCYCLOVIR SENSITIVE

Figure 1.31 The generation of new length VNTR alleles by unequal crossing over. Misalignment of the tandem arrays, followed by reciprocal recombination, results in the formation of two new length alleles, containing novel numbers of repeat units. It is thought that this type of process occurs most often through unequal sister-chromatid exchange. The repeat arrays shown are flanked by a site for a restriction endonuclease (E) (modified from Jarman, 1989).

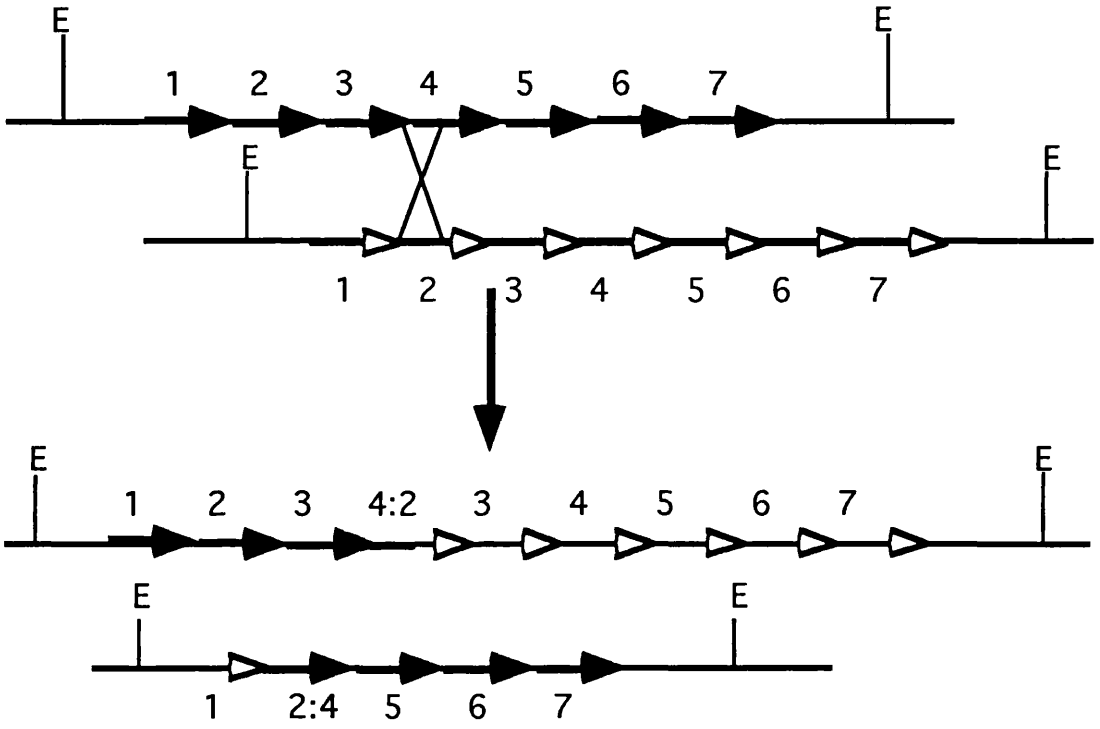

E
E
1
2
3
4
5
6
7
E
E
1
2
3
4
5
6
7
E
E
1
2
3
4:2
3
4
5
6
7
E
E
1
2:4
5
6
7

Figure 1.32 VNTR hypervariability at the human MUC1 gene locus. Genomic DNA isolated from peripheral blood from random unrelated individuals was digested with the restriction endonuclease HinfI and size-fractionated by agarose gel electrophoresis. DNA was transferred to nylon membranes and hybridised to P^{32} labelled human MUC1 VNTR probe pMUC7. Two hybridising fragments were observed in most individuals due to two alleles containing different numbers of tandem repeats.

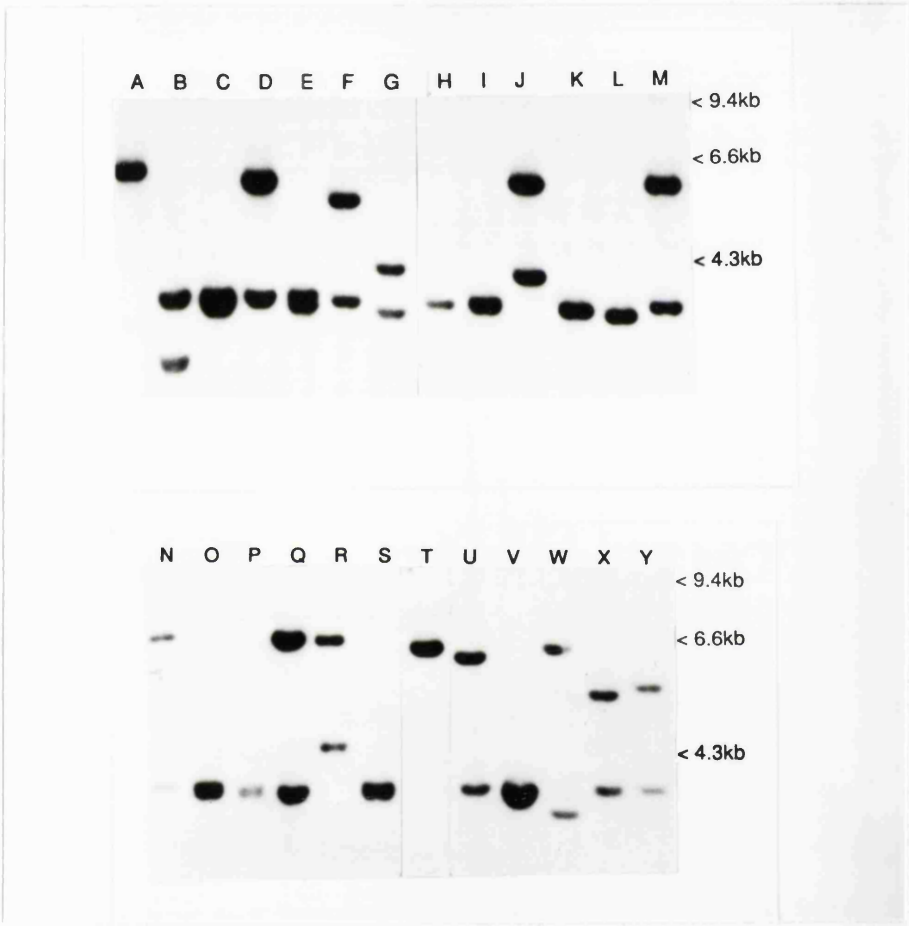

A B C D E F G H I J K L M
< 9.4kb
< 6.6kb
< 4.3kb
N O P Q R S T U V W X Y
< 9.4kb
< 6.6kb
< 4.3kb

Table 1 Mucin tandem repeat sequences. Amino acid sequence of the consensus tandem repeats for the mucin genes cloned to date. Serine, threonine and proline residues are indicated in bold. Where the map position and/or polymorphism has been determined it is indicated. References for the relevent cloning papers are included on page 18.

MUC1: Human MUC1 gene; expressed in the majority of simple secretory epithelial organs; Integral membrane protein;

Muc-1: Mouse Muc-1 gene; expressed by the majority of simple secretory epithelial organ; Integral membrane protein;

MUC2: Human MUC2 gene; expressed primarily by cells of the small and large intestine; Secreted, classical disulphide bonded mucin;

MUC3: Human MUC3 gene; expressed primarily by cells of the small and large intestine;

MUC4: Human MUC4 gene; expressed primarily by cells of the lungs and trachea;

MUC5: Human MUC5A-5C; Currently it is not known whether or not these clones represent transcripts from three diferent genes or ≥ 1 gene, expressed primarily by cells of the lungs and trachea;

MUC6: Human MUC6; expressed by the stomach

Rat intestinal: Thought to be the rat homologue of the human MUC2 intestinal mucin gene;

PSM: Porcine submaxillary mucin gene: expressed by the submaxillary glands; Secreted, classical disulphide bonded mucin;

BSM: Bovine submaxillary mucin gene

FIM-A.1, FIM-B.1 and FIM-C.1: *Xenopus laevis* frog integumentary mucins, expressed in secretory glands in the frog skin

PSGP: Apo-polysialoglycoprotein of rainbow trout eggs

Sgs-3 and Sgs-4: *Drosophila melanogaster* larval glue protein genes

GP-Iba: Human glycoprotein Ibα gene

abbreviations
ref to pages

TABLE 1
CLONED MUCIN AND MUCIN-LIKE GENES AND THEIR RESPECTIVE REPEAT SEQUENCES

GENE	REPEAT SEQUENCE	VNTR	MAP POSITION
MUC1	GSTAPPAHGVTSAPDTRPAP (20AA)	YES	1q21
Muc-1	DSTSSPV(A)HSGTSSPATSAPX (20-21AA)	NO	3
MUC2	PTTTPITTTTTVTPTPTPTGTQT (23AA)	YES	11p15
MUC3	HSTPSFTSSITTTETTS (17AA)	YES	7q22
MUC4	TSSA/VSTGHATPLPVTD (16AA)	YES	3q29
MUC5	(8AA)	ND	11p15
MUC6	SPFSSTGPMTATSFQTTTTYPTPSHPQTTLPTHVPPFSTSLVTPSTGTVITPTH AQMATSASIHSTPTGTIPPPTTLKATGSTHTAPPMTPTTSGTSQAHSSFSTAK TSTSLHSHTSSTHHPEVTPTSTTTITPNPTSTGTSTPVAHTTSATSSRLPTPFT THSPPTGS (169AA)	YES	11p15
Rat Intestinal	TTTPDV (6AA)	ND	ND
PSM	GAGPGTTASSVGVTETARPSVAGSGTTGTVSGASGSTGSSSGSPGATGASIG QPETSRISVAGSSGAPAVSSGASQAAGTS (81AA)	YES	ND
BSM	GTTVAPGSSNT (11AA)		
FIM-A.1	VPTTPETTT (9AA)	ND	ND
FIM-B.1	GESTPAPSETT (11AA)	YES	ND
FIM-C.1	KATTTTPTTTTTTPTTTTT (19AA)	YES	ND
PSGP	DDATSEAATGPSG (13AA)	YES	ND
Sgs-3	PTTTK (5AA)	YES	ND
Sgs-4	TCKTEPP (7AA)	YES	X
GP-Iba	SEPAPSPTTPEPT (13AA)	YES	ND

CHAPTER TWO:

MATERIALS AND METHODS

MATERIALS

2.1 Chemicals and solvents

All chemicals used were of analytical grade and were obtained from either BDH or Sigma Chemicals Ltd, except for the following:

acrylamide: bis-acrylamide (38:2)	BioRad
acrylamide: bisacrylamide (37.5:1)	Fisher, USA
SeaKem agarose	FMC Bioproducts, USA
SeaKem NuSieve low melting point	
GTG agarose	FMC Bioproducts, USA
ammonium persulphate	BioRad
ampicillin	Boehringer Manneheim
bacto-agar	Difco
bacto-tryptone	Difco
bromophenol blue	BioRad
caesium chloride	Boerhinger Manneheim
ethidium bromide (tablets)	BioRad
guanadinium isothiocyanate (GTC)	Fluka
methylene blue	BioRad
phenol (Tris-buffer saturated)	Gibco-BRL
repelcote	BioRad
xylene cyanole	BioRad
yeast extract	Difco

2.2 Radiochemicals

[α-^{32}P]dCTP	(3000 Ci/mmol)	Amersham International
[α-^{35}S]dATP	(1000 Ci/mmol)	Amersham International

2.3 Enzymes

Restriction endonucleases were purchased from New England Biolabs, Boehringer Manneheim or Promega.

AmpliTaq DNA polymerase	Perkin-Elmer Cetus
Calf Intestinal Phosphatase (CIP)	Boehringer Manneheim
Klenow DNA polymerase (5U/μl)	Amersham International
Lysozyme (powder)	Sigma Chemicals Ltd
M-MuLV reverse transcriptase	Boehringer Manneheim
RNase A	Boehringer Manneheim
Pronase	Sigma Chemicals Ltd
Proteinase K (powder)	Boehringer Manneheim
Sequenase version 2.0	United States Biochemical
T4 DNA Ligase (1U/μl)	Gibco-BRL
Taq DNA polymerase	Boehringer Manneheim

2.4 Miscellaneous

Biodyne nylon membrane	Pall Biodyne, USA
DEAE NA45 membrane	Schleicher and Schuell, Germany
Fuji Medical X-ray film	Fuji, Japan
Filtration units (0.2 and 0.45 μM)	Nalge Company, USA
Geneticin (G418)	Gibco-BRL
Gancyclovir (GanC)	Syntex Corporation
Gestyl (human chorionic gonadotropin)	Diosynth BV, Holland
Methyl green	Mallinckrodt, Germany
pBluescript SKII+ and KSII+	Stratagene, USA
pd(N)$_6$	Pharmacia, UK
Polaroid film types 57 and 553	Polaroid
Profasi (pregnant mare's serum gonadotropin)	Serono Laboratories, USA
Rainbow protein markers	Amersham International

Sequenase™ version 2.0 sequencing kit	United States Biochemical Corporation
Tissue culture plasticware	Falcon, USA
Vectastain ABC Kit	Vector Laboratories, USA

2.5 Buffers

All solutions were prepared using sterile deionised water and stored at room temperature unless otherwise stated.

10X restriction endonuclease buffers were supplied with the respective enzymes. Similarly, AmpliTaq DNA polymerase, Taq DNA polymerase, T4 DNA Ligase and calf intestinal phophatase (CIP) were supplied with the recommended buffers.

TE buffer pH 8.0: 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0.

20X SSC (saline sodium citrate):
3 M NaCl, 0.3 M sodium citrate. Adjusted to pH 7.0 with NaOH and made up to 1 litre.

10X MOPS: 200 mM MOPS in 400 ml 50 mM NaOAc pH 5.5. Adjusted to pH 7.0 with NaOH. 10 ml 0.5 M EDTA pH 8.0 added and volume adjusted to 500 ml. Sterile filtered and protected from light.

Electrophoresis running buffers:

a) 10X Tris-Boric acid-EDTA (TBE):
1.3M Tris base, 440 mM boric acid, 25 mM EDTA pH 8.0;

b) Formaldehyde-northern gel running buffer:
1X MOPS;

c) Glyoxal-northern gel running buffer:
10 mM NaH_2PO_4: Na_2HPO_4 pH 7.0;

d) Protein gel running buffer:

25 mM Tris base, 192 mM glycine, 0.1% (w/v) SDS.

DNA sample buffer: 33% glycerol, 0.375 X TBE buffer, 125 mM EDTA pH 8.0, 0.27% (w/v) SDS, bromophenol blue and xylene cyanole to colour.

Formaldehyde-northern gel sample buffer:

50% (v/v) glycerol, 1 mM EDTA pH 8.0, bromophenol blue to colour.

Protein gel sample buffer:

2% (w/v) SDS, 10% (v/v) glycerol, 80 mM Tris-HCl pH 6.8, 100 mM DTT, 0.2% bromophenol blue. Stored at -20°C.

2X PCR buffer: 20 mM Tris-HCl pH 8.3 (room temp), 100 mM KCl, 3 mM $MgCl_2$.

OLB Buffer:

Solution O: 1.2 M Tris-HCl pH 8.0, 125 mM $MgCl_2$;

Solution A: 1 ml of O, with 18 µl 2-mercaptoethanol, and 5 µl each of dATP, dGTP and dTTP (each nucleotide triphosphate dissolved in TE pH 7.0 at a concentration of 100 mM);

Solution B: 2 M Hepes pH 6.6, 4 M NaOH;

Solution C: Hexadeoxyribonucleotides evenly suspended in TE pH 8.0 at 90 OD_{260} units/ml;

Solutions A:B:C mixed in ratio of 100:250:150 to make OLB Buffer. Stored at -20°C.

Bacterial transformation buffers:

Tfb1: 100 mM RbCl, 100 mM $CaCl_2$, 30 mM KOAc, 50 mM $MnCl_2$, 15% (v/v) glycerol. Adjusted to pH 5.8 with 0.2 M glacial acetic acid;

Tfb2: 10 mM PIPES, 75 mM $CaCl_2$, 10mM $RbCl_2$, 15% (v/v) glycerol. Adjusted to pH 6.5 with KOH.

Genomic DNA Extraction Buffers:

Mouse tail and tissue extraction buffer:

50 mM Tris-HCl pH 8.0, 100 mM EDTA, 100 mM NaCl, 1% (w/v) SDS, 500 μg/ml Proteinase K. Stored at -20°C;

Cell extraction buffer: 10 mM Tris-HCl pH 7.5, 1 mM EDTA, 100 mM NaCl, 1% (w/v) SDS, 500 μg/ml Proteinase K. Stored at -20°C;

Plasmid mini-preparation buffers:

TENS buffer: 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 0.1 M NaOH, 0.5% (w/v) SDS;

CTAB solutions:

STET buffer: 8% (w/v) sucrose, 0.1% (v/v) Triton X-100, 50 mM EDTA pH 8.0, 50 mM Tris-HCl pH 8.0. Stored at 4°C;

CTAB solution: 5% (w/v) solution of CTAB in distilled water. Heated at 37°C to dissolve.

10X medium salt restriction endonuclease buffer:

100 mM Tris-HCl pH 7.5, 100 mM $MgCl_2$, 500 mM NaCl, 10 mM DTT. Stored at -20°C.

Bacteriophage λ library screening prewash buffer:

50 mM Tris-HCl pH 8.0, 1 M NaCl, 1 mM EDTA pH 8.0, 0.1% (w/v) SDS.

Cosmid/ large scale DNA preparation alkaline lysis buffers:

Buffer I: 50 mM glucose, 10 mM EDTA pH 8.0, 25 mM Tris-HCl pH 8.0;

Buffer II: 0.2 M NaOH, 1.0% (w/v) SDS;

Buffer III: 60 ml 5M KOAc, 11.5 ml glacial acetic acid, 28.5 ml distilled water. Final concentration of KOAc is 3 M.

4 M GTC (guanidinium isothiocyanate) solution:

4 M GTC, 25 mM sodium citrate pH 7.0, 0.2% (v/v) N-lauryl sarkosine. Sterile filtered. Immediately before use, 2.9 ml 2-mercaptoethanol were added to 197.1 ml of 4 M GTC.

5.7 M CsCl: 5.7 M CsCl, 0.1 M EDTA pH 8.0. Weighed before and after autoclaving to check for loss of weight.

Calcium phosphate transfection buffers:

10X HB-Salts: 1.37 M NaCl, 50 mM KCl, 7 mM $Na_2HPO_4.7H_2O$, 60 mM dextrose;

Solution H: 2X HB-Salts, 40 mM Hepes buffer adjust to pH 7.0 with HCl;

Solution C: 0.5 M $CaCl_2$;

Solution D: 160 μg/ml sheared salmon testes DNA 1 mM Tris-HCl pH 8.0, 0.1 mM EDTA pH 8.0. Stored at -20°C.

Phosphate-buffered saline (PBS):

1 tablet (Sigma) added to 200 ml sterile distilled water to make a solution of 1X PBS at pH 7.4.

2.6 Growth media and tissue culture reagants

A) Bacterial/bacteriophage:

L-Broth: 1.0% (w/v) bacto tryptone, 0.5% (w/v) bacto yeast extract, 1.0% (w/v) NaCl. For phage work pH was adjusted to 7.5 with NaOH. Sterilised by autoclaving;

LB-Agar: Bacto-agar added to LB to 0.7% (w/v);

Tetracycline/Ampicillin LB-Agar plates:

Ampicillin added to 100 μg/ml, tetracycline added to 40 μg/ml.

If IPTG and X-gal were added before plates were poured they were supplemented at respective concentrations of 5 mM and 40 μg/ml;

ψa medium: 2.0% (w/v) bacto tryptone, 0.5% (w/v) yeast extract, 20 mM $MgSO_4$. Adjusted to pH 7.6 with KOH. 1.4% (w/v) bacto agar. Sterilised by autoclaving;

ψb medium: ψa without bacto-agar;

Top-Agar: LB pH 7.5 with 0.7% (w/v) bacto-agar;

Bottom Agar: LB pH 7.5 with 1.5% (w/v) bacto-agar;

2X YT Broth 1.6% (w/v) bacto tryptone, 1.0% (w/v) yeast extract, 0.5% (w/v) NaCl. Sterilised by autoclaving;

SM medium: 100 mM NaCl, 8 mM $MgSO_4.7H_2O$, 50 mM Tris-HCl pH 7.5. Sterilised by autoclaving.

B) Mammalian tissue culture:

All tissue culture media and reagents were purchased from GIBCO-BRL with the exception of gancyclovir which was obtained from Syntex Corporation. All tissue culture media was stored at 4°C.

Mouse embryonic stem (ES) cell medium:

Dulbecco's modified Eagles medium (DMEM), minus pyruvate, with 4500 mg/ml glucose, 15% (v/v) batch tested foetal bovine serum (FBS), 5mM L-glutamine, 1% of a 100X stock of non-essential amino acids, 0.1 mM 2-mercaptoethanol, 1X penicillin-streptomycin, 10^3 units/ml leukaemia inhibitory factor (LIF).

For selection purposes, 400 μg/ml G418 (active constituent) was added to the above medium. Where gancyclovir was used, the respective concentration was 2 μM.

STO-neo fibroblast medium:

DMEM, minus pyruvate, with 4500mg/ml glucose, 10% (v/v) FBS, 5 mM L-glutamine;

During mitomycin C treatment of STO-neo cells, 5% (v/v) FBS was utilised rather than the standard 10% (v/v), and mitomycin C, prepared in 1X PBS, was added to the medium to a final concentration of 10 μg/ml.

Other cell lines:

HP-1, hamster pancreatic carcinoma:

Iscove's modified Eagle's medium, 10% (v/v) foetal calf serum (FCS), 5 mM L-glutamine;

ZR-75-1, human pancreatic carcinoma, HT1080, human fibrosarcoma and T47D, human mammary carcinoma:

DMEM, 10% (v/v) FCS, 5 mM L-glutamine;

C57MG, mouse mammary carcinoma:

DMEM, 10% (v/v) FCS, 10 μg/ml insulin;

HC11, mouse mammary carcinoma:

RPMI 1640, 10% (v/v) FCS, 5 μg/ml insulin, 10 μg/ml epidermal growth factor.

C) Embryo culture and animal solutions:

From Hogan, 1986.

M2 medium: 94.66 mM NaCl, 4.78 mM KCl, 1.71 mM $CaCl_2.2H_20$, 1.19 mM KH_2PO_4, 1.19 mM $MgSO_4.7H20$, 4.15 mM $NaHCO_3$, 20.85 mM HEPES, 23.28 mM sodium lactate, 0.33 mM sodium pyruvate, 5.56 mM glucose, 0.4% (w/v) BSA (Fraction V), 0.006% (w/v) penicillin G. potassium salt, 0.005% (w/v) streptomycin sulphate; 0.001% (w/v) phenol red;

Whitten's medium:

94.66 mM NaCl, 4.78 mM KCl, 1.71 mM $CaCl_2.2H_20$, 1.19 mM KH_2PO_4, 1.19 mM $MgSO_4.7H20$, 25.00 mM $NaHCO_3$, 23.28 mM sodium lactate, 0.33 mM sodium pyruvate, 5.56 mM glucose, 0.4% (w/v) BSA (Fraction V), 0.006% (w/v) penicillin G. potassium salt, 0.005% (w/v) streptomycin sulphate; 0.001% (w/v) phenol red;

Avertin: 50% (w/v) 2,2,2-tribromoethanol, 50% (v/v) TERT-amyl alcohol. The injection solution was made by diluting 30 µl stock solution with 2.5 ml PBS. In most cases 700 µl were injected intra-peritoneally (i.p.) per mouse;

PMSG (pregnant mare's serum gonadotropin):

25 units Profasi in Hank's buffered saline solution. Aliquotted into 1 ml and stored at -20°C;

HCG (human chorionic gonadotropin):

25 units Gestyl in Hank's buffered saline solution. Aliquotted into 1 ml and stored at -20°C.

2.7 Histochemicals

Methacarn fixative: 60% (v/v) methanol, 30% (v/v) chloroform, 10% (v/v) glacial acetic acid.

LacZ staining solutions:

Fixative:

Cells: 2% (v/v) formaldehyde, 0.2% (v/v) gluteraldehyde in PBS;

Tissues: 2% (v/v) formaldehyde, 0.2% (v/v) gluteraldehyde, 0.02% (v/v) NP-40, 0.01% (v/v) sodium deoxycholate;

Staining reaction mix: 1 mg/ml X-gal, 5 mM potassium ferricyanide, 5 mM potassium ferrocyanide, 2 mM $MgCl_2$. Made up in 1X PBS. For staining monolayers of cells, to minimise detachment of cells from substrate, molten agar was added to 1% (w/v) immediately before staining. For staining tissues Nonidet P-40 (NP-40) and sodium deoxycholate were supplemented as for the fixative.

2.8 Miscellaneous solutions

100X Denhardt's solution:

2% (w/v) bovine serum albumin (BSA), 2% (w/v) Ficoll, 2% (w/v) polyvinylpyrrolidone. Sterile filtered and stored in aliquots at -20°C.

Standard hybridisation mix:

5X Denhardt's, 5X SSC, 50 mM NaH_2PO_4: Na_2HPO_4 pH 6.5, 0.1% (w/v) SDS, 250 µg/ml denatured salmon sperm DNA, 50% (v/v) deionised formamide. Stored at -20°C.

Church hybridisation mix:

1% (w/v) BSA (Fraction V), 1 mM EDTA pH 8.0, 0.5 M NaH_2PO_4: Na_2HPO_4 pH 7.2, 7% (w/v) SDS.

Formaldehyde-northern hybridisation mix:

0.2 M NaH_2PO_4: Na_2HPO_4 pH 7.2, 1 mM EDTA pH 8.0, 7% (w/v) SDS, 45% (v/v) deionised formamide. Warmed prior to use to ensure even mixing.

Methylene blue staining solution for northerns:

0.5 M NaOAc pH 5.2, 0.04% (w/v) methylene blue. Used repeatedly.

Standard stringency wash buffers:

Low stringency, room temp wash buffer: 2X SSC, 0.1% (w/v) SDS; high stringency, 50°C wash buffer: 0.1X SSC, 0.1% (w/v) SDS.

Church wash buffers:

Low stringency, room temp wash buffer: 0.5% (w/v) BSA (Fraction V), 1 mM EDTA, 40 mM NaH_2PO_4: Na_2HPO_4 pH 7.2, 2% (w/v) SDS; High stringency, 65°C wash buffer: 1 mM EDTA, 40 mM NaH_2PO_4: Na_2HPO_4 pH 7.2, 1% (w/v) SDS.

Formaldehyde-northern wash buffer:

40 mM NaH_2PO_4: Na_2HPO_4 pH 7.2, 1 mM EDTA pH 8.0, 1% (w/v) SDS.

Poly-acrylamide gels:

Sequencing gel: 15 ml acrylamide stock (38% acrylamide: 2% bis acrylamide), 10 ml 10X TBE, 46 g urea, sterile distilled water to 100 ml. Warmed to 37°C to dissolve urea. Immediately before pouring, 500 µl 10% (w/v) ammonium persulphate and 110 µl TEMED were added;

Protein gel:

5% Running gel: 5 ml acrylamide stock (37.5:1); 11.2 ml 1 M Tris-HCl pH 8.8, 150 µl 20% (w/v) SDS, 13.7 ml sterile distilled water. 100 µl 10% (w/v) ammonium persulphate and 20 µl TEMED were added prior to pouring;

3% Stacking gel: 1.0 ml acrylamide stock (37.5:1), 1.25 ml 1 M Tris-HCl, pH 6.8, 50 µl 20% (w/v) SDS, 7.7 ml sterile distilled water. 50 µl 10% (w/v) ammonium persulphate and 10 µl TEMED were added prior to pouring.

METHODS

2.9 MOLECULAR BIOLOGY METHODS

2.91 Genomic DNA isolation

A) Isolation of genomic DNA from cell lines

Genomic DNA was prepared using a variety of methods. In general, two different methods were utilised for the isolation of high molecular weight genomic DNA from cell lines. These were as follows:

Method 1: Cells growing in tissue culture flasks were washed twice with cold PBS which was aspirated off after each wash. Ten mililitres of fresh PBS were added to the flask and the cells were scraped off the bottom of the flask with a cell scraper into a centrifuge tube. The resultant cell suspension was centrifuged at 1,000 rpm for 8 minutes at 4°C in order to pellet the cells. The cell pellet was resuspended in 3 mls of an ice cold solution consisting of 150 mM NaCl and 25 mM EDTA. Fifteen microlitres of a solution of 10% (w/v) SDS with 300 μg of Proteinase K were added to the cell suspension which was subsequently incubated at 50°C for between 1 and 4 hours.

After incubation at 50°C, the solution was extracted twice with Tris-buffer equilibrated phenol followed by two extractions with chloroform: iso-amyl alcohol (24: 1). The samples being extracted were mixed for 5 minutes on a vertical rotor in each case. The tubes containing the respective samples were spun for 5 minutes at 5,000 rpm to separate the two phases, after which the aqueous phase was removed into a fresh tube using a 10 ml wide-bore pipette. After the final extraction, 2.0 volumes of ice cold 100% ethanol with 0.1 volumes of 0.4 M NaCl were added to the sample. The sample was mixed by hand at which stage DNA became visible as a stringy white precipitate. A pasteur pipette was fashioned with a small hook at the tip, and this was used in order to retrieve the DNA precipitate from solution. The DNA was spooled out onto the pipette and transferred to a 1.5 ml microcentrifuge tube containing 70% ethanol in order to wash the precipitate. The DNA, still on the pipette, was allowed to air-dry then transferred to a further microcentrifuge tube containing 50 μl sterile distilled water, in which it

was dissolved. The concentration of the DNA was estimated through measurement of the optical density (OD) of a 1 in 200 dilution at 260 nm (assuming a 50 μg/ml solution of DNA to have an OD_{260} value of 1.0), and the purity was determined through measurement of the OD at 280 nm and subsequent calculation of the ratio OD_{260} to OD_{280} (assuming that this ratio is 1.80 for a pure solution of DNA). The integrity of the DNA was checked by running a small amount out on a 0.7% (w/v) agarose gel, alongside λHindIII molecular weight markers.

Method 2: This method largely superceded method 1 and was initially used in order to isolate genomic DNA from mouse embryonic stem cells growing in 24-well plates. Cells were washed twice with PBS. The buffer was aspirated off and DNA extraction buffer was added. In general, cells growing in 24-well plates received 500 μl of extraction buffer. Those growing in 6-well plates received 1 ml of extraction buffer, whereas cells growing in, for instance, a 75 cm^2 (T75) flask received between 3 and 5 mls of extraction buffer. Cells were digested overnight at 37°C, removed to a microcentrifuge or 15 ml tube, then extracted once with an equal volume of Tris-buffered phenol: chloroform: iso-amyl alcohol (25: 24: 1). Tubes were mixed well, allowed to stand for 5 minutes then centrifuged at a minimum of 3,000 rpm for 10 minutes in the benchtop or microcentrifuge. The aqueous phase was removed to a fresh tube and the DNA precipitated by the addition of 0.6 volumes of ice cold propan-2-ol. At this stage the DNA became visible as a stringy white precipitate which was collected by centrifugation. The resultant DNA pellet was washed once with 70% ethanol, air-dried then dissolved in an appropriate volume of sterile TE buffer pH 8.0, or sterile distilled water. From a single well of a 24-well plate the yield of DNA was sufficient for one restriction endonuclease digestion (approximately 10 to 20 μg total yield). From cells growing in larger areas the concentration, purity and integrity of the DNA preparations was determined as described above.

B) Isolation of high molecular weight genomic DNA from tissue

This method was initially developed for the isolation of high molecular weight genomic DNA from transgenic mouse tails (Hogan, 1986), but it was found to be an effective method for the isolation of genomic DNA from any tissue.

Seven hundred microlitres of tissue DNA extraction buffer were added to the tissue samples (< 1 gram) in 1.5 ml microcentrifuge tubes. The tubes were incubated overnight at 55°C. Tubes were spun for 1 minute at maximum speed in a microcentrifuge in order to pellet debris. The supernatant was removed into a fresh tube and extracted twice with an equal volume of Tris-buffered phenol: chloroform: iso-amyl alcohol (25:24:1). For each extraction, tubes were mixed by vortexing, allowed to stand on the bench for 5 minutes then microcentrifuged for 10 minutes. Point six volumes of ice-cold propan-2-ol were added to each tube which was then mixed by vortexing and microcentrifuged for 10 minutes. The resultant pellet was washed once with 70% ethanol, then allowed to air-dry on the bench. The dried pellet was dissolved in an appropriate volume of sterile TE buffer pH 8.0 or sterile distilled water. As above, the concentration, purity and integrity of the DNA was determined through measurement of the OD_{260} and OD_{280} values and through agarose gel electrophoresis.

2.92 Restriction endonuclease digestion of plasmid DNA and genomic DNA

Restriction endonuclease digestion of both plasmid and genomic DNA was typically carried out in a total volume of 30 µl. 10X buffers for the respective restriction endonucleases were supplied by the manufacturer, and used at a final working concentration of 1X. For digests in which more than one restriction endonucease was utilised simultaneously the optimal buffer was determined according to the manufacturer's instructions. If it was determined that both enzymes would not digest completely together, for instance if SmaI, which cuts best at room temperature was to be used in combination with an enzyme that is optimally active at 37°C, the two respective digests were carried out sequentially. Genomic DNA digests were of between 10 and 20 µg total genomic DNA and proceeded overnight at the optimal temperature for the particular restriction endonuclease being utilised. Plasmid DNA (typically 1-5 µg) was digested for a minimum of one hour at the optimal temperature. Where larger amounts of plasmid were digested, the reaction was allowed to proceed according to published activity values for the respective restriction endonucleases. All restriction endonuclease digests were typically carried out with a 3-5 fold excess of enzyme, at a final glycerol concentration not exceeding 10% (v/v).

2.93 Agarose gel electrophoresis of DNA

Agarose gels were prepared by dissolving agarose at 0.8-3.0% (w/v) in 1X TBE buffer in a microwave oven. The solution was allowed to cool, after which ethidium bromide was added to a final concentration of 0.1 μg/ml, and the gel was poured into a gel mould containing a comb well former. Gels were allowed to set at room temperature. DNA samples were prepared by the addition of DNA sample buffer to 1/5th final volume, and introduced into the wells of the gel submerged in 1X TBE buffer. Electrophoresis was carried out in 1X TBE buffer prepared from 10X TBE stock solution. Gels were electrophoresed at 5 to 7 V/cm at room temperature during the day or 1 to 2 V/cm overnight, until the desired range of separation of the DNA fragments was achieved. DNA was visualised by illumination over a long wave UV light box, and photographed with polaroid film. The sizes of fragments were estimated by comparison of their mobility relative to molecular weight markers of known size. Molecular weight markers utilised were as follows:
a) HindIII restriction endonuclease digestion of bacteriophage λ DNA;
b) HaeIII restriction endonuclease digestion of bacteriophage ϕX174 DNA;
c) BRL 1 kilobase pair ladder.

2.94 Isolation of DNA fragments from gels

A) Isolation of DNA fragments from low melting point agarose

Low melting point (LMP) agarose gel purification was utilised for the preparation of all the DNA probes described in this thesis. Low melting point agarose gels were prepared in 1X TBE as for standard agarose gels, except that LMP gels were allowed to set at 4°C. As LMP gels tend to be much more fragile than standard agarose gels, LMP gels of greater than 1.0% (w/v) were routinely utilised. DNA samples were loaded into the gel and electrophoresed at 4°C in 1X TBE buffer at 5 V/cm. After suitable separation, fragments to be isolated from LMP gels were cut out of the gel as a thin gel slice. This was carried out over long wave UV illumination. Excess agarose was trimmed from around the DNA within the gel slice, and the slice was transferred to a sterile 1.5 ml microcentrifuge tube that had been preweighed. The tube containing the gel slice was then reweighed and the weight of the gel slice was calculated. Sterile distilled water was added to the tube at the ratio of 3 ml

water to 1 gram of gel. The tube was subsequently boiled for 10 minutes after which time the probe was aliquotted into separate tubes, the concentration of DNA was estimated and recorded on the tubes in nanograms/μl, and the tubes were stored at -20°C.

B) Isolation of DNA fragments using DEAE paper

This method was used for the isolation and purification of DNA fragments for cloning purposes. It was used successfully to isolate DNA fragments up to and greater than 10 kilobase pairs. However, in general, smaller DNA fragments were more efficiently purified by this method.

Using a sterile scalpel, a small cut was made in the gel immediately ahead of the DNA fragment to be isolated. This procedure was performed over UV illumination. Squares of DEAE (NA45) ion-exchange paper (Schleicher and Schuell, Dassell, Germany) were cut slightly larger than the width of one of the wells, and then soaked in TE buffer pH 8.0. Using forceps and a sterile scalpel, the DEAE paper was carefully inserted into the incision that had been made in the gel. The ion-exchange paper was positioned such that once electrophoresis resumed, the DNA fragment to be isolated would run onto the paper. The correct position of the ion-exchange paper was verified over UV illumination. Electrophoresis resumed and the gel was allowed to run for a further 15-20 minutes. After this period of time, the gel was removed and again examined over UV light. At this time the DNA fragment being isolated was almost invariably entirely located on the ion-exchange paper. If not, the gel was returned to the electrophoresis chamber and electrophoresis was continued. Once it was confirmed that the DNA fragment had migrated onto the paper, the paper was removed from the gel and washed briefly in TE buffer pH 8.0. Excess paper was trimmed from around the DNA, and the paper was transferred to a 1.5 ml microcentrifuge tube containing 400 μl 1 M NaCl. The tube was then placed at 70°C for 30 minutes, vortexing every 5-10 minutes. After 30 minutes, the paper was removed from the tube, washed in TE buffer pH 8.0 and examined over UV light to determine if the DNA had come off the paper into the salt solution. If this was found to be so, 1.0 ml of ice-cold 100% ethanol was added to the tube which was subsequently vortexed briefly to mix. For DNA fragments known to be low in concentration, 10 μg yeast tRNA (transfer RNA) were added to the tube

prior to ethanol precipitation. The tube was transferred briefly to dry ice or -20°C, after which time it was centrifuged for 10 minutes at top speed in the microcentrifuge. The resultant DNA pellet was washed with 70% ethanol and air-dried. The pellet was dissolved in sterile distilled water and one tenth of the total volume was analysed by agarose gel electrophoresis. The concentration of the purified DNA fragment was estimated through reference to DNA molecular weight marker fragments which were run on the same gel.

2.95 Labelling DNA fragments

A) Labelling DNA fragments to high specific activity by random priming

This method was used for the preparation of all labelled DNA probes described in this thesis and follows the method of Feinberg, 1984. DNA was isolated in low melting point agarose as described above and aliquotted at a final concentration of between 2 and 5 nanograms/µl. Approximately 50 nanograms of DNA in low melting point agarose were boiled for 10 minutes, centrifuged briefly to bring down any condensate, and placed at 37°C for 5 to 10 minutes. Ten microlitres of OLB buffer and 2 µl BSA (10 mg/ml solution) were added to the tube and the volume was raised to 44.5 µl. Two point five units of Klenow DNA polymerase were added to the tube followed by 50 µCi [α^{32}P]dCTP (5 µl volume). Labelling was carried out for at least 4 hours at room temperature, but was generally allowed to proceed overnight. The reaction was terminated by the addition of 2 µl of 0.5 M EDTA pH 8.0. Labelling by this method typically resulted in probes labelled to a specific activity of > 1 X 10^8 dpm/µg.

The level of incorporation of radioactivity was measured as described (Maniatis, 1982). Five microlitres of the labelled probe were diluted up to a total of 50 µl with sterile distilled water. Five microlitres of the diluted probe were spotted onto the centre of a Whatman glass fibre disc. An equal volume of the diluted probe was added to a tube containing 100 µl of a solution of salmon sperm DNA (500 µg/ml in 20 mM EDTA pH 8.0). Five mililitres of a solution of ice-cold 10% (w/v) trichloroacetic acid (TCA) were added to the tube. The resultant precipitate was collected by filtering through another glass fibre filter disc which was subsequently washed several additional times with 5 ml ice-

cold 10% (w/v) TCA. After washing with TCA, the filter was washed with 95% ethanol. Both filters were dried under a heat lamp. Filters were counted by liquid scintillation, the first filter giving a measure of the total radioactivity in the sample, and the second a measure of the radioactivity specifically incorporated into the probe.

B) Preparation of radioactively labelled molecular weight markers;

Radioactively labelled DNA molecular weight markers were prepared for reference purposes on Southern blots. Bacteriophage λ DNA was digested with the restriction endonuclease, HindIII. Klenow DNA polymerase specifically incorporates nucleotides into 3' recessed ends. The first available nucleotide of the 3' recessed end created by the restriction endonuclease HindIII is an adenosine (A). In order to label HindIII digested λ DNA, 1 μg of DNA in a total volume of 44 μl, was heated to 65°C for 10 minutes then cooled on ice. Five microlitres of 10X medium salt buffer, 5 μCi of [α^{35}S]dATP, and 2.5 units of Klenow DNA polymerase were added to the tube and the labelling reaction was incubated at 37°C for a minimum of 30 minutes. After 30 minutes, 2 μl of 0.5 M EDTA pH 8.0 were added to the tube to terminate the labelling reaction. For Southern blots, 5 μl of labelled λ HindIII DNA were added to 2 μg cold (unlabelled) λ HindIII DNA.

2.96 Southern blotting

A) Blotting onto nylon membranes

Gels were run until the desired range of separation was achieved, as described above. At this time gels were documented by photography over long wave UV illumination. For reference, a ruler was placed alongside the gel which could be compared in the photograph to the respective positions of the molecular weight warkers utilised. Gels were agitated for 30 minutes in a solution of 0.5 M NaOH, 1.5 M NaCl in order to denature the DNA. DNA was transferred overnight onto Biodyne B nylon membranes by capillary elution in the presence of 20X SSC, as described in Maniatis, 1982. After overnight transfer, the positions of the wells were marked onto the nylon membrane with pencil, the efficiency of transfer was checked by holding the membrane over the UV illuminator, and the membrane was baked at 80°C for 1 hour.

B) Hybridisation with ^{32}P radiolabelled DNA probes

Baked membranes were soaked in 6 X SSC for 5 minutes before prehybridisation in standard hybridisation mix. Hybridisation mix was pre-heated to 95°C-100°C for 10 minutes to denature the salmon sperm DNA, chilled on ice, then added to the membrane in a heat-sealed plastic bag at a ratio of 4 ml hybridisation solution per 100 cm^2 of membrane. Membranes were prehybridised for a minimum of 1 hour at 42°C. [$\alpha^{32}P$]dCTP radiolabelled probe was boiled for 5 minutes, spun briefly, then chilled on wet ice. Probe was added to an activity of approximately 5 X 10^5 dpm/ml, to fresh hybridisation solution that had been treated as described above. The probe-hybridisation mix was added to the membrane in a fresh heat-sealed plastic bag at a ratio of 2 ml hybridisation mix to 100 cm^2 membrane. Bubbles were removed from the bag by carefully rolling a pipette along the bag. Membranes were hybridised overnight for a minimum of 16 hours at 42°C on a rocking platform.

C) Stringency washes and autoradiography

Membranes were removed from the plastic bags in which they were hybridised and washed 4 times in room temperature wash buffer for 5 minutes. Approximately 250 ml of wash buffer were utilised per 100 cm^2 of membrane. Washes were performed in plastic trays on a rotor-shaker (Luckham, UK). After the room temperature washes, high stringency wash buffer, pre-heated to 50°C, was added at 250 ml/ 100 cm^2 membrane and the trays were placed in a shaking water bath at 50°C. Membranes were washed twice for 15 minutes at high stringency. After washing, the membranes were wrapped in clingfilm and specific DNA-DNA duplexes were visualised by autoradiography at -70°C in the presence of intensifier screens. Autoradiography proceeded from several hours to several days, depending upon the intensity of the signal obtained.

D) Rehybridisation of membranes

A number of methods were used to remove labelled probe from membranes in order that they could could be rehybridised. They included: washing in boiling in 0.1% (w/v) SDS (100 ml/ 100 cm^2

membrane) for 2 X 15 minutes, washing in boiling water (100 ml/ 100 cm^2 membrane) for 2 X 10 minutes, and incubating in a solution of 10 mM NaH_2PO_4: Na_2HPO_4 pH 6.5 with 50% (v/v) deionised formamide (100ml/100 cm^2 membrane) at 65°C for 1 hour. In each case, before rehybridising, membranes were autoradiographed overnight, as described above, in order to determine the efficiency of probe removal.

2.97 DNA ligations

A) Preparation of vector DNA

The cloning vectors utilised throughout for all cloning procedures were pBluescript SK and KS II+ (Stratagene, USA). In order to reduce the self-ligation of vector DNA when a single restriction endonuclease enzyme was used to prepare the vector, the terminal 5' phosphate groups were removed using calf intestinal phosphatase (CIP). Typically, 1-2 μg plasmid DNA were digested with the appropriate restriction endonuclease and then precipitated by the addition of 0.1 volumes 3 M NaOAc pH 5.5 and 2.5 volumes 100% ethanol. DNA was recovered by centrifugation in a bench-top microfuge, the pellet was washed with 70% ethanol, dried and redissolved in CIP buffer. Calf intestinal phosphatase was added to the DNA at the ratio of 0.01 units of enzyme per 1 pmole (picomole) of 5' protruding ends or 1 unit of enzyme per 1 pmole of blunt or 5' recessed ends. For 5' protruding ends the tubes were incubated at 37°C for 30 minutes, whereas CIP treatment of 5' recessed or blunt ends required incubation at 50°C for 1 hour. After CIP treatment, 0.1 volumes of 0.5 mM EDTA pH 8.0 were added to each tube and the tubes were heated at 65°C for 45 minutes in order to deactivate the enzyme. DNA was purified by extraction with Tris-buffered phenol: chloroform: iso-amyl alcohol (25: 24: 1), ethanol precipitated, washed with 70% ethanol, dried and resuspended in TE buffer pH 8.0 at a final concentration of approximately 10-25 ngs/μl.

B) Conversion of 5' and 3' overhanging ends to blunt ends using Klenow DNA polymerase

Where necessary, 5' and 3' overhanging hands were blunted using the Klenow fragment of DNA polymerase I. Klenow fills in 3' overhanging ends, whereas it blunts 5' protruding ends by removing the

overhang. For conversion of 3' overhanging ends to blunt ends, 1-2 μg restricted DNA were incubated for 30 minutes at 37°C in a final volume of 50 μl, in the presence of 2 mM dNTPs, in medium salt restriction buffer with Klenow polymerase at 2.5 units/μg DNA. Five-prime protruding termini were removed under the same conditions as that used for filling in 3' overhanging ends, except that dNTPs were omitted from the reaction mixture. DNA was purified by extraction with Tris-buffered phenol: chloroform: iso-amyl alcohol (25: 24: 1), ethanol precipitated, washed with 70% ethanol, dried and resuspended in TE buffer pH 8.0.

C) Ligation of DNA fragments

Ligations were routinely carried out with 10-25 ng vector DNA and an equimolar ratio of insert DNA, derived as a gel-purified restriction fragment or PCR product. Ligations were generally performed in a final volume of 10-20 μl with 1-2 units of T4 DNA Ligase, and were incubated for a minimum of 4 hours at room temperature. For blunt-ended ligations, reactions were incubated overnight at 16°C.

2.98 Bacterial transformation

A) Bacterial strains

Escherichia coli DH5α or XL-1 Blue (Stratagene, USA) bacterial strains were used throughout to propagate and amplify plasmid DNAs. All plasmids described herein contain the β-lactamase gene which allows the selection of bacterial transformants on L-Agar containing ampicillin at a concentration of 50 to 100 μg/ml.

B) Preparation of competent bacteria

Bacteria were streaked out on ψa plates and incubated upside down at 37°C overnight. A single colony was picked and inoculated into 5 ml of ψb broth and incubated overnight at 37°C with shaking. The culture was subcultured 1: 20 into 100 ml of prewarmed ψb broth and grown until an OD_{550} of approximately 0.48 was reached. Cells were chilled on wet ice for five minutes then recovered by centrifugation in prechilled corex tubes at 4,000 x g at 4°C. The resultant cell pellet was resuspended in

0.40 volumes Tfb1 buffer and left on wet ice for 5 minutes. Cells were recovered again, as before, and the resultant cell pellet was resuspended in 0.04 volumes Tfb2 buffer. Cells, resuspended in Tfb2 buffer, were left on wet ice for 15 minutes then aliquotted into freezing vials, 200 µl of cells per vial, and snap frozen on dry ice. For long term storage, cells were stored under liquid nitrogen.

C) Transformation of competent bacteria

Competent cells were thawed on wet ice for approximately 15 minutes. For each transformation, a minimum of 40 µl of cell suspension was added to 10-20 µl of each DNA sample, and the tubes were placed on wet ice for approximately 30 minutes, mixing gently every 10 minutes. After incubation on ice, the cells were heat shocked at 42°C for 90 seconds. One hundred and sixty microlitres of LB were added to the cell-DNA and the tubes were incubated at 37°C for 30 to 45 minutes. Transformed cells were spread onto pre-dried L-agar plates containing 100 µg/ml ampicillin which were then incubated upside down overnight at 37°C.

The vector pBluescript is designed such that antibiotic resistant colonies carrying inserts can be identified by inclusion of the galactose analogue, 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal), in addition to isopropyl-β-D-pyranoside (IPTG) in the culture medium. Colonies carrying plasmids with empty vectors (no inserts) almost invariably grow as blue colonies, due to the action of the bacterial β-galactosidase (LacZ) enzyme on the X-gal. The multi-cloning site polylinker of pBluescript vectors is inserted into the LacZ gene in such a way that ligation of an insert into the polylinker will generally result in a frame-shift mutation within the LacZ gene. Thus, bacteria transformed by plasmids containing inserts are usually unable to synthesise β-galactosidase and, therefore, can be selected on X-gal/IPTG plates due to their white colour. It should be pointed out, though, that only certain strains of *E. coli* (those that lack an endogenous LacZ gene) can be utilised in this way. XL-1 Blue cells can be used for blue/white colour selection, whereas DH5α cannot be used as they possess an endogenous β-galactosidase gene.

2.99 Preparation of plasmid DNA

A) Preparation of mini-prep DNA

Two methods were utilised to isolate mini-prep DNA. These methods were based on those of Del Sal, 1988, and Zhou, 1990, and each produced between 5 and 20 μg of plasmid DNA, depending upon the plasmid vector. Both methods were used for the rapid isolation of plasmid DNA used for diagnostic restriction endonuclease digestion and in direct plasmid sequencing.

CTAB method (Del Sal, 1988): One and a half mililitres of bacterial overnight culture were spun for 10-20 seconds in a bench-top microcentrifuge and the culture medium was aspirated leaving the cell pellet. The pellet was resuspended in 200 μl of STET buffer by vortexing, and incubated for 5 minutes at room temperature after the addition of 4 μl of a 50 mg/ml solution of lysozyme in 10 mM Tris-HCl pH 8.0. The resultant suspension was boiled for 45 seconds and centrifuged for 10 minutes at room temperature. The pellet of cell debris was removed from the tube using a sterile toothpick or yellow disposable pipette tip, and the plasmid DNA was precipitated from the supernatant by the addition of 8 μl of a 5% (w/v) solution of cetyltrimethylammonium bromide (CTAB). The DNA was spun down by centrifugation for 5 minutes at room temperature. The DNA pellet was resuspended in 300 μl of 1.2 M NaCl and reprecipitated by the addition of 660 μl ice-cold 100% ethanol followed by centrifugation for 10 minutes at room temperature. The final pellet was washed with 70% ethanol, air-dried and resuspended in 20 μl TE buffer pH 8.0. For restriction enzyme digest analysis, typically 2-5 μl was used. 5 to 10 μl of the total sample was used for direct plasmid sequencing.

TENS method (Zhou, 1990): One and a half mililitres of bacterial overnight culture was spun for 10-20 seconds in a bench-top microcentrifuge and the culture medium was poured off leaving 50-100 μl of medium behind. The cell pellet was resuspended by vortexing. Three hundred microlitres of TENS buffer were added to the tube which was then mixed by vortexing for 2-5 seconds. One hundred and fifty microlitres of a solution of 3 M NaOAc pH 5.5 were added to the tube, which was vortexed briefly and spun for 2 minutes in the bench-top

microcentrifuge to pellet cell debris. The supernatant was transferred to a fresh tube and DNA was precipitated by the addition of 900 µl of ice-cold 100% ethanol, vortexing and centrifugation for 2 minutes in the bench-top microcentrifuge. The resultant DNA pellet was washed with 70% ethanol, air-dried and resuspended in a total volume of 20µl TE buffer pH 8.0 or an equivalent volume of sterile distilled water. Typically, 2-5 µl of plasmid DNA was used for restriction digest analysis, whereas for direct plasmid sequencing 5 µl was utilised.

B) Large scale preparation of plasmid DNA

Again, two methods were utilised for large scale preparation of plasmid DNA. Both methods typically yielded between 500 µg and 3 mg of supercoiled plasmid DNA, and was used to prepare plasmid DNA for long term sequencing projects and for transfection into mammalian cells.

Large scale LiCl plasmid preparation: A single colony was picked from a freshly plated culture, and inoculated into 5 ml LB. After incubation for greater than 5 hours, the culture was subcultured into 400 ml of LB supplemented with 100 µg/ml ampicillin. Cultures were incubated overnight at 37°C with shaking. Bacterial cells were harvested by centrifugation at 7,000 x g for 5 minutes. The cell pellet was resuspended in 10 ml TE buffer pH 8.0. Twenty mililitres of large scale DNA preparation solution ii were added to the resuspended cells, followed by 10 ml of solution iii. The cell suspension was mixed by shaking vigorously then transferred to a 50 ml sterile disposable centrifuge tube (Falcon). Cell debris was pelleted by centrifugation for 20 minutes at > 2,500 rpm in a bench-top centrifuge. The supernatant was transferred into two fresh 50 ml centrifuge tubes by filtering through sterile gauze. Point six volumes of ice-cold propan-2-ol were added to each tube which was mixed well and centrifuged for 15 minutes at >2,500 rpm in the bench-top centrifuge. The resultant nucleic acid pellet was washed with 70% ethanol, resuspended in 5 ml TE buffer pH 8.0 and transferred to a corex tube. An equal volume of ice-cold 5 M LiCl was added to each tube. Tubes were covered with parafilm, inverted to mix and spun at 10,000 rpm for 5 mins at 4°C. After centrifugation, the supernatant was poured into a corex tube and DNA was precipitated by the addition of 2.5 volumes of 100% ethanol. DNA was pelleted by centrifugation at 10,000 rpm for 10 minutes. The pellet was washed with

70% ethanol and dissolved in 1 ml of TE buffer pH 8.0 containing 40 μg/ml pre-boiled RNase A. An equal volume of a solution of 13% polyethylene glycol (PEG) 6000 in 1.6 M NaCl was added and the tubes were mixed and incubated at 4°C for 1 hour. DNA was pelleted by centrifugation at 10,000 rpm for 5 minutes, dissolved in 400 μl TE buffer pH 8.0 and transferred to a 1.5 microcentrifuge tube. Each DNA sample was extracted three times with an equal volume of Tris-buffered phenol: chloroform: iso-amyl alcohol (25: 24: 1). DNA was precipitated from the final aqueous phase by the addition of 0.1 volumes 3 M NaOAc pH 5.5 and 2.5 volumes of ice-cold 100% ethanol, and centrifugation in the bench-top microcentrifuge for 10 minutes at room temperature. The resultant DNA pellet was washed with 70% ethanol, dried and dissolved in 300 μl of sterile distilled water. The concentration, purity and integrity of the plasmid DNA was determined as previously described. Plasmid DNA was diluted to a final working concentration of 1 μg/μl, aliquotted and stored at -20°C.

Large scale CsCl plasmid preparation: This method proceeded as for the LiCl method described above, up to the propan-2-ol precipitation. After precipitation, the resultant nucleic acid pellet was resuspended in 5 ml of TE buffer pH 8.0. The DNA solution was added to a preweighed sterile universal (Sterilin, UK) and made up to a total of 9 grams by the addition of TE buffer pH 8.0. Ten grams of ultrapure caesium chloride (CsCl) with 50 μl ethidium bromide solution (10 mg/ml) were added to each tube. The solutions were loaded into Beckman polyallomer quick seal tubes, balanced by addition of a balance solution of 9.5 ml TE buffer pH 8.0 with 10 g CsCl, and spun at 64,000 rpm at room temperature for at least 18 hours in a Beckman ultracentrifuge. After spinning, supercoiled plasmid DNA was removed through the side of the tube with a 19-guage hypodermic needle and syringe. An equal volume of distilled water was added to the DNA which was extracted several times with an equal volume of water-saturated butan-1-ol, in order to remove the ethidium bromide. Plasmid DNA was precipitated by the addition of 2.5 volumes of 100% ethanol (at room temperature), recovered by centrifugation and washed with 70% ethanol. DNA was dissolved in TE buffer pH 8.0 and the concentration, purity and integrity was determined as previously described. DNA was diluted to a working concentration of 1 μg/μl, aliquotted and stored at -20°C.

2.910 Bacteriophage λ library screening

A) Bacterial host strains and the preparation of plating cells

The choice of *Escherichia coli* strains was dependent upon the bacteriophage being utilised. For λgt10, strain NM514 was used. For λgt11, Y1090 cells were used as the host strain, whereas the λZAPII bacteriophage work described herein was carried out with XL-1 Blue cells as the host. All plating cells were prepared in the same way. Bacteria were streaked onto L-Agar plates and incubated upside down at 37°C overnight. A single colony was inoculated into 10 ml LB pH 7.5, supplemented with 0.4% maltose, and incubated with shaking overnight at 37°C. The presence of maltose in the growth medium stimulates the expression of the *E. coli* bacteriophage λ receptor protein. After overnight culture, 1 ml was subcultured into 50 ml prewarmed LB pH 7.5 with 0.4% maltose and incubated further until an OD_{600} of approximately 0.5 was reached. This typically took between 2.5 and 3.0 hours. Cells were chilled on wet ice and pelleted by centrifugation at 3,000 rpm for 10 minutes at 4°C. The cell pellet was resuspended in 15 ml of ice-cold 10 mM $MgSO_4$. Cells were stored at 4°C until ready for infection with bacteriophage.

B) Phage adhesion and plating

Bacteriophage libraries were diluted in SM medium to an appropriate concentration. For a 9 cm plate approximately 3×10^4 plaque forming units (pfus) were sufficient to give a good plaque density without causing plate lysis. The respective numbers of pfus for 15 cm plates and 24 x 24 cm plates were 8.4×10^4 and 3×10^5 pfus. Bottom agar plates were ideally prepared 2 to 3 days in advance and were dried and prewarmed in the incubator immediately before use. Sterile top agar was prepared, $MgSO_4$ was added to a final concentration of 10 mM and the top agar was placed in a 45°C water bath for 20 minutes. One hundred microlitres of phage, at the appropriate concentration, were added to 0.3 ml (9 cm plate), 0.6 ml (15 cm plate), or 1.6 ml (24 x 24 cm plate) of plating cells and incubated at 37°C for 20 minutes with occasional agitation. Two point five to three mililitres (9 cm plate), 7.5-8.0 ml (15 cm plate) or 22.0 ml (24 x 24 cm plate) of top agar were added to each culture and poured onto the prewarmed LB bottom agar plates. Plates were incubated overnight, upside down at 37°C.

C) Membrane lifts

After overnight growth, membrane lifts were taken as follows: Biodyne nylon membranes (Pall Biodyne, Glen Cove, New York) were purchased at the correct size or, alternatively, filters were cut to the desired size from a roll. Double lifts were taken off each plate. Filters were numbered accordingly, on the side of the filter that would contact the plaques. Filters were placed carefully onto the surface of the plate. After the membrane had become moistened by the agar, the filter was marked by puncturing at least three times through the membrane and agar with an 18-gauge needle. The plate was then turned over on a light box and the position of the three marks was recorded on the back of the plate with a fine-point permanent marker pen. The first filter was left on the plate for approximately 1 minute. The procedure was repeated once for each plate. In each case the orientation marks made in the plate during the first lift were transferred to the second lift. The plate, with the second membrane on the surface of the agar, was placed on a light box such that the pen marks made on the back of the plate could be seen through the agar and membrane. Using a sterile 18-gauge needle the position of the marks was recorded on the second lift. After lifts were taken, the membranes were placed plaque side up on Whatman 3MM paper until denaturation. DNA was denatured by placing the respective filters on Whatman 3MM paper soaked in denaturation solution (0.5 M NaOH, 1.5 M NaCl) for 1 minute. Filters were then neutralised by transferring to Whatman 3MM paper soaked in a solution of 1.5 M NaCl, 0.5 M Tris-HCl pH 7.5 or 8.0, for 5 minutes. After neutralisation, filters were rinsed in 2 X SSC and placed DNA side up on Whatman 3MM paper. All denaturation, neutralisation and washing steps were carried out in plastic trays. Once all filters had been treated in this way, DNA was fixed to the membrane by baking at 80°C for 1 hour.

D) Prewash, prehybridisation and hybridisation

Baked filters were rehydrated by submersion in a solution of 6 X SSC for 5 minutes. Filters were subsequently transferred to a plastic tray containing prewash solution and incubated for a minimum of 1 hour at 42°C. Usually, however, filters were incubated overnight at 42°C in prewash buffer. After prewashing, membranes were prehybridised in standard hybridisation mix for a minimum of 1 hour at 42°C, except

where special conditions were used. Filters were prehybridised and hybridised in plastic petri dishes equivalent in size with the filters. Typically, 10 filters were prehybridised and hybridised together. The volume of hybridisation solution utilised was simply the minimum amount that was required to cover all filters. Each filter was added individually into the prehybridisation solution. In order to prevent drying of the top filter, and also as a control for the hybridisation, a blank filter was used as the top filter in each dish. Radiolabelled probe was treated as previously described and added to the dish containing the hybridisation solution. Filters were removed from the solution, the probe was added and the filters were added back sequentially. Hybridisation proceeded overnight at 42°C on a rocking platform. Stringency washes were performed as previously described. Washed filters were placed DNA side up on Whatman 3MM paper cut to the size of an autoradiography cassette. Filters were fixed to the paper with tape, and orientation markers were placed on the paper using India Ink to which a small amount of P^{32} had been added. The paper and filters were wrapped in clingfilm and were autoradiographed overnight at -70°C in the presence of intensifier screens.

E) Identification of positive plaques and replating

Double lifts were taken such that if a spot was identified on the first lift, it was only considered to correspond to a truly positive plaque if it aligned with a spot on the second lift that hybridised with lower signal intensity. The position of the puncture marks that had been made in the filters were marked onto the autoradiographs and used to align the films with the marks on the backs of the agar plates. Plugs of agar were taken from the plates corresponding to the positions of the identified positive plaques. Agar plugs were transferred to 1 ml of SM with a couple of drops of chloroform. Phage particles were allowed to diffuse into the SM medium at room temperature then stored at 4°C. It was assumed that agar plugs contained approximately 1×10^7 pfus. Phage were serially diluted in SM buffer for replating. Three dilutions were made up, such that approximately 25-50, 100 and 500 pfus were plated on each of three plates. Three concentrations were used in order that one would yield a suitable plaque density for use in rescreening. Discrete plaques were desired at this stage. Phage adhesion and plating were carried out as previously described. The number of plaques that arose from the three

respective dilutions permitted a calculation of the phage titre. Lifts were taken, as before, from plates with the desired density of plaques, and were hybridised to the screening probe. Discrete positive plaques were identified, picked and replated. This procedure was repeated until hybridisation resulted in all plaques present on one plate being identified as positive. At this point, such a clone was considered to be pure.

F) Preparation of bacteriophage λ DNA

Two hundred microlitres of plating cells of the appropriate host strain were added to approximately 1 X 10^7 plaque forming units (pfus) of bacteriophage and incubated at 37°C for 20 minutes with occasional agitation. In general, two approximate concentrations of phage were utilised, which ranged from 5 X 10^6 to 2 X 10^7 pfus, depending upon the titre. Ten mililitres of LB pH 7.5 containing 10 mM $MgSO_4$ were added to each tube and the tubes were cultured overnight at 37°C with shaking. After overnight culture, a large amount of cell debris was present in the medium, indicative of cell lysis. One hundred microlitres chloroform were added to each tube and the tubes were returned to the incubator at 37°C for an additional 10 minutes. Cells and cell debris were pelleted by centrifugation at 10,000 rpm for 10 minutes. The supernatant, containing bacteriophage particles was removed, loaded into polyallomer tubes and phage particles were recovered by ultracentrifugation at 40,000 rpm for 40 minutes in a Beckman 50 Ti rotor. The resultant phage pellets were resuspended in 250 µl TE buffer pH 8.0 and transferred to microcentrifuge tubes. Two point five microlitres of 0.5 M EDTA pH 8.0 and 2.5 µl of 10% (w/v) SDS were added to each tube and the tubes were incubated at 68°C for 15 minutes to break open the phage particles. Tubes were cooled to 37°C and 30 µl of a 10 mg/ml solution of predigested pronase were added. Tubes were incubated at 37°C for at least 1 hour. Proteins were removed by one extraction with an equal volume of Tris-buffered phenol, followed by two extractions with chloroform: iso-amyl alcohol (24:1). Bacteriophage λ DNA was precipitated from the aqueous layer by the addition of 100 µl 7.5 M NH_4OAc and 1.0 ml 100% ethanol, and collected by centrifugation in the bench-top microcentrifuge for 20 minutes. The pellet was allowed to dry and was redissolved in 200 µl TE buffer pH 8.0. DNA was reprecipitated by the addition of 0.1 volumes of 3 M NaOAc pH 5.5 and 2.5 volumes of 100% ethanol. DNA was recovered by centrifugation as described above. The final pellet was washed in 70%

ethanol, dried and dissolved in a final volume of 15 µl TE buffer pH 8.0. Four microlitres of bacteriophage λ DNA were sufficient for restriction enzyme analysis.

G) Direct plaque PCR of bacteriophage λ

Plaques (not necessarily pure) were picked as agar plugs into 250 µl of 2X PCR buffer in microcentrifuge tubes. Tubes were mixed briefly by vortexing and 25 µl aliquots were taken for PCR amplification in a final volume of 50 µl. Conditions used in the PCR reactions were as follows: 10 mM Tris-HCl pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 1 µM each oligonucleotide primer, 200 µM dNTPs, 2% (v/v) deionised formamide, 2.5 units Taq DNA polymerase, overlaid with mineral oil (Sigma). Prior to the addition of enzyme, each tube was heated to 95°C for 10 minutes, then cooled on wet ice. A variety of primer combinations were utilised to specifically amplify the entire insert or portions of it.

H) λ ZAP II excision

The Lambda ZAP II vector (Stratagene, USA) has been constructed in such a way as to allow the *in vivo* excision and recircularisation of any cloned insert, contained within the lambda vector, to form a pBluescript SK- phagemid containing the cloned insert. This process is dependent on co-infection of host cells with helper phage. The helper phage provide specific proteins that bind to sequences present within the λ ZAP vector, initiating DNA replication of the pBluescript phagemid sequence in addition to any insert that may be present. As the pBluescript vector contains M13 sequences, the helper phage also packages the phagemid as a filamentous phage. Thus, co-infection of a host bacteria with a λ ZAP clone and helper phage allows the rescue of pBluescript phagemid with any insert that is present. The phagemid can be used to transform *E. coli*, and cells carrying the pBluescript phagemid can be directly selected for by culture on LB-Agar supplemented with ampicillin. This procedure allows the rapid sub-cloning of inserts from phage clones and means that bacteriophage λ DNA preparations are not necessary with this system.

Bacteria were streaked out on LB-Agar plates and incubated upside down overnight at 37°C. For λ ZAPII, two bacterial host strains were used. XL-1 Blue (Stratagene, USA) was used for infection, and SOLR

(Stratagene, USA) was used for the subsequent transformation with rescued phagemid. Single colonies of both bacterial strains were inoculated into 5 ml of LB pH 7.5, supplemented with 0.2% maltose and 10 mM $MgSO_4$, and were incubated overnight at 30°C with continuous shaking. The lower temperature was chosen so that cells would not overgrow. Cells were cultured until they reached an OD_{600} of approximately 1.0, then were pelleted by centrifugation at 2,000 x g for 10 minutes at 4°C. The pellet of cells was resuspended in 0.5 volumes of 10 mM $MgSO_4$. An agar plug containing a postive plaque was taken from an agar plate and transferred to a sterile microfuge tube containing 500 µl of SM buffer with 20 µl chloroform. Phage particles were allowed to diffuse into the buffer for 1-2 hours at room temperature. In a 15 ml sterile tube (Falcon, USA), 200 µl of OD_{600}= 1.0 XL-1 Blue cells were added to 100 µl of phage stock (containing > 1 X 10^5 phage particles) and 1 µl of ExAssist helper phage (Stratagene, USA) (>1 X 10^6 pfu/ml). The mixture was incubated at 37°C for 15-20 minutes before adding 3 ml of 2X YT medium and incubating at 37°C for 3-5 hours with continuous shaking. After 3-5 hours at 37°C, the mixture was placed at 70°C for 20 minutes to kill bacteria. Bacteria and cell debris was pelleted by centrifugation at 4,000 x g for 15 minutes. The supernatant containing rescued phagemid was decanted to a fresh tube. In order to plate the rescued phagemid, 1 µl and 50 µl of the phage stock were added to each of two tubes containing 200 µl of SOLR cells (at OD_{600}=1.0). Tubes were incubated at 37°C for 15 minutes. 100 µl of cells from each tube were plated onto LB-Agar plates supplemented with 100 µg/ml ampicillin. Plates were incubated upside down at 37°C overnight. Single ampicillin resistant bacterial colonies were inoculated into 5 ml LB with ampicillin (100 µg/ml) and incubated overnight at 37°C with continuous shaking (approximately 300 rpm). Plasmid mini-prep DNA was prepared as previously described.

2.911 Screening cosmid libraries

A) Plating out the library

Cosmid libraries prepared from both Balb/c and 129 Sv genomic DNA were screened in order to obtain genomic clones containing the mouse Muc-1 gene. Libraries were screened according to described methods (Little, 1985; Poustka, 1985). Libraries were plated onto nylon membranes (Pall Biodyne, Glen Cove, New York) on 15 cm or 24 x 24 cm

LB-Agar plates, supplemented with ampicillin at 100 μg/ml. Plates were made three days in advance and stored at 4°C. Prior to plating the library, plates were dried at 37°C. Nylon membranes were cut to the correct size and laid onto the surface of the agar. Libraries were diluted in LB with ampicillin, such that 1 ml of library would yield approximately 1 X 10^5 bacterial colonies. This number of colonies was found to be sufficient to give a high density of small colonies on a megaplate (24cm x 24 cm). One mililitre of the library was spread onto single 24 cm x 24 cm plates with a sterile flamed glass spreader and the plates were incubated upside down at 37°C overnight.

B) Replica plating

For every plate, three extra LB-Agar/ampicillin plates were poured. These were required for incubation of the replica filters and the master filter after the lifts had been taken. Plates were dried at 37°C before use. Replica filters were obtained from the master filters as follows: The master filters were lifted carefully off the plates, using blunt-ended forceps, and placed colony side up on a sheet of Whatman 3MM paper. A replica filter was carefully positioned on the master filter and an additional sheet of moistened Whatman 3MM paper was placed over both filters. Pressure was applied over the filters using a 2 litre glass conical flask with soft cleaning cloths taped to the base. The flask was moved about over the filters, applying presssure. As with λ cDNA library screening, two lifts were taken. The first lift was left for one minute on the master filter. After one minute the Whatman paper was removed and several holes were made through both filters using a sterile disposable needle to provide orientation markers for identification of positive colonies after screening. Each needle hole was touched with the sharp tip of a fine-pointed marker pen such that the marks were transferred through to the master filter. The first lift was carefully lifted off the master and placed colony side up onto a fresh LB-Agar/ampicillin plate. The second lift was taken using the same procedure, except that it was left on the master for 2 minutes. The positions of the orientation markers were transferred to the second lift by laying both the master and the second lift together on a light box. The black marks that were made to mark the needle holes could be seen through both filters. A needle was used to mark the holes on the second lift, and again the needle holes were marked over with a black pen. As for the first lift, the second lift

was carefully removed from the master and subsequently transferred colony side up to a fresh LB-Agar/ampicillin plate. All three plates (master and first and second lifts) were returned to the 37°C incubator for a further 4 hours.

C) Prehybridisation, hybridisation and identification of positive colonies

After growth, the master plate was sealed with Parafilm and stored at 4°C. Alternatively, master filters can be stored frozen at -70°C. The replica filters were carefully lifted off the plates and treated individually as follows: Filters were denatured by placing them onto the surface of Whatman 3MM paper, presoaked in 0.5 M NaOH, for 7 to 10 minutes. After denaturation, filters were neutralised by first transferring them to the surface of 3MM paper presoaked in 1.0 M Tris-HCl pH 7.4 for 2 minutes, then transferring them to the surface of 3MM paper presoaked in 1.5 M NaCl for 10 minutes. Filters were then placed on dry Whatman 3MM paper and were wiped gently with sterile gauze in order to remove bacterial cell debris. DNA was fixed to the membrane by baking at 80°C for 1 hour. Prior to prehybridisation, filters were presoaked in a solution of 6 X SSC for 5 minutes.

i) Prehybridisation and hybridisation by the method of Church, 1984

Filters were prehybridised at 65°C for 4 hours in approximately 20 ml hybridisation solution per filter (this volume was used for 24cm x 24cm filters). Radiolabelled probe, labelled to a specific activity of $>1 \times 10^8$ dpm/μg, was added to the hybridisation solution. As for the λ library screening, each filter was removed from the hybridisation solution, the probe was added to the solution and each filter was added back in turn, such that each filter received good coverage of probe. Filters were hybridised overnight at 65°C in a volume of approximately 10 ml per filter. After overnight hybridisation, filters were washed four times at room temperature in Church low stringency wash buffer. Filters were then washed twice at 65°C in high stringency wash buffer. Filters were exposed to X-ray films overnight, at -70°C with intensifier screens. Double lifts allowed the unambiguous identification of positive colonies (See Figure 3.4).

ii) Prehybridisation and hybridisation using standard Southern hybridisation conditions

Filters were prehybridised and hybridised as described previously for the λ library filters. Prehybridisation proceeded for 4 hours at 42°C with approximately 20 ml of standard hybridisation solution per 24 x 24 cm filter. Radiolabelled probe was added to the hybridisation solution as previously described (approximately 10 ml per filter), and hybridisation was carried out overnight at 42°C. Stringency washes were as described for Southen blots. Filters were exposed to X-ray films overnight at -70°C, with intensifier screens. Positive colonies were identified as described previously (See Fig. 5.51).

D) Replating primary clones

Small squares approximately 0.1 cm^2 in area were cut from the master filter from areas that corresponded to positively hybridising colonies. Squares of membrane were transferred to microcentrifuge tubes containing 1 ml LB with 100 μg/ml ampicillin. From this 1 ml of solution, 2 μl, 5 μl, 10 μl, 30 μl and 50 μl aliquots were plated directly onto fresh LB-Agar/ampicillin plates and incubated overnight at 37°C. Double lifts were taken off the plates and the filters were treated prior to hybridisation, as described above. In general, a total of three rounds of colony purification were utilised before a clone could be demonstrated as being pure.

E) Preparation of cosmid DNA

Bacteria carrying cosmids were streaked out onto LB-Agar/ampicillin plates and incubated overnight at 37°C. Single colonies were inoculated into 50 ml LB with ampicillin and cultured overnight in 500 ml glass conical flasks. After overnight culture, cells were pelleted by centrifugation at 10,000 rpm for 15 minutes at 4°C. Cell pellets were resuspended in 1 ml cosmid alkaline lysis buffer I. Three mililitres of buffer II were added slowly to each bottle, the bottles were mixed and left on wet ice for 20 minutes. Two point five mililitres of buffer III were added to each tube which was then mixed well and left on wet ice for between 30 and 60 minutes. Cell debris was pelleted by centrifugation at 15,000 rpm for 1 hour at 4°C. The supernatant from each tube was

transferred to a fresh tube, and 0.6 volumes of ice-cold propan-2-ol were added. Tubes were mixed briefly, left on wet ice for 5 minutes, then nucleic acids were recovered by centrifugation at 8,000 rpm for 10 minutes at 4°C. The resultant pellets were each resuspended in 700 µl of 2 M NH_4OAc, transferred into individual microcentrifuge tubes, and left on wet ice for 5 minutes. Contaminating cell debris was pelleted by centrifugation in the bench-top microcentrifuge, the supernatant was transferred to a fresh tube and 420 µl ice-cold propan-2-ol were added to precipitate DNA. Tubes were mixed by vortexing, then DNA was recovered by microcentrifugation for 10 minutes. The DNA pellet was washed with 70% ethanol, dried and finally dissolved in 50 µl TE buffer pH 8.0. Typical yields of cosmid DNA from a 50 ml culture were in the range of 40 to 75 micrograms.

As an alternative to the above approach for the purification of cosmid DNA, the TENS plasmid mini-prep protocol was also found to yield high quality cosmid DNA. The only slight modification of the TENS plasmid mini-prep method for use in cosmid DNA preparation was that the initial step of spinning down the bacteria was repeated once. Briefly, 1.5 ml overnight culture were poured into a microcentrifuge tube. Bacteria were recovered by pulse spin in the microcentrifuge, and the growth medium was poured off. An additional 1.5 ml overnight culture were added to the same tube and cells were pelleted again. The TENS mini-prep procedure, as described above, was followed from this point. The average yield of cosmid DNA using this method was in the range of 10 to 25 µg.

2.912 Chromosomal localisation studies

Genomic DNA was prepared from tissue from C3H/HeJ-*gld*, *Mus spretus* and [(C3H/HeJ-*gld* x *Mus spretus*)F_1 x C3H/HeJ-*gld*] backcross mice as described. Approximately 10 µg genomic DNA were digested overnight with the appropriate restriction endonuclease and size-fractionated by electrophoresis through a 0.9% (w/v) agarose gel. DNA was transferred to nylon membrane and hybridised under standard Southern conditions to the respective probes. Washing conditions were carried out under high stringency. Probes were labelled by random priming with [α-32P] dCTP. The reference locus probes utilised were *Cacy* (Ferrari, 1987), *Cd1* (Balk, 1989), *D3Tu51* (Vincek, 1989), *Fcgr1* (Sears, 1990),

Gba (Ginns, 1985), *Pklr* (Tani, 1988) and *Thbs-3* (Vos, 1992). Informative restriction fragment length variants (RFLVs) for the reference locus probes were as previously described (Oakey, 1992b; Kingsmore, in press). Samples were typed by segregation analysis of unique *M. spretus* RFLV detected with these probes (Green, 1981). This work was carried out as part of a formal collaboration with Dr. Stephen F. Kingsmore and Dr. Michael Seldin, Duke University, North Carolina, USA.

2.913 Isolation of RNA

A) Isolation of RNA from cell cultures

Total RNA was prepared by the guanidinium isothiocyanate extraction method as described (Chirgwin, 1979). Cells growing as a monolayer were washed twice with ice-cold PBS. 6 ml of 4 M GTC solution were added per 75 cm^2 flask to dissolve the cells, and the dissolved cell solution was loaded onto a cushion of 5.7 M CsCl in open-topped polyallomer tubes (Beckman). Tubes were placed in their respective buckets and balanced by the addition of 4 M GTC solution. RNA was pelleted by ultracentrifugation in an SW28 rotor (Beckman, USA) at 25,000 rpm overnight at 20°C. After overnight spin, the solution was poured from the tubes and the tubes were inverted to drain. The bottom of each tube was cut off with a heated disposable scalpel. The area around the RNA pellet was carefully washed with 70% ethanol and the tube bottoms were inverted again to drain. One hundred microlitres of sterile distilled water were added to each tube to dissolve the RNA pellet and the solution was then transferred to a sterile RNase free microcentrifuge tube. From this point onwards, tubes were kept on wet ice. RNA was precipitated with 2 volumes of ice-cold RNase free 100% ethanol, recovered by centrifugation, washed with 70% ethanol, air-dried briefly, and finally dissolved in sterile distilled water. The concentration, purity and integrity of the RNA samples was determined by taking OD readings at 260 nm and 280 nm. A small amount of the RNA sample was loaded on a 1% (w/v) agarose-TBE gel and was electrophoresed at room temperature at 5 V/cm in 1X TBE buffer. RNA was diluted to a final concentration of 1 μg/μl, was aliquotted and then stored at -70°C.

B) Isolation of RNA from liquid nitrogen (LN_2) frozen tissues

Frozen tissue samples were wrapped in silver foil and manually crushed with a heavy weight. Small pieces of tissue were transferred to a teflon bomb (Braun, Melsuugen, Germany) containing a tungsten ball. Teflon bombs were soaked overnight in 3% (v/v) hydrogen peroxide solution, then washed several times with sterile distiled water before use. Tungsten balls were soaked overnight in 100% ethanol than washed several times with sterile distilled water before use). Teflon bombs, containing tissue samples, were wrapped with tape and returned to a dewar of LN_2. In order to reduce the tissue to a powder, one teflon bomb at a time was removed from the LN_2, the tape was removed and the bomb was placed in a Braun Mikro-dismembranator II (Melsuugen, Germany). Tissue was pulverised in the dismembranator for 1 minute. After one minute, the status of the tissue was ascertained. If the tissue was not yet powdered effectively the bomb was wrapped in tape again, returned to LN_2 and the pulverisation process was repeated. Once the tissue had been pulverised, the powder was scraped into an open-topped polyallomer tube containing 15 ml 4 M GTC solution and homogenised gently. The homogenate was loaded onto a cushion of 5.7 M CsCl and RNA was recovered by centrifugation as described above.

2.914 Agarose gel electrophoresis of RNA

A) Glyoxal method

RNA samples (10-15 μg) were added to 5 μl of freshly prepared 10 M urea, 15 μl sterile DMSO, and 5 μl glyoxal mix in a microfuge tube. Tubes were incubated at 50°C for 1 hour then allowed to cool to room temperature. One microlitre of 0.1% (w/v) bromophenol blue solution was added to each tube and the samples were introduced into the wells of a 1.3% (w/v) agarose gel prepared in 10 mM sodium phosphate pH 7.0. The gel was cast in a gel-tray and tank that had been previously soaked in a solution of 3% (v/v) hydrogen peroxide and rinsed with sterile distilled water, in order to remove any contaminating RNases. RNA was fractionated by electrophoresis for 3 to 4 hours at 5-7 V/cm in glyoxal-northern gel running buffer. Glyoxal dissociates from RNA at > pH 8.0 and, therefore, in order to maintain the pH of the running buffer, the buffer was continually recirculated after the dye front had entered the gel.

B) Formaldehyde method

RNA was fractionated by electrophoresis through a 1.0% (w/v) agarose gel containing 6% (v/v) formaldehyde. One gram of agarose was dissolved in 62 ml sterile distlled water by boiling in a microwave oven, and after cooling to 60°C was added to 20 ml of 5X formaldehyde gel running buffer and 18 ml formaldehyde (37% (v/v) solution in water). The gel was cast in a gel-tray and tank that had been previously soaked in hydrogen peroxide as described. RNA samples (approximately 15 µg) were denatured by addition of 1 µl 10X MOPS, 3.5 µl formaldehyde (37% (v/v) solution) and 10 µl deionised formamide. Each sample was incubated at 65°C for 15 minutes then chilled on wet ice. After the addition of 2 µl formaldehyde-northern sample buffer, each sample was loaded into the agarose gel and electrophoresed at 5 V/cm for approximately 4 hours in 1X MOPS buffer.

2.915 Northern blotting

At the completion of electrophoresis, RNA was transferred overnight to Biodyne nylon membranes by capillary elution in the presence of 20 X SSC as described (Maniatis, 1982). After transfer, the location of the wells was marked on the membrane with pencil and RNA was immobilised onto the membrane by baking at 80°C for 1 hour. The distance of migration of the RNA and the quality of the RNA were assessed by staining the baked membrane in methylene blue staining solution for 5-10 minutes (Wilkinson, 1991). The stained blot was wrapped in clingfilm and photocopied. Prior to hybridisation the blot was destained in a solution of 1% (w/v) SDS for 10-15 minutes. Glyoxal northerns were prehybridised and hybridised in a heat-sealable plastic bag in standard hybridisation solution at 55°C. Hybridisation proceeded overnight at 55°C. Stringency washes were carried out in standard wash buffers, except that high stringency washes were carried out twice for 30 minutes at 65°C. Specific RNA-DNA duplexes were visualised by autoradiography at -70°C in the presence of intensifier screens. Formaldehyde-northerns were prehybridised and hybridised in heat-sealable plastic bags in formaldehyde-northern hybridisation mix at 42°C. Membranes were prehybridised for 30 minutes, whereas hybridisation proceeded overnight. After overnight hybridisation, membranes were washed three times for 30 minutes at 65°C in formaldehyde-northern

wash buffer. Visualisation of specific RNA-DNA complexes was carried out as described above.

2.916 Polymerase chain reaction (PCR)

Standard PCR amplification reactions used the supplied buffer at 1X concentration. Reactions were generally assembled in a 0.5 ml eppendorf tube (when Hybaid Thermal Reactor was utilised) or in Perkin-Elmer Gene-Amp tubes (Perkin-Elmer Cetus, USA). Standard conditions used were as follows:

10.0 µl 10X PCR buffer
1 µl dNTP mix (stock solution of all four dNTPs at 20 mM); final concentration was 200 µM
5 µl each oligonucleotide primer (stock at 20 µM); final concentration was 1 µM
2% (v/v) deionised formamide
10 ng plasmid template or 100 ng total genomic DNA
0.5 µl Taq DNA polymerase (2.5 units)

Total volume was raised to 100 µl by the addition of sterile distilled water.

The reactions were overlaid with two drops of mineral oil and heated to 95°C for 5 minutes prior to the addition of the Taq polymerase. Reactions were performed in either a Hybaid Thermal Reactor (Hybaid, UK) or a Perkin-Elmer Gene-Amp 9600 (Perkin-Elmer Cetus, USA). Reaction conditions were varied depending upon the calculated annealing temperatures of the respective primers and on the length of the expected product. Similarly, where no product was detectably amplified under normal buffer conditions, the Mg^{2+} concentration of the buffer was modified. Amplified products were analysed by agarose gel electrophoresis. Where no product could be visualised after the first set of cycles, fresh enzyme was added to the tube and the reaction was continued for another 10-15 cycles. If a product was still undetectable at this stage the agarose gel was blotted and the membrane was hybridised with a radiolabelled probe.

2.917 Reverse transcriptase-PCR (RT-PCR)

RT-PCR was utilised to specifically amplify fragments of Muc-1 cDNA from a variety of species, directly from total RNA. Reverse transcriptase PCR reactions were carried out as described in Braga, 1992, except that different primers were utilised. RNA samples (10-15 μg) were heated to 95°C for two minutes in a total volume of 25 μl in the presence of 5 mM Tris-HCl pH 7.5. After heating, tubes were chilled on wet ice then spun briefly in the microfuge to collect the condensate. One tube of RNA (25 μl) was then used for two first-strand cDNA synthesis reactions. Reaction conditions utilised for the reverse transcription were as follows:

> 50 mM KCl, 10 mM Tris-HCl pH 8.3, 2.5 mM MgCl2, 0.8 mM each dNTP, 12.5 pmol primer, 10 units Moloney murine leukaemia virus reverse transcriptase (M-MuLV-RT).

A master mix, containing all the above components, was prepared and aliquotted into tubes on ice, such that after heating, 12.5 μl of the RNA solution could be directly added to the tubes and reverse transcription could proceed immediately. Tubes were incubated at 42°C for at least one hour. At the completion of first-strand synthesis, 75 μl of a PCR master mix were added to each tube to make a total volume of 100 μl. The final conditions utilised for the specific PCR amplification of first-strand cDNA were identical to those described above for standard PCR except that the final concentration of Mg^{2+} in the reaction was 1.75 mM, as opposed to the standard 1.5 mM. Again, as for the PCR conditions described above, the cycles of the PCR amplification were set according to the respective annealing temperatures of the primers being utilised and the length of the expected product.

2.918 Direct cloning of PCR products into PCR T-vectors

Initially, PCR primers were designed with sites for specific restriction endonucleases at their 5' termini such that amplified products could be digested and ligated into cloning vectors cut with the same enzymes. However, many restriction enzymes have been shown to cut sites present at the ends of DNA fragments very inefficiently (Kaufman, 1990) and, therefore, a different strategy was adopted for the direct cloning

of PCR amplified products. Taq DNA polymerase has been found to possess a template-independent terminal transferase activity which results in the addition of a single nucleotide at the 3' end of the amplified fragment (Clark, 1988; Mole, 1989). This nucleotide has been demonstrated to be almost invariably an adenosine, due to the strong preference of Taq polymerase for dATP (Clark, 1988). This template-independent activity was exploited by Marchuk, 1991, to create a direct PCR cloning scheme with the efficiency of sticky-end cloning.

In order to directly clone PCR amplified products, a T-vector was prepared as described (Marchuk, 1991). pBluescript vector DNA was digested with the restriction endonuclease EcoRV, which creates blunt ends, and ethanol precipitated. One-twentieth of the digested plasmid DNA was analysed by agarose gel electrophoresis to determine if the vector had been completely linearised. The linearised vector DNA was incubated with Taq DNA polymerase (1 unit/μg plasmid/20 μl volume) using standard PCR buffer conditions in the presence of 2 mM dTTP for 2 hours at 70°C. The absence of any other nucleotide in the reaction resulted in the addition of a single thymidine at the 3' end of each fragment. The T-vector plasmid DNA was purified by a single phenol: chloroform: iso-amyl alcohol extraction and precipitated with 2.5 volumes ice-cold ethanol. The resultant DNA pellet was resuspended in 20 μl TE buffer pH 8.0 and 1/10th was analysed by agarose gel electrophoresis to estimate the amount of DNA. T-Vector DNA was diluted to a final concentration of 10 ng/μl, aliquotted and stored at -20°C.

PCR amplified products were efficiently ligated into the T-vector directly from the unpurified PCR reaction. Alternatively, fragments were gel-purified, as described, and ligated into the T-vector. Gel-purification of specifically amplified PCR products was only necessary when small amounts of non-specifically amplified products were observed to have been amplified in addition to the specific product. Ligated DNA was used to transform competent XL-1 Blue cells which were subsequently plated onto LB-Agar/ampicillin plates supplemented with X-gal and IPTG.

2.919 Sequencing of plasmid DNA: Sequencing reactions and electrophoresis

A) Sequencing reactions

Double-stranded DNA to be sequenced was denatured in one of two ways, to allow annealing of a specific oligonucleotide primer. In general, approximately 5 μg plasmid DNA was denatured by incubating in a total volume of 20 μl 0.2M NaOH for 5-10 minutes at room temperature. Alternatively, mini-prep plasmid DNA was denatured in the presence of 0.2 M NaOH at 65°C for 10 minutes. The denatured DNA was ethanol precipitated by the addition of 8 μl of 5 M NH_4OAc pH 5.4 and 100 μl ice-cold 100% ethanol and recovered by centrifugation in a bench-top microfuge for 10 minutes. After washing in 70% ethanol, the pellet was resuspended in 10 μl of 1X Sequenase reaction buffer with 5 ng of a specific sequencing primer. The primer was annealed to the template by heating at 65°C for 5 minutes followed by gradual cooling to room temperature. Sequencing reactions were performed using a Sequenase™ version 2.0 kit (United States Biochemical Corp.) as recommended by the manufacturer. This method is a modification of the original dideoxy chain-termination protocol described by Sanger, 1977, but replaces the Klenow fragment with a modified DNA polymerase from the T7 bacteriophage. The kit was used in conjunction with [α-^{35}S]dATP (1000 Ci/mmol) (Amersham International).

B) Electrophoresis and autoradiography

The products of the sequencing reactions were analysed by electrophoresis through denaturing polyacrylamide gels. The gel mix was prepared as described, and poured between two glass plates separated by 0.2 mm thick wedge spacers and a 0.2 mm thick well former. The gel was allowed to polymerise for at least 3 hours at room temperature, but was generally left overnight to ensure complete polymerisation. A sharkstooth comb was carefully inserted between the two glass plates, to form the wells for loading. The notched glass plate was siliconised with Repelcote (BioRad) to ensure that the gel didn't adhere to this plate when they were separated. Formamide stop solution, from the Sequenase kit, was introduced into alternate wells in order to check for any leaks that may have been present and also to identify any potential problem wells.

Gels were prerun for 20-30 minutes in 1X TBE buffer at a constant 45 Watts. Sequencing reaction samples were heated for 2 minutes at 95°C before loading. Reactions were electrophoresed at a constant 50 Watts for the desired length of time. In general, gels were run until the bromophenol blue dye front had run off the gel and then for a further 15-30 minutes. After electrophoresis, the glass plates were separated, leaving the gel on the non-notched plate. The gel was fixed for 20 minutes in a solution of 10% (v/v) methanol, 10% (v/v) glacial acetic acid, then transferred to a sheet of Whatman 3MM paper. The gel was vacuum-dried at 80°C for 30-60 minutes and the radiolabelled products were visualised by autoradiography.

2.920 SDS polyacrylamide gel electrophoresis of proteins

Sample preparations and electrophoresis were performed according to Laemmli, 1970. Vertical slab gels were prepared from two solutions, forming the running, or resolving gel, and the stacking gel, respectively. The running gel routinely contained 5% (w/v) acrylamide and was prepared as described. The gel mixture was poured between two glass plates separated by 0.75 mm spacers, with sufficient space left for the stacking gel, then overlaid with water-saturated butan-1-ol. After polymerisation, the butan-1-ol was poured off and the gel was rinsed several times with sterile distilled water. The stacking gel mix, containing 3% (w/v) acrylamide, was poured onto the top of the running gel and was allowed to polymerise around a well-former. Protein samples were added to an equal volume of 2X SDS-protein sample buffer and boiled for 2-5 minutes before loading into the wells. Rainbow molecular weight protein markers (Amersham International) were run alongside the protein samples to provide a reference. Gels were run at low voltage (30-50 V) in SDS-PAGE buffer until the bromophenol blue dye front reached the running gel, at which point the voltage was raised to >100 V. The gel was run until the desired separation was achieved.

2.921 Detection of milk-fat-globule associated Muc-1 protein by silver staining

Detection of Muc-1 glycoprotein in SDS-PAGE gels proceeded according to the method described by Morrissey, 1981. All gel handling was carried out wearing gloves that were washed with water before

touching the gel. After electrophoresis, the two glass plates were separated and the gel was prefixed in a solution of 50% (v/v) methanol, 10% (v/v) glacial acetic acid for 30 minutes, followed by a solution of 5% (v/v) methanol, 7% (v/v) glacial acetic acid for 30 minutes. The gel was then fixed for 30 minutes in 10% (v/v) gluteraldehyde. After fixation, the gel was rinsed in several changes of distilled water and generally left to soak overnight in an excess of water. The gel was soaked for 30 minutes in a 5 μg/ml solution of dithiothreitol (DTT) which was then poured off and replaced by a solution of 0.1% (w/v) silver nitrate. The gel was treated with silver nitrate solution for 30 minutes then rinsed with distilled water. Developing solution consisted of 50 μl of formaldehyde (37% v/v) solution) in 100 ml 3% (w/v) Na_2CO_3. Gels were placed in developing solution until protein bands became visible (generaly about 10-15 minutes). The staining was stopped by the addition of 5 ml 2.3 M citric acid per 200 ml developer and agitating for 10 minutes. Gels were soaked in 0.03% (w/v) Na_2CO_3 (to prevent bleaching) and then either photographed directly or dried down under vacuum onto Whatman 3MM paper at 80°C for one hour.

2.10 CELL METHODS

2.101 Growth and maintenance of cells

Cells were routinely grown as monolayer cultures on Falcon tissue culture dishes at 37°C in a humidified atmosphere supplemented with 10% CO_2. On reaching sub-confluence, cell cultures were subcultured. The growth medium was removed by aspiration and the cell monlayer was washed once with PBS. A solution of prewarmed (37°C) 1X trypsin/EDTA (Gibco-BRL) was added to the cell cultures which were subsequently returned to the incubator. After 2-5 minutes, cultures were removed from the incubator and tapped to dislodge cells. Cell clumps were dispersed by pipetting up and down several times with a sterile cotton-plugged glass pasteur pipette, and trypsin was inactivated by the addition of ≥ 1 volume of prewarmed complete medium. The cell suspension was again pipetted up and down before the cells were reseeded into fresh tissue culture dishes.

Cells were prepared for freezing by pelleting at 1,000 rpm in the bench-top centrifuge and then resuspended at the desired concentration.

An equal volume of 2 X freezing medium, consisting of 60% complete medium supplemented with 20% (v/v) dimethyl sulphoxide (DMSO) and 20% (v/v) FBS, was added to the cell suspension which was then pipetted into prelabelled cryovials (Nunc, UK). Vials were wrapped in paper to prevent rapid freezing and placed at -70°C overnight before transfer to liquid nitrogen for long-term storage.

Cells were recovered from liquid nitrogen by rapid thawing at 37°C in a water bath. Freezing vials were placed in a 37°C water bath to thaw, swabbed with 70% ethanol, and the cells were either directly added to tissue culture flasks containing an excess of the appropriate medium (to dilute out the toxic effects of the DMSO) or added to prewarmed medium in a sterile disposable tube (Falcon, USA) and pelleted by centrifugation at 1,000 rpm for 5 minutes. The cell pellet was then resuspended in an appropriate volume of prewarmed medium before adding the cells to culture flasks.

2.102 Growth and maintenance of ES cells

The mouse embryonic stem cell lines, E14TG2a, derived from a male blastocyst of mouse strain 129/Ola (agouti/chinchilla) (Hooper, 1987) (provided by Dr. Martin Hooper, University of Edinburgh), and GK129, derived from a male blastocyst of mouse strain 129/Ola/Hsd (agouti/chinchilla) (Philpott, 1992) (provided by Dr. Graham Kay, Medical Research Council, Clinical Research Centre, Harrow, UK), were routinely cultured on a layer of mitotically inactivated STO-neo fibroblast feeders on gelatin coated (0.1% (w/v) in sterile distilled water) tissue culture plates, in ES medium supplemented with 10^3 units/ml of recombinant murine leukaemia inhibitory factor (LIF). Medium was replaced daily. Cells were routinely maintained in 75 cm^2 (T75) flasks and seeded at an approximate cell density of 3 X 10^6 ES cells per flask. Under these conditions, cells that were seeded on Day 1, were generally at a sub-confluent density and ready to be subcultured on Day 3. ES cells were frozen at a concentration of approximately 3.5 X 10^6 ES cells/ml/cryovial. Upon thawing, it was assumed that approximately 3.0 X 10^6 cells would recover.

2.103 STO-neos and the production of feeder cells

STO-neo fibroblasts are a neomycin resistant subline of the thioguanine and ouabain resistant line of SIM mouse fibroblasts originally isolated by Dr. A. Bernstein (Martin, 1975). STO-neo cells (provided by Dr. Mark Hertle, University of Chicago) were routinely cultured in Dulbecco's modified Eagle's medium (DMEM) containing 10% (v/v) foetal bovine serum (FBS) and 5 mM L-glutamine. Cells were grown on gelatinised tissue culture plates. STO-neo cells were routinely seeded into 175 cm^2 (T175) tissue culture flasks at a density of 1 X 10^6 cells and grown for a further three days. Cells were grown at a density such that they were maintained as a monolayer in culture. If STO-neo cells were observed to overgrow and start to pile-up, losing contact inhibition, the cultures were discarded and a fresh vial of cells, from an earlier passage, was recovered from LN_2. When STO-neos were growing at the desired growth rate, T175s yielded approximately 1.2 X 10^7 cells. STO-neos were routinely frozen at a concentration of 1 X 10^6 cells/ml/vial.

STO-neo feeder cells were prepared as follows: The medium was aspirated from cultures growing at approximately 75-80% confluence (approximately 1.2 X 10^7 cells per T175) and replaced with DMEM containing 5% (v/v) FBS and 10 μg/ml mitomycin C (Sigma Chemicals Ltd). Cells were cultured in the presence of mitomycin C at 37°C for between 2 and 6 hours. After mitotic inactivation, cultures were washed with an excess of PBS and trypsinised at room temperature with prewarmed trypsin/EDTA. Cells were dislodged and typsin was inactivated by the addition of an equal volume of prewarmed DMEM/10% FBS/glutamine. In general, 10 T175s were treated at the same time. Cells from 5 T175s were pooled into a single 50 ml sterile disposable tube (Falcon, USA) recovered by centrifugation at 1,000 rpm for 5 minutes and washed three times with PBS (i.e. the cell pellet was resuspended in PBS then recovered by centrifugation at 1,000 rpm for 5 minutes for each wash) then resuspended in 5 ml DMEM/10% FBS/glutamine. An equal volume of 2X freezing medium was added to the cell suspension and cells were frozen at an approximate concentration of 6 X 10^6 cells/ml/vial. This concentration of feeder cells was set, assuming a recovery from freezing of approximately 65-70%. For growth of ES cells, feeder cells were plated onto gelatin-coated tissue culture plastic at a density of approximately 5 X 10^4 cells/cm^2. A

percentage recovery of feeder cells of 65% would give a viable feeder cell number of approximately 4 X 10^6 (from 6 X 10^6 frozen cells) and this was found to be sufficient to provide a good feeder monolayer for a T75 flask, at a density of approximately 5 X 10^4 feeder cells/cm^2. Feeder cells were normally plated onto gelatin-coated plates one day before ES cells were to be plated onto them. However, once plated, feeders remained viable for plating ES cells for about 7 days.

2.104 Electroporation of ES cells and selection of resistant colonies

A) Optimisation of electroporation conditions for ES cells

ES cells were trypsinised and recovered by centrifugation at 1,000 x g for 5 minutes. The cell pellet was resuspended in PBS and the cell density was determined by counting on a Neubauer haemocytometer. In order to determine the absolute number of ES cells present, a figure, corresponding to the estimated number of feeder cells present, was subtracted from the total cell count. For these purposes it was assumed that T75 and T175 flasks would yield approximately 4 X 10^6 feeder cells and 1 X 10^7 feeder cells, respectively.

Prior to the electroporation experiments, embryonic stem cell cultures were grown for 10 days in the presence of Geneticin (G418) at concentrations ranging from 100 μg/ml (active constituent) up to 500 μg/ml. This was in order to determine the optimal concentration of G418 that would be necessary for selection purposes. Electroporation conditions were optimised, around published values, by transfection of a firefly luciferase gene reporter construct (Andreason, 1989) into ES cells. Approximately 5 μg of reporter plasmid was electroporated into ES cells in PBS, at a concentration of 4 X 10^7 ES cells/800 μl using a BioRad Gene-pulser at a constant capacitance of 250 microFarads (μF), and voltages ranging from 150 to 400 V. After electroporation, cells were placed on wet ice for 10 minutes then plated onto 90 mm dishes containing feeders. The efficiency of electroporation was determined 48 hours after electroporation, by a calculation of luciferase activity of the respective cell cultures, relative to total protein (determined with a BioRad protein assay kit). After optimisation of electroporation conditions, mouse embryonic stem cell line, E14TG2a, was routinely electroporated in PBS

using a BioRad Gene-pulser at 350 V and 250 μF, at an ES cell density of approximately 4 X 10^7 ES cells/800 μl.

In preparation for the electroporation, plasmid DNA was digested overnight as described. One-tenth of the digest was analysed by agarose gel electrophoresis to determine if the plasmid DNA had been completely linearised. Linearised plasmid DNA was purified by extraction with an equal volume of Tris-buffered phenol: chloroform: iso-amyl alcohol (25: 24: 1) and ethanol precipitated. DNA was recovered by centrifugation in a bench-top microfuge, washed with 70% ethanol, dried, and dissolved in sterile TE buffer pH 8.0. Linearised plasmid DNA was electroporated into the ES cells at a concentration of 5 nM (nanomolar). After electroporation, cells were placed on wet ice for 10 minutes before plating onto ten 90 mm tissue culture plates of fresh feeders (approximately 4 X 10^6 electroporated ES cells per plate).

Mouse embryonic stem cell line, GK129, was electroporated using a BTX electroporator (BTX, USA). 4 X 10^7 ES cells were electroporated in 800 μl PBS at a voltage of 250 V and a capacitance of 500 μF. After electroporation, cells were treated as described above.

B) Selection

Twenty-four hours after the electroporation, ES medium, containing G418 at an active concentration of 400 μg/ml, was added to the cultures. In some cases gancyclovir, at a concentration of 2 μM, was also added to the culture medium in an attempt to enrich for homologous recombinants. When gancyclovir was utilised, two of the ten plates received medium with G418 only. This was to enable a calculation of the relative enrichment factor achieved using gancyclovir. As a control, ES cells were electroporated under the same conditions as described above, but with buffer only. Selection proceeded for 9-11 days. For the first four days, fresh selective medium was added daily. After this point, fresh selection was added every other day.

C) Picking resistant embryonic stem cell colonies and identification of correctly targeted embryonic stem cell clones by Southern blotting

Resistant embryonic stem cell colonies started to become visible after 6-7 days of selection and were picked after 10-11 days. Plates were washed with PBS and colonies were picked into individual wells of a 96-well plate, containing 100 µl prewarmed trypsin/EDTA solution, by scraping around the colony and sucking the colony up into a volume of 10 µl PBS using a Gilson pipetman. After twelve colonies were picked in this manner, cells were dispersed in the trypsin solution by pipetting up and down several times. Trypsin was inactivated by the addition of 100 µl of prewarmed ES medium, and the cells from a single 96-well were transferred to a single well of a 24-well plate containing fresh feeders in selective medium. After 4-5 days of growth, wells reached sub-confluent levels, at which time the ES cells were subcultured into two separate wells of a 24-well plate. One of these wells contained fresh feeder cells, whereas the other was gelatin-coated only. Cells were grown for a further 2-3 days. Upon reaching sub-confluent levels, cells growing in the presence of a feeder layer were frozen in a single freezing vial as described. Cells growing without feeders were used to make genomic DNA as described. These cells were grown in the absence of a feeder layer so that genomic DNA preparations would consist primarily of ES-derived DNA.

Genomic DNA preparations were dissolved in a final volume of 20 µl sterile distilled water and were digested with the restriction endonuclease, EcoRI, in a final volume of 30 µl overnight at 37°C under conditions recommended by the manufacturer. DNA digests were fractionated by agarose gel electrophoresis through a 0.7-0.8% (w/v) agarose gel, alongside λHindIII molecular weight markers. DNA was transferred onto nylon membrane as described and hybridised, as described, with a specific flanking radiolabelled probe. Positive diagnosis of correctly targeted ES clones was performed after autoradiography and was dependent upon the presence of specifically hybridising fragments of the predicted molecular weight. ES clones that were identified as being correctly targeted were thawed and plated into single wells of a 24-well plate containing fresh feeders. Cells were gradually expanded, until they were cultured in T75 flasks. At this point, several vials of each ES clone were frozen.

2.105 *In vitro* differentiation of ES cells

ES cells are pluripotential and, as such, they have the ability to differentiate to form a wide variety of cells characteristic of certain tissues. The pluripotentiality of ES cells is normally maintained in tissue culture by growth on a layer of mitotically inactivated fibroblast feeders and/or growth in the presence of the differentiation inhibitor leukaemia inhibitory factor (LIF). However, ES cells can also be specifically induced to differentiate in culture.

The induction of differentiation of specific ES clones was induced by seeding freshly trypsinised ES cells into 90 mm bacteriological culture plates at a density of 2 X 10^6 cells in ES medium without 2-mercaptoethanol and LIF. Under these conditions, ES cells form clumps and grow in suspension culture. After 3-4 days in suspension culture, ES cells differentiated to form structures described as simple embryoid bodies, and upon further culture, differentiated to form structures described as cystic embryoid bodies. Both the simple embryoid bodies and the cystic embryoid bodies were returned to gelatin-coated tissue culture plates and cultured further. According to Hebert, 1990, 3-6 days in suspension culture is roughly equivalent to days 4.5-6.5 of embryonic development (vaginal plug= day 0.5), 12-15 days in suspension culture is roughly equivalent to days 6.5-8.5 of embryonic development, and 25 days in culture (plated onto gelatin-coated plates at approximately day 10-12) is roughly equivalent to days 8.5-10.5 of embryonic development.

In addition to being utilised as a method for assessing the *in vitro* differentiation potential of specific ES clones, embryoid bodies were utilised as a good source of ES genomic DNA.

2.106 Karyotype analysis of embryonic stem cells

In order to determine if targeted embryonic stem cell clones retained the normal 2n=40 diploid mouse chromosome constitution, the karyotype was determined.

A) Preparation of chromosome spreads

Fresh medium was added to exponentially growing embryonic stem cell cultures. After 2-3 hours, colcemid (stock 5 μg/ml stored at -20°C) was added to embryonic stem cell cultures to a final concentration of 0.05 μg/ml. Cultures were returned to the incubator for approximately 1 hour. After colcemid treatment, cultures were trypsinised and cells were recovered by centrifugation at 1,000 rpm for 5 minutes in a bench-top centrifuge. The supernatant was removed by aspiration and the cell pellet was disrupted by tapping the tube. One mililitre of a 0.56% (w/v) KCl solution was added to the cells which were resuspended by flicking the tube. An excess of KCl (6 ml) was added to the tube which was inverted to mix, then left at room temperature for 10 minutes. Cells were recovered by centrifugation at 500 rpm for 5 minutes. KCl solution was removed by aspiration and the cell pellet was disrupted by flicking the tube. Five mililitres of ice-cold fixative (75% (v/v) methanol, 25% (v/v) glacial acetic acid) were added to the cell pellet from a pasteur pipette while flicking the tube to prevent clump formation. The cells were fixed for 5 minutes at room temperature then recovered by centrifugation at 500-1,000 rpm for 5 minutes. The fixation procedure was repeated three times and the final cell pellet was resuspended in a volume of 0.5-1.0 ml fixative. The cell suspension was added, dropwise, onto grease-free slides (soaked in 1 N HCl overnight, rinsed with several changes of sterile distilled water, and wiped with 70% ethanol immediately before use) and allowed to air-dry.

B) Staining chromosomes

To count chromosomes only, slides were stained in a solution of 3% (v/v) Gurr's Giemsa in PBS for 10-15 minutes. After staining, slides were rinsed in two changes of sterile distilled water and allowed to air-dry. For photography, slides were mounted with a coverslip in DepeX mountant, and photographed using an oil immersion objective.

To G-band chromosomes, slides were first incubated at 60°C in a bath of 2X SSC for 1 hour. Slides were rinsed extensively in several changes of sterile distilled water and stored under water in a metal staining rack until ready to be used. Slides were immersed individually in trypsin solution (0.25% trypsin in Gurr's buffer pH 6.8) for 4-5 minutes

then rinsed with Gurr's buffer pH 6.8. Chromosomes were stained by immersing in a solution of 3% (v/v) Gurr's Giemsa in PBS for 5 minutes, then rinsed in two changes of Gurr's buffer pH 6.8 followed by two changes of sterile distilled water, and allowed to air-dry. For photography, slides were treated as described above.

2.107 Calcium-phosphate transient transfection of mammalian cells

DNA was introduced into ZR-75-1, HP-1 and HT1080 cells by calcium-phosphate precipitation (Graham, 1973). Cells were seeded, 18-24 hours prior to transfection, at an approximate density of 1.5 X 10^6 cells per 90 mm dish and fresh medium was added to the cultures about 4 hours before the transfection. DNA calcium-phosphate precipitate was prepared as follows: the transforming plasmid DNA (2-5 μg) was diluted in 150 μl of solution D. Fifty microlitres of 0.5 M calcium chloride were added to each tube which was mixed by vortexing then left at room temperature for 10 minutes. Tubes were mixed again then placed on wet ice for 10 minutes. The DNA calcium chloride solution was added dropwise to 300 μl solution H at an approximate rate of 1 drop per second, whilst air was continuously passed through the liquid to aid mixing. The tubes were vortexed immediately for 5 seconds, left on wet ice for 10 minutes, then incubated at room temperature for at least 30 minutes to allow the precipitate to form. The precipitate was applied directly to the dishes of cells. After 4-6 hours of incubation, the medium was aspirated and the cells were washed several times with prewarmed medium, to effectively remove the precipitate. Fresh medium was then added to the cells and they were returned to the incubator.

2.108 LacZ staining of transfected cells

Twenty four to forty eight hours after transfection, cells were screened for the expression of β-galactosidase according to described methods (Sanes, 1986). Medium was aspirated from the cells which were rinsed with PBS then fixed at 4°C for 5 minutes in a solution of 2% (v/v) formaldehyde, 0.2% (v/v) gluteraldehyde made up in PBS. After fixation, cells were washed with PBS and overlaid with staining solution. In order to prevent dislodging of cells growing as a monolayer, molten agar was added to a final concentration of 1% (w/v), and the reaction mixture was overlaid onto the cell monolayer. Staining proceeded overnight at 37°C.

2.11 ANIMAL METHODS

2.111 Production and recovery of mouse blastocysts

Inbred strain C57Bl/6J was used as the donor mouse strain throughout. Female mice were superovulated to control ovulation and to produce a larger number of embryos for injection. On day one, four to six week old females were injected intra-peritoneally (i.p.) with 5 IU of PMSG (pregnant mare's serum gonadotropin). On day three the same mice were injected with 5 IU of HCG (human chorionic gonadotropin) and mated immediately to stud males. The following day, females were checked for the presence of vaginal plugs. In general, 40 matings were set up for each days worth of blastocyst injections.

On day 3.5 post-coitum (p.c.), female mice were sacrificed by cervical dislocation. Blastocysts were recovered by flushing M2 medium through the two respective uterine horns using a 27-gauge needle and 5 ml syringe. Blastocysts were flushed into a glass dish of M2 medium. After all uteruses had been flushed, embryos were collected and transferred to micro-drop cultures of M2 medium under paraffin oil.

2.112 Microinjection of embryonic stem cells into blastocysts

Injecting and holding pipettes were first pulled to an appropriate diameter prior to further refinement on the microforge. Glass pipettes (outer diameter 1.0 mm, internal diameter 0.75 mm) were pulled using a Sutter P-87 micropipette puller (Sutter Instrument Company, USA) then treated further using a microforge (Nikon, Japan). The injecting pipette was ground to an angle and a small spike was created at the tip on the microforge. Injecting pipettes were siliconised prior to use. Both the injecting and holding pipettes were bent twice in order that they would clear the edge of the injection dish and fit straight into the micromanipulator arms.

On the day of the injection, injection pipettes were back-filled with fluorinert (Sigma Chemicals Ltd) before loading into the micromanipulator arm. The micromanipulator injection arm was filled with paraffin oil and flushed immediately before use to remove any air

bubbles that might have been present. Air bubbles within the system can drastically reduce the control of the injection needle.

Dishes were set up with microdrop cultures of M2 medium (60 µl microdrops) under paraffin oil and blastocysts (10 to 20 at any one time) were carefully introduced at the bottom of a drop. Mouse embryonic stem cell cultures growing in a T75 flask were re-fed on the morning of the injection. Cells were trypsinised and resuspended in a final volume of approximately 10-15 mls ES medium. For injection, cells were thawed on day 1 into flasks containing freshly plated feeder cells. Medium, without selection, was replaced daily until day 4 when cells were trypsinised for the injection. Between 5 and 10 µl of the embryonic stem cell suspension was introduced into the microdrop containing the blastocysts. Each blastocyst was injected with between 10-20 embryonic stem cells. Injected blastocysts were transferred to the top of the microdrop until all the blastocysts within that drop had been injected. After injection, blastocysts were carefully transferred to microdrop cultures of Whitten's medium and incubated in a humidified atmosphere of 10% CO_2 at 37°C for approximately 1 hour.

2.113 Vasectomy, preparation of pseudopregnant females and embryo transfer

Vasectomised males were required to produce pseudopregnant females for uterine transfer of microinjected blastocysts. Two month old males of the strain CD1 were vasectomised and used for between 6 and 8 months. Males were anaesthetised by injection i.p. with 0.015 ml Avertin/gram body weight. Vasectomy was performed by single transverse ventral incision. A short length of the vas deferens was removed to ensure that sperm could no longer pass between the severed ends. The incision in the skin was closed with suture clips and removed after 7 days. Vasectomised males were housed singly and test-mated after 14 days to check that they were indeed sterile.

Pseudopregnant females of mouse strain CD1 or CBA/C57Bl F_1 hybrid were produced by mating them with vasectomised males. Matings were set up such that pseudopregnant recipient females would be 2.5 days pseudopregnant when injected blastocysts were transferred into them. Although the blastocysts were 3.5 days old, a delay of one day is though to

allow the blastocyst time to equilibrate with the new uterine environment prior to implantation.

Pseudopregnant females were anaesthetised with Avertin and fur was shaved from a patch on the back. A small transverse incision was made along the lower spinal cord and the skin was moved around until the incision was over the ovary and fat pad. Another incision was made in the body just over the ovary. By grasping the ovarian fat pad, the ovary, oviduct and a small length of the uterus was pulled out through the body wall and held in place with serafine. A small hole was made in the top of the uterus, near the utero-tubule junction, using a 30-gauge needle. Embryos were collected with a transfer pipette and carefully deposited in the uterus. An average of 9 blastocysts were transferred in turn to each side of the uterus. After transfer, the skin was closed with wound clips.

2.114 Breeding programme of chimaeras

Between 18 and 19 days after embryo transfer, mice were born. It was found that mice that exhibited extensive ES contribution to the coat colour could be easily identified between 3 to 4 days, on the basis of skin pigment. Otherwise, mice were examined at between 7 and 10 days when the fur had started to grow. Chimaeras were identified through the presence of agouti/chinchilla fur colour. At this stage the relative percentage contribution of the ES cells to the coat was estimated.

Embryonic stem cell contribution to the germline was screened for by back-crossing all chimaeric animals against C57Bl/6 mice. Male chimaeras were paired up with 3 C57Bl/6 female mice when they reached 4-6 weeks of age. Female chimaeras were paired with a single C57Bl/6 male when they reached 4-6 weeks of age. The presence of agouti offspring was diagnostic of germline transmission. A mouse that had sired or given birth to five litters without transmitting the ES derived coat colour was considered to be non-germline. Offspring that were agouti were screened for the presence of the designed Muc-1 mutation through the preparation of tail DNA and PCR analysis utilising diagnostic oligonucleotide primers.

2.115 Sub-cutaneous injection of mouse embryonic stem cells into athymic nude mice

In an attempt to analyse the differentiation of specific embryonic stem cell clones carrying a targeted mutation in the Muc-1 gene, cells were injected sub-cutaneously (s.c.) into athymic nude mice. This type of analysis was also utilised to investigate the expression pattern of a Muc-1/LacZ fusion protein that was created at the targeting site.

Mice were anaethetised with Avertin and 1.5-2.0 X 10^6 ES cells were injected s.c. in a volume of 0.1 ml sterile PBS. For independently derived targeted ES clones, cells from the same clone were injected into three separate mice. After 2-3 weeks, mice were sacrificed by CO_2 and tumours were removed.

2.116 Histochemical analysis of nude mouse teratocarcinomas

Nude mouse ES-derived teratocarcinomas were treated in two different ways. Each tumour was removed from the mouse, measured and weighed. Tumours were then cut in two. One half of the tumour was placed into methacarn fixative for one hour then subsequently transferred to 70% (v/v) ethanol. This half would be prepared for immunohistochemistry. The second half was washed first with sterile PBS then fixed for between 30-60 minutes at 4°C in LacZ tissue fixative. After fixation, these tumours were washed twice briefly with an excess of PBS, then incubated overnight at room temperature, in the dark, in LacZ staining solution. After overnight staining, tumours were washed again with PBS then fixed at 4°C in 4% (w/v) paraformaldehyde.

For histochemical analysis, all teratocarcinomas were dehydrated through an ascending series of ethanol, then cleared with xylene before paraffin embedding. At least 5 sections were cut at 5 μ intervals at 3 levels within the tumour. Prior to staining, sections were deparaffinised in xylene and rehydrated through a descending series of ethanol ending in a 10 minute rinse in sterile distilled water.

A) Counterstaining sections for morphology, and analysis of LacZ staining

Sections to be analysed were treated with two different counterstains in order to investigate the morphology and the LacZ staining pattern. For an investigation of the morphology of the respective tumours, sections were stained with Gomori's trichrome. Rehydrated sections were stained in preheated (60°C) Bouin's fixative for 3 minutes then rinsed in distilled water before staining for an additional 3 minutes in a preheated solution of Weigert's iron haematoxylin. After staining, slides were rinsed well with several changes of sterile distilled water then stained in a preheated solution of Gomori's trichrome with light green for 10 minutes. At the completion of staining, the sections were again rinsed well, dehydrated through an ascending series of ethanol, cleared in xylene then mounted in Permount (Fisher Scientific, USA).

As Gomori's trichrome stain is fairly blue, it was decided that for the purposes of the investigation of the LacZ staining pattern, sections pre-stained for LacZ were counterstained with eosin only. Briefly, sections were rehydrated as described above, then stained in Eosin Y solution for 1 minute. Sections were dehydrated, cleared then mounted in Permount as previously described.

B) Immunohistochemical staining for the detection of Muc-1 protein.

In order to detect the presence of endogenous Muc-1 protein, sections were stained with the polyclonal antiserum CT1 which is directed to the last 17 amino acids of the human MUC1 cytoplasmic domain (Pemberton, 1992). Slides were deparaffinised and hydrated as described. Staining procedures were carried out as recommended by the manufacturer (Vector Laboratories, USA). Endogenous peroxidase activity was blocked by incubation of the slides for 30 minutes in a solution of 0.5% (v/v) hydrogen peroxide in 80% (v/v) methanol. Slides were washed with PBS then incubated for 20 minutes in dilute normal goat serum (75 µl stock in 5 ml PBS). Excess serum was blotted from the sections which were then incubated for 30 minutes in the primary antiserum diluted to the appropriate concentration in normal goat serum blocking solution. As a negative control, pre-immune or non-

immune serum was utilised. After incubation with the primary antiserum, sections were washed with PBS and incubated for 30 minutes in biotinylated secondary antibody solution (goat anti-rabbit IgG), prepared according to the manufacturer's instructions. While sections were being incubated with the secondary antibody, ABC reagent was prepared and left to stand at room temperature for 30 minutes. Sections were washed for 10 minutes with PBS and incubated in Vectastain ABC reagent for 30 minutes. The ABC reagent consists of an avidin DH: biotinylated horseradish peroxidase H complex. The complex has at least one free binding site for biotin. Sections treated with the ABC reagent were washed with PBS and immune complexes were visualised by incubation in a solution of 0.02% (v/v) hydrogen peroxide and 0.1% (w/v) diaminobenzidine tetrahydrochloride (DAB) in 0.1 M Tris-HCl pH 7.2. When the desired level of staining had been achieved, sections were washed, dehydrated in 70% (v/v) ethanol and counterstained in a 1% (w/v) solution of methyl green (Mallinkrodt, Germany) in methanol. Slides were washed with absolute alcohol, cleared in xylene and mounted in Permount as previously described.

2.117 Immunohistochemical analysis of spontaneous mouse mammary carcinomas

Formalin fixed paraffin-embedded mouse mammary carcinomas were kindly provided by Dr. Clive Dickson, ICRF, London. Sections were taken at 5 μ intervals and were stained for the presence of mouse Muc-1 utilising a monoclonal antibody, LB2, that recognises epitopes present within the oligosaccharide component of the human mammary MUC1 protein (Moss, 1988). In addition, the monoclonal antibody, HMFG2, which recognises the core protein epitope, DTR, present within the human MUC1 tandem repeat was utilised. Staining procedures were carried out as described, utilising a rabbit-anti-mouse secondary antibody. After immunohistochemistry, slides were counterstained with haematoxylin and treated as described above before mounting in Permount.

CHAPTER THREE:

CLONING THE MOUSE HOMOLOGUE OF THE HUMAN TUMOUR-ASSOCIATED MUC1 GENE.

3.1 Introduction

The human polymorphic epithelial mucin gene, designated MUC1, has been cloned, fully sequenced and characterised. The gene spans an average of between 4.5 and 6.5 kilobase pairs and is made up of seven exons and six introns (Fig 3.1). The majority of the coding capacity of the gene is comprised of variable numbers of a highly GC-rich 60 bp tandem repeat encoding a 20 amino acid repeat motif. This repetitive portion is located entirely within the second exon and gives the gene the characteristics of an expressed variable number tandem repeat (VNTR), or minisatellite locus. As such, this gene is one of the few minisatellite loci that encodes a protein product. The MUC1 gene codes for an integral membrane protein with a large external portion, made up of the repetitive domain, a 31 amino acid membrane-spanning domain and a 69 amino acid cytoplasmic domain.

The human MUC1 gene was cloned through screening cDNA expression libraries utilising antibodies that had been previously characterised as being tumour-specific. An investigation of the tissue distribution of the human MUC1 protein revealed that the protein was expressed by the majority of simple secretory epithelial tissues. In these tissues the protein was observed to be exclusively localised to the apical surface of cells lining the lumen. In addition many human carcinomas appear to overexpress the MUC1 protein product. An investigation of the embryonic profile of expression revealed that in the developing mouse embryo the presence of the mouse Muc-1 protein correlated well with the differentiation status of the respective epithelial organs. Investigation of the expression patterns of both the human and mouse Muc-1 genes and

their protein products has led to the proposal of numerous functions for the Muc-1 molecule, yet its biological role in the normal and tumour tissues in which it is expressed remains uncertain.

This chapter describes the cloning and characteristion of the mouse homologue of the human tumour-associated MUC1 gene. The mouse homologue was sought for two reasons. Firstly, the isolation of a homologue of the human MUC1 gene would provide a starting point for a study of the evolution of the minisatellite characteristics of the Muc-1 gene locus. Secondly, the isolation and characterisation of the mouse Muc-1 gene would permit an investigation of the function of the mouse Muc-1 gene through gene targeting experiments in mouse embryonic stem cells. These experiments will be discussed in turn in the following chapters.

3.2 DNA probes

Human MUC1 DNA probes, for example pMUC7 (Gendler, 1987) from within the variable number tandem repeat (VNTR) region, were observed to cross-hybridise only with genomic DNA from human and primate samples (Pemberton, 1992). It was reasoned that if the membrane-spanning and cytoplasmic domains of the MUC1 protein play a functional role in the cells in which the MUC1 protein is expressed, then this functional importance would be reflected in sequence conservation. For this reason, the human MUC1 cDNA clone, pGEM-PEM16, designated 16.2, which contains sequence encoding the membrane-spanning and cytoplasmic domains (Gendler, 1990a) was used as a probe against mouse genomic DNA digests. Genomic DNA was prepared from two mouse mammary carcinoma cell lines, C57MG (Vaidya, 1978) and HC11 (Ball, 1988) and also from the human mammary carcinoma cell line T47D (Keydar, 1979), to act as a positive control for the probe. Approximately 10-15 μg of genomic DNA from the respective cell lines were digested separately with the restriction endonucleases EcoRI, HinfI and BamHI and size-fractionated through a 0.7% (w/v) agarose gel, alongside λHindIII molecular weight markers. The DNA samples were transferred onto nylon membrane and hybridised with radiolabelled probe 16.2 overnight in hybridisation mix with 43% formamide (as opposed to the standard 50% (v/v) formamide hybridisation mix recipe). It was found that even at high stringency a positive signal was obtained

on mouse genomic DNA samples that corresponded to DNA restriction endonuclease fragments migrating at approximately 11 kilobase pairs (kbps) (EcoRI) and 15 kilobase pairs (BamHI) (Fig 3.2). The mouse HinfI digests failed to demonstrate any significant hybridising fragments. This was presumed to be a reflection of the presence of HinfI restriction endonuclease sites in the three-prime end of the mouse Muc-1 gene that would reduce this region of the mouse Muc-1 gene to numerous small fragments. The human DNA samples, acting as a positive control, displayed the characteristic doublet for all three endonucleases; the result of two alleles with different numbers of the 60 bp tandem repeat per allele.

3.3 Screening λgt10 library: cDNA cloning

Cells of the lactating mammary gland express the MUC1 protein at one of the highest levels observed in humans (Patton, 1986; Zotter, 1988) and this protein can be detected in the milk of all mammalian species thus far characterised (Patton, 1986; Patton, 1989; Patton, 1990; Welsch, 1990; Spicer, 1991; Campana, 1992; Patton, 1992). For this reason, a λgt10 cDNA library was constructed from mRNA prepared from Balb/c mouse lactating mammary gland tissue (Stubbs, 1990). 6 X 10^6 plaque forming units (pfus) of the amplified library were plated out, transferred to nylon membranes and hybridised with the human probe 16.2, under the same conditions previously used for the Southern blot. Twenty-two positive plaques were identified and taken through three successive rounds of plaque purification to arrive at pure clones. λ DNA was prepared from each clone, digested with the restriction endonuclease EcoRI, and fractionated through a 1% (w/v) agarose gel. All twenty-two clones appeared to contain the same insert, characterised by DNA fragment sizes of 1.2 and 0.8 kilobase pairs (presumably because the library had been amplified). A Southern blot of the EcoRI digested λ DNA preparations, utilising the probe 16.2, revealed that only the band migrating at 1.2 kilobase pairs hybridised to the probe.

Both insert fragments and the full-length insert of 2.0 kilobase pairs were purified from agarose gels by DEAE ion-exchange paper and sub-cloned into the EcoRI site of the vector, pBluescriptSKII+, for further characterisation and DNA sequencing. The respective fragments, when cloned into pBluescript, were designated as pMuc10-I (2.0 kbp), II (1.2 kbp)

and III (0.8 kbp). Double-stranded plasmid DNA was in all cases sequenced by the di-deoxy chain termination method (Sanger, 1977). Alignment of DNA sequences with the human MUC1 sequence was performed using the Intelligenetics SEQ-ALIGN programs on a VAX computer. The sequence for pMuc10-II aligned with sequence of the human MUC1 gene immediately downstream of the tandem repeat domain (Fig 3.31) at its 5' end and contained a polyA tail and poly-adenylation sequence at its 3' end. Sequence obtained from the 800 base pair fragment, pMuc10-III, failed to align with known human MUC1 sequence. However, a search of the GenBank Sequence database, utilising the Intelligenetics FastDB program, indicated a close to 100% match with the published sequence for the gene encoding mouse ε-casein (Hennighausen, 1982). Presumably, this transcript is so abundant in mouse lactating mammary gland RNA that the resultant reverse transcribed cDNA was a particularly common species when the library was made, and thus was ligated non-specifically to the end of the Muc-1 cDNA clone prior to ligation into the λgt10 vector. As a result of these findings, the clone pMuc10-II was re-designated as simply pMuc10.

All twenty-two of the initial cDNA clones obtained were identical, with the casein insert attached at one end, and were, therefore, the result of specific amplification of one primary clone. In an effort to obtain clones containing sequence further 5' of that present in pMuc10, the primary library was screened with the pMuc10 insert. The total primary library of approximately 1 X 10^5 pfus was plated out at a lower density than that used previously, in order that plaques would be discrete, and then screened with the radiolabelled probe pMuc10. Five positive plaques were identified and half of each plaque was picked into standard SM buffer for re-plating and further plaque purification, whereas the other half was picked into 250 μl of a solution of 2X PCR buffer of composition 20mM Tris-HCl pH 8.7, 100mM KCl and 3mM $MgCl_2$. Twenty-five microlitre aliquots of each clone, in 2 x PCR buffer, were amplified by the polymerase chain reaction (PCR) utilising synthetic λgt10 oligonucleotides directed to sequence flanking the EcoRI cloning-site. These oligos were used in combination with an antisense oligonucleotide directed to sequence specific to the Muc-1 gene, 5'-CCA AGC TTG ACT AGA CTG GTA GCT GAG CC-3' (as obtained from the clone pMuc10), containing a site for the restriction endonuclease HindIII. It was reasoned that such an arrangement would result in the specific

amplification of a product only if the clone in question contained sequence 5' of the Muc-1 oligo utilised (Fig 3.32). A fragment was specifically amplified from two of the five positive plaques; the largest fragment amplified being approximately 500 base pairs in size. This fragment was sub-cloned into pBluescript KSII+ and designated pMuc2TR (Fig 3.31 and 3.32) where TR denotes tandem repeat. The entire insert, of approximately 1.7 kilobase pairs, from the original phage clone was also sub-cloned into pBluescriptKSII+ and designated pMuc2. Sequencing revealed that the clone pMuc2TR was made up entirely of degenerate tandem repeats 60-63 base pairs in size.

3.4 Screening Balb/c cosmid library

Although the entire primary cDNA library had been screened, none of the Muc-1 cDNA clones isolated represented the full-length transcript. This absence of full-length cDNA clones may have been a reflection of the difficulty of reverse transcribing extended stretches of repetitive GC-rich mRNA. In an effort to obtain sequence of the 5' end of the mouse Muc-1 gene, the cDNA clone pMuc2TR was utilised as a probe to screen a Balb/c genomic cosmid library. This library was constructed from a partial Sau3A digest of mouse liver genomic DNA cloned into the BamHI site of cos203, an Epstein Barr virus (EBV) based shuttle vector (Kioussis, 1987). Replica filters of this library, constructed by Dr. Dimitris Kioussis (National Institute for Medical Research, London, UK), were kindly provided by Dr. Alastair Lammie (Imperial Cancer Research Fund Laboratories, London). The mouse cosmid library was screened with the Muc-1 cDNA clone, pMuc2TR, according to the method of Church, 1984. Positive colonies (e.g. Fig 3.4) were subjected to three successive rounds of colony-purification. Purified clones were referred to as cosmo (cosmid-mouse) 1.21, 1.22, 1.23 and 2.21, 2.22 and 2.23, respectively. Cosmid DNA was digested with various restriction enzymes and, subsequently, fragments of approximately 15 kilobase pairs (BamHI) and approximately 11 kilobase pairs (EcoRI) containing the mouse Muc-1 gene (see Fig 3.2 and 3.4) were sub-cloned into the vector pBluescript. These plasmids were designated pMucBam and pMucEco, respectively. The plasmid pMucEco was sub-cloned further through PstI and TaqI digests. PstI and TaqI restriction fragments containing all or part of the repetitive domain were cloned into pBluescriptKSII+ for sequencing.

3.5 5' cDNA cloning by RT-PCR

Sequence of the putative translation start site was obtained from the mouse Muc-1 genomic clones using oligonucleotides previously synthesised according to sequence of the human MUC1 gene. This putative sequence was used to synthesise a specific mouse Muc-1 oligonucleotide of the following sequence 5'-CCC GAA TTC ATG ACC CCG GGC ATT CGG GCT-3' (corresponding to nucleotides +69 to +89 in Fig 3.61) containing a site for the restriction endonuclease EcoRI at its 5' end. Total RNA was isolated from C57Bl/ICRF one day post-partum mouse lactating mammary tissue and was reverse transcribed with a mouse Muc-1 antisense oligonucleotide directed to the furthest 5' sequence obtained from the mouse Muc-1 cDNA clone pMuc2TR. This oligonucleotide was directed to part of repeat number seven and had the sequence, 5'-CCC AAG CTT GTC TGG AGA GCT GGT GGA GTC-3' (corresponding to nucleotides +1314 to +1294 on the antisense strand in Fig 3.61) and contained a site for the restriction endonuclease HindIII at its 5' end. The product of the first strand cDNA synthesis reaction was specifically amplified by the polymerase chain reaction (PCR) utilising the two oligonucleotides described above. An amplified fragment migrating at approximately 300 base pairs was obtained. This product was digested with EcoRI and HindIII and ligated into pBluescriptKSII+. Upon bacterial transformation and plasmid preparation it was observed that, rather than there being clones containing inserts exclusively of 300 base pairs in size, the insert size ranged from between approximately 200 base pairs up to approximately 400 base pairs. Subsequent sequencing of these variant clones indicated that the insert size variation was the result of differing numbers of repeat units. Clones contained either one, two or four repeats, respectively, with the most common clone being one containing only a single repeat. Although the 3' anti-sense oligonucleotide utilised was directed to repeat number seven, it had preferentially annealed and extended off a number of other repeats, notably repeats number two, three and five.

This type of phenomenon is often observed when attempting to amplify repetitive sequence. Many times the resultant amplified product is much smaller than that expected. This is thought to be due to the slippage of repeat units relative to one another, subsequent mis-priming of the polymerase, and eventual selection for products with the smallest

number of repeat units (Jeffreys, 1988b). Presumably, these smaller fragments with a lower number of repeat units are inherently more stable and are synthesised more rapidly by the DNA polymerase eventually competing out the larger full-length fragments.

When working with genes possessing repetitive domains, the determination of the total number of repeat units that may be present is not trivial. Sequencing oligonucleotides and PCR oligonucleotides, if made to any part of the repetitive sequence, may efficiently anneal at a number of different positions along the repeat array. In order to elucidate the correct number of repeat units present in the mouse Muc-1 gene, it was necessary to sub-clone various portions of the repetitive domain from the Muc-1 genomic clone pMucEco. To this end, TaqI, PstI and SacI sub-fragments of the repeat domain were cloned and their sequence was determined. In this way it was determined that the total number of repeat units was sixteen (Fig 3.32 and 3.61). All sequence obtained from PCR derived clones was confirmed in each case through sequencing three independent clones and also through comparison with the sequence from the genomic clones, which were obtained through standard 'non-PCR' sub-cloning.

To confirm that all three classes of cDNA clones, 3' (pMuc10), tandem repeat (pMuc2TR), and 5' (pMuc5') were indeed part of the same transcript, northern blots of mouse lactating mammary gland total RNA were hybridised with each clone, respectively. Each clone was observed to hybridise to a single transcript migrating at approximately 2.2 kilobase pairs (Fig 3.32). This confirmed that the 2.2 kilobase pairs of cDNA sequence that had been obtained represented the full-length cDNA sequence of the mouse Muc-1 gene.

3.6 DNA and protein sequence analysis of the mouse Muc-1 gene

A combination of cDNA and genomic clones led to a determination of the full-length sequence for the mouse Muc-1 gene. Sequence was determined from both DNA strands from approximately 600 base-pairs upstream of the mouse Muc-1 translation start site to just downstream of the Muc-1 poly-adenylation signal (Fig 3.61). Comparison of the cDNA sequence with genomic sequence allowed the exon-intron structure of the gene to be elucidated. The mouse Muc-1 gene, like its

Upon close inspection of the sequence at the end of the mouse Muc-1 first intron it became apparent that there were several in-frame stop codons within the last 200 base pairs.

human counterpart, was deduced to be made up of seven exons and six introns (Fig 3.62). Unlike the human gene, however, the mouse Muc-1 first intron was found to be over 200 base pairs longer. Sequence alignment indicated that this extra sequence corresponds to sequence that makes up part of the beginning of exon 2 in the human gene. Accordingly, the human MUC1 predicted core protein sequence includes 67 amino acids between the signal peptide and repetitive domain that share no homology with the mouse Muc-1 core protein.

A dot-matrix plot, comparing the human and mouse Muc-1 genomic sequences, illustrated the overall high similarity of the two genes (Fig. 3.63). Using this type of analysis, a straight line through the diagonal is expected if the two genes share perfect homology. Gaps in the main diagonal, corresponding with areas with little or no significant homology, were observed to line up with the intronic sequences of the two genes. Lines parallel to the main diagonal forming a 'box' were also observed. This type of result is indicative of the presence of tandemly repeated elements.

The deduced amino acid sequence of the mouse mucin gene (Fig. 3.61 and 3.64) encodes an integral membrane protein with 44% of its coding capacity made up of serine, threonine and proline. The protein appears to consist of four distinct regions: (a) an amino-terminal region containing a hydrophobic signal sequence preceeding a short stretch of unique sequence; (b) a tandem repeat region encoding sixteen degenerate repeats (underlined), five of which are 21 amino acids in length, the remaining 11 repeats being 20 amino acids long; (c) a carboxy-terminal region containing unique sequence followed by a hydrophobic membrane spaning domain of 31 amino acids; and (d) a carboxy-terminal cytoplasmic tail of 69 amino acids. According to the predicative method of von Heijne, 1986, the signal peptide is 11 amino acids long and is cleaved between the glycine and phenylalanine residues.

The sequence of the 20-21 amino acid tandem repeat domain corresponds to what might be expected for a protein that is extensively O-glycosylated. On average there are 9 serine/threonine residues per repeat with eight of these being found as doublets. The predicted molecular mass for this core protein is 65 kDa (kilodaltons), yet mouse milk fat globule proteins when run on SDS-PAGE gels indicated that the

glycosylated protein present in mouse milk is at least 200 kDa in size (Fig. 4.51). This would imply that as much as 75% of the molecular weight of the fully glycosylated protein is made up of carbohydrate. In addition to there being multiple potential O-linked glycosylation sites, the mouse Muc-1 core protein was found to contain 10 possible N-linked glycosylation sites (Asn-X-Ser/Thr) in the extracellular domain, five of which were identified within the last five repeats.

Alignment of the amino acid sequence of the human and mouse mucin genes revealed the most significant homology to be centred around the membrane-spanning and cytoplasmic tail domains, 90% and 87%, respectively. Other significant areas of homology noted were the amino-terminal signal sequence, the serines and threonines (potential attachment sites for O-linked carbohydrate) prolines and histidines within the repetitive domain, and the potential N-linked glycosylation sites. The most significant difference between the two sequences occurred at the amino-terminal end where 67 amino acids of the human MUC1 protein were found to share no homologous sequence with the mouse Muc-1 protein. As discussed previously, this appears to be due to the incorporation of over 200 extra base pairs of sequence into the first intron of the mouse gene.

Figure 3.64 depicts the alignment between the mouse and human MUC1 predicted core protein sequences, and Figure 3.65 summarises the homology levels, taking into account the various domains of the protein. In order to achieve an optimal alignment, a human MUC1 allele possessing 12 consensus repeats was chosen for comparison with the mouse Muc-1 amino acid sequence. Alignments were carried out on a Vax computer utilising the Genetics Computer Group (GCG) GAP program. Although the overall homology between the two Muc-1 protein sequences was only 53%, a two-dimensional structural prediction indicated that the predicted structures for the two respective proteins was very similar (Fig 3.66).

3.7 Promoter and expression analysis

Several potential hormone responsive elements have been identified within 500 base pairs of the human MUC1 transcription start site, including potential progesterone, glucocorticoid (Lancaster, 1990)

and oestrogen (Tsarfaty, 1990) responsive elements. An alignment of the corresponding sequence obtained for the mouse Muc-1 gene, revealed no significant homology, suggesting that these potential elements probably do not contribute to the regulation of expression of the MUC1 gene. However, overall homology between the human and mouse Muc-1 sequences was high. In particular, the sequence recently identified by Kovarik, 1993, as being responsible for binding factors controlling the tissue-specific expression of the human MUC1 gene, was found to share close to 100% identity with the corresponding sequence of the mouse Muc-1 gene (Fig. 3.71). In addition, the tissue-specific enhancer sequence recently identified by Abe, 1993, was also found to be highly conserved. This sequence has been demonstrated to specifically bind a 45 kDa protein.

High homology between the promoters of the human and mouse mucin genes suggests that expression of the two genes is regulated in a similar manner. The human MUC1 gene was initially identified due to the fact that it was found to be expressed at high levels by human carcinomas, particularly those of the mammary gland and pancreas. In an effort to determine if the mouse Muc-1 gene was also highly expressed in carcinomas, RNA samples isolated from spontaneous mouse mammary tumours were analysed by northern blotting. RNA samples were kindly provided by Dr. Clive Dickson, Imperial Cancer Research Fund, London, UK. Ten micrograms of each RNA sample were size-fractionated through a formaldehyde agarose gel. As a control, an equivalent amount of total RNA isolated from 1 day post partum mouse lactating mammary gland tissue was run alongside the tumour RNAs. Northern blots were screened for the expression of mouse Muc-1 through hybridisation with radiolabelled probe pMuc2TR. Autoradiography revealed that most tumours demonstrated high levels of Muc-1 expression. Indeed, several of the tumours exhibited levels of Muc-1 expression as high as that observed in mouse lactating mammary gland (Fig. 3.72). In an attempt to detect the presence of the mouse Muc-1 protein in sections derived from spontaneously arising mouse mammary carcinomas, sections were stained with a monoclonal antibody, LB2, that is specific for the oligosaccharide component of the human MUC1 mammary mucin (Moss, 1988). This antibody has been demonstrated to cross-react with the oligosaccharide moieties present on the mouse Muc-1 mammary mucin (Parry, 1992). Staining of mouse mammary carcinoma sections resulted

in strong but heterogeneous staining, reminiscent of the staining pattern often observed on sections of human carcinoma stained for the presence of MUC1 protein. Heterogeneity is thought to be partly due to the masking of carbohydrate epitopes by sialic acid and a variation in the extent of sialylation on a cell-to-cell basis (Moss, 1988).

3.8 Conclusions

The mouse Muc-1 gene was isolated and characterised at both the cDNA and genomic levels. The mouse gene, like its human homologue, is made up of seven exons and six introns. The data show that the amino acid composition of the mouse mucin gene is typical of that for a mucin, with serine, threonine and proline making up a large proportion of the total amino acid content. Alignment between the human and mouse sequences revealed the most significant homologies to be centred around the transmembrane and cytoplasmic tail domains. Interaction of this domain with the actin cytoskeleton has been demonstrated (Parry, 1990), and in combination with this data, this would imply an important role for this region in the function of the protein. It is feasible that this region of the protein is involved in trafficking of the protein to the apical surface of the epithelial cell in which the protein is expressed, and may also be binding a so-called link protein/s which in turn may bind to elements of the actin cytoskeleton (Parry, 1990; Spicer, 1991). The sequence of the cytoplasmic tail of the MUC1 protein contains numerous tyrosine residues and a potential site for phosphorylation by protein kinase C. All of these sites were found to be conserved in the mouse Muc-1 sequence. Although it is yet to be shown, it is possible that this portion of the mucin protein is phosphorylated under certain conditions.

Analysis of the mouse Muc-1 sequence revealed the presence of what can be described as a degenerate tandem repeat domain, repeats being 60 or 63 base pairs in size. On average, each repeat contains 9 serine/threonine residues which would allow for up to half the amino acids in the repetitive domain to be O-glycosylated. The predicted size of the mouse Muc-1 core protein is 65 kDa, yet SDS protein gels indicated that the glycosylated protein present in mouse milk is greater than 200 kDa, suggesting that as much as 75% of the molecular mass of the mature mouse milk mucin is made up of carbohydrate. Alignment of the human and mouse Muc-1 protein repetitive domains yielded an overall

homology of only 37%. However, of the amino acids that were conserved between the two species, 75% were either serine, threonine or proline (85% if histidine was also considered). Gendler, 1990a, hypothesised that the tandem repeats of mucin proteins exist as a scaffold for the attachment of carbohydrate. Reinforcement of this hypothesis comes from the observation that it is primarily the potential O-glycosylation sites within the repeats that are conserved between the human and mouse Muc-1 sequences. The fact that antibodies that have epitopes present within the oligosaccharide component of the human MUC1 mammary mucin also cross-react with epitopes present on the mouse Muc-1 mammary mucin (Fig. 3.73 and Parry, 1992) suggests that the profile of O-glycosylation in the mammary gland is similar in these two species. Therefore, although the sequence of the tandem repeat protein core may have substantially diverged, the oligosaccharide side chains have remained similar. This observation provides additional support for the hypothesis that mucin repeat domains function primarily as a scaffold for the attachment of carbohydrate.

The sequence of the human MUC1 tandem repeat portion has recently been demonstrated to form an extended poly-proline β-turn helix (Fontenot, manuscript submitted). Computer-modelling revealed that the amino acid side-chains within each tandem repeat radiate outward and are completely exposed to the solvent. The authors propose that this side-chain orientation facilitates the accessibility of potential O-glycosylation sites to the glycosylation machinery. An investigation of peptides made up of two and three repeat units revealed that in all cases the single histidine residue, present within each repeat, was in exactly the same conformation and local environment. It is interesting to find, therefore, that this residue is also conserved between the two species. A precise repeat structure predicts that each potential O-linked carbohydrate attachment site is equally accessible to elements of the glycosylation machinery.

Although, overall, the two homologues were found to share only 53% identity at the protein level, a two-dimensional plot of the predicted structures of the human and mouse Muc-1 protein cores revealed their similarity (Fig.3.66). It was seen from such a plot that the repetitive portions of both the human and mouse Muc-1 core proteins formed similar extended rod-like structures with numerous β-turns, and that the

sequence C-terminal of the repetitive domain was also very similar in conformation.

In addition to the numerous potential sites for O-linked glycosylation, the mouse Muc-1 protein was also found to possess ten potential sites for N-linked glycosylation. This number can be compared with the five identified within the human MUC1 sequence. Four of these five sites were found to be precisely conserved between the mouse and human sequences, the other one having a potential site in the mouse sequence located within one amino acid residue of its position in the human sequence.

The pattern of expression of the mouse and human MUC1 genes has been shown to be very similar (Zotter, 1988; Pemberton, 1992; Braga, 1992) and, accordingly, the sequence obtained for the mouse Muc-1 promoter displays high homology with the human MUC1 promoter sequence (Fig. 3.71). Of particular interest is the close to 100% conservation of the E-MUC1 sequence identified by Kovarik, 1993, as being involved in controlling the tissue-specific expression of the human MUC1 gene. Indeed, homology of the two sequences in this region led us to postulate, in 1991, that this region of the promoter may play a role in the tissue-specific regulation of expression of the Muc-1 gene (Spicer, 1991).

Also conserved, between human and mouse, is a sequence that has been recently identified by Abe, 1993, which overlaps with a sequence that was found to exhibit high homology to a sequence element identified as being able to bind a transcription factor, MPBF, present in the lactating mammary gland (Watson, 1991). The sequence present within the human MUC1 promoter appears to act as a tissue-specific enhancer and specifically binds a 45 kDa protein present in the human breast adenocarcinoma cell line MCF-7. In the mouse Muc-1 promoter the corresponding sequence displays an 18 out of 21 nucleotide identity with the human enhancer sequence. In addition, a motif present in the mouse Muc-1 gene, overlapping the enhancer sequence identified by Abe, 1993, was found to demonstrate high homology with an element present within the promoter of the mouse whey-acidic protein (WAP) gene which has been shown to be capable of binding to the lactating mammary gland transcription factor MPBF (Fig. 3.71) (Watson, 1991).

The majority of human carcinomas express high levels of the human MUC1 gene. Studies have shown that in normal tissues the pattern of expression of human and mouse Muc-1 is the same but little is known regarding the expression of Muc-1 in mouse carcinomas. In an attempt to determine whether or not the mouse Muc-1 gene, like its human counterpart, is highly expressed by mammary carcinomas, total RNA was obtained from several spontaneously arising mouse mammary carcinomas. Northern blots revealed that the expression level of mouse Muc-1 in mouse mammary tumours was elevated and was close to the level observed in the lactating mammary gland (Fig. 3.72). Immunohistochemical analysis of sections prepared from spontaneously arising mouse mammary tumours also revealed strong heterogeneous staining (Fig. 3.73). The high expression of the Muc-1 gene in mammary carcinomas in both humans and mice suggests that the protein may play a similar role in these tissues in the two species. A strong similarity in the expression profile of the Muc-1 gene in both humans and mice suggests that for an investigation of the biological function of this protein the mouse represents a good model organism.

Figure 3.1 Genomic structure of the human tumour-associated MUC1 gene. The sizes of exons and introns are given in base pairs (bps). The tandem repeat region (striped box) varies in size from 2 to 6 kbps depending upon the number of 60 bp repeats. Hatched area = 39 bp signal sequence; stippled area= 93 bp transmembrane sequence; Arrow indicates the start of transcription.

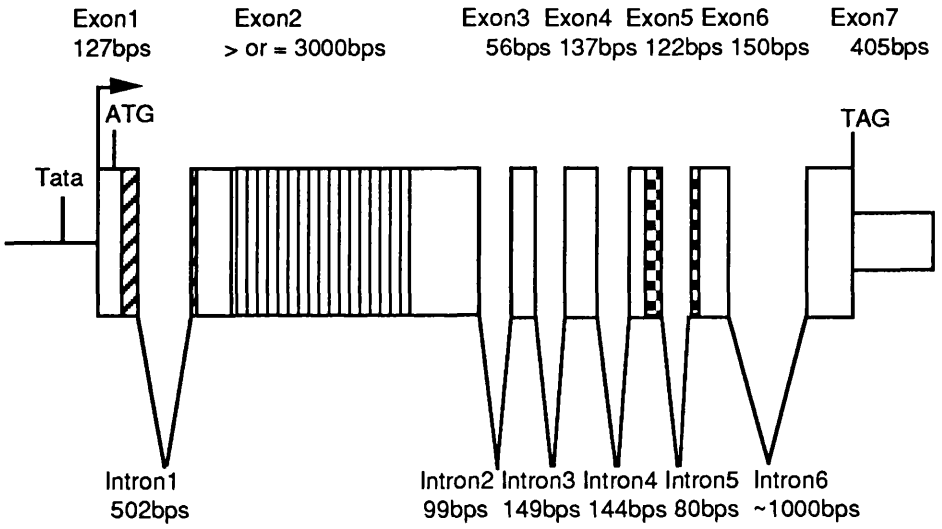

Exon1
127bps
Exon2
> or = 3000bps
Exon3 Exon4 Exon5 Exon6
56bps 137bps 122bps 150bps
Exon7
405bps
ATG
TAG
Tata
Intron1
502bps
Intron2 Intron3 Intron4 Intron5 Intron6
99bps 149bps 144bps 80bps ~1000bps

Figure 3.2 Southern blot of mouse and human genomic DNAs hybridised to the human MUC1 cDNA probe pGEM-PEM16. λ HindIII= radiolabelled HindIII digested bacteriophage λ DNA; E (EcoRI); B (BamHI); H (HinfI). 11 kilobase pair (EcoRI) and 15 kilobase pair (BamHI) fragments were observed to positively hybridise to the human MUC1 cDNA probe.

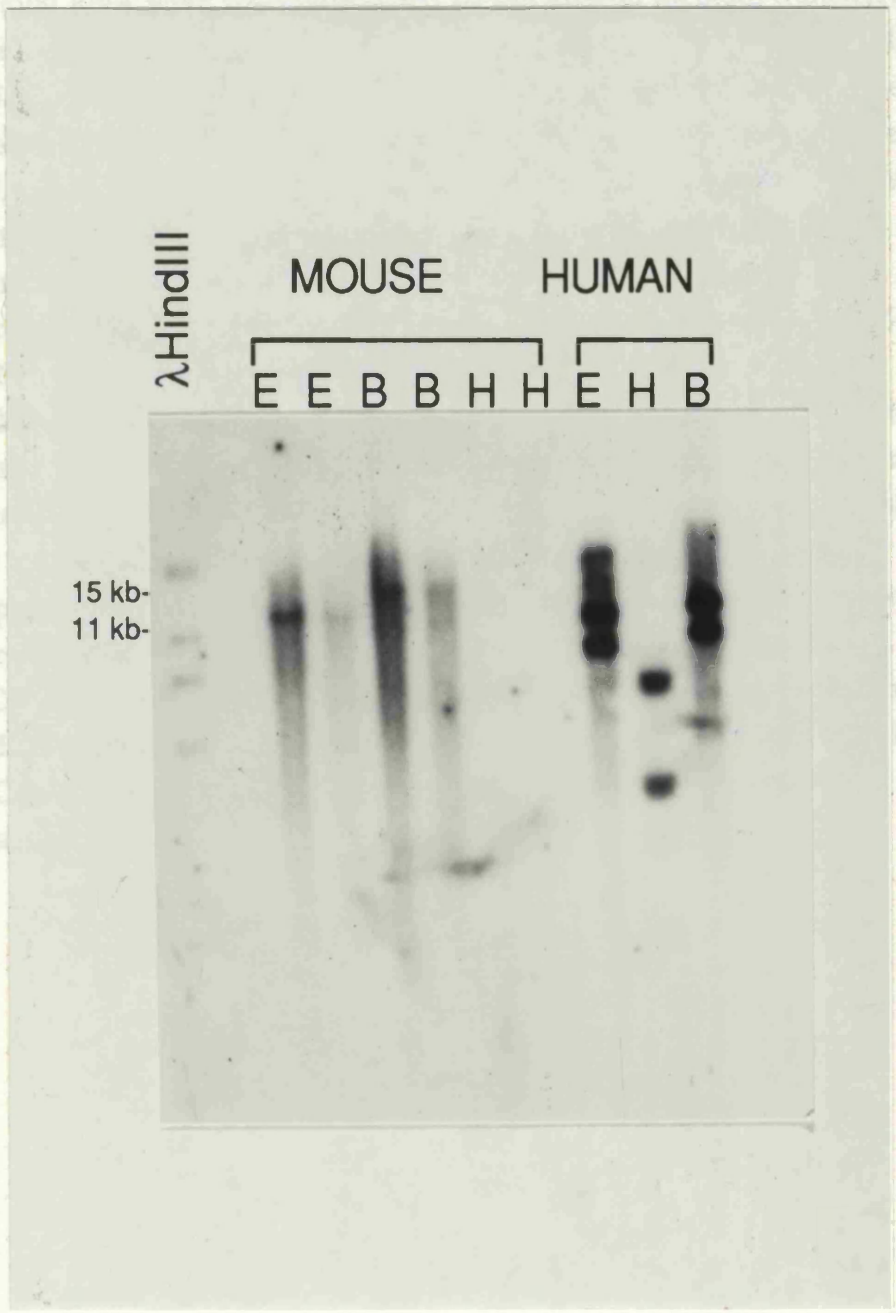

λHindIII
MOUSE
HUMAN
E E B B H H E H B
15 kb-
11 kb-

Figure 3.31 A) Northern blots of mouse lactating mammary gland RNA hybridised to three mouse Muc-1 probes derived from different regions of the transcript. Probe a represents a 5' cDNA clone; probe b is the clone pMuc2TR and probe c represents the clone pMuc10. Approximately 10-15 μg total RNA were size-fractionated by the glyoxal method, transferred onto nylon membrane and hybridised to the respective radiolabelled DNA probe overnight. Lane 1, T47D human mammary carcinoma cell line RNA; lane 2, ICRF-23 human embryonic lung fibroblast cell line RNA; lane 3, C57Bl/ICRF mouse lactating mammary gland RNA. With all three probes a single band was detected in the mouse lactating mammary gland RNA corresponding to a messenger RNA of approximately 2.2 kilobase pairs in size. In addition, the mouse Muc-1 3' cDNA probe, probe c, cross-hybridised at high stringency with RNA isolated from the human mammary carcinoma cell line T47D. **B) Mouse Muc-1 cDNA cloning strategy and restriction map of full-length mouse Muc-1 cDNA.** The sixteen repeats are indicated by empty boxes. All sites for the restriction enzymes, M (MspI); T (TaqI); H (HinfI); P (PstI); K (KpnI); S (SacI); and N (NcoI) are shown. Below this line the horizontal bars indicate the extent of the respective characterised cDNA clones. 'PCR' refers to clones obtained as a PCR product/s, whereas 'subcloned' indicates clones obtained through conventional methods. pMuc5'-clones (PCR) designates three 5'-clones differing in number of repeats, as discussed in text.

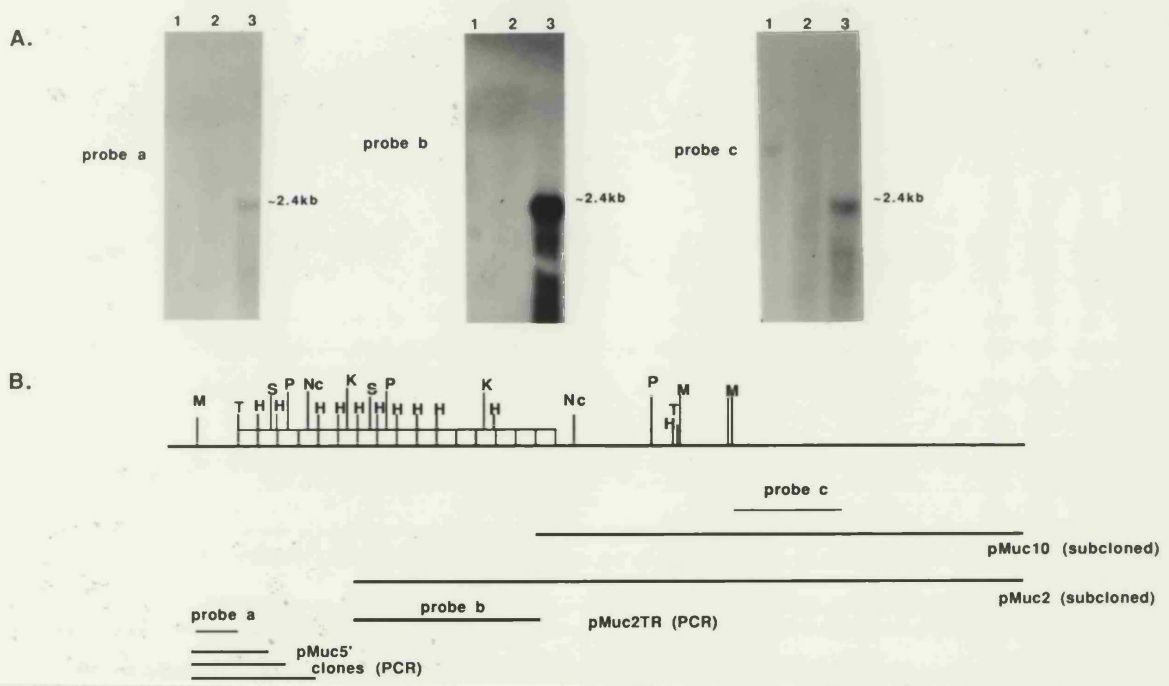
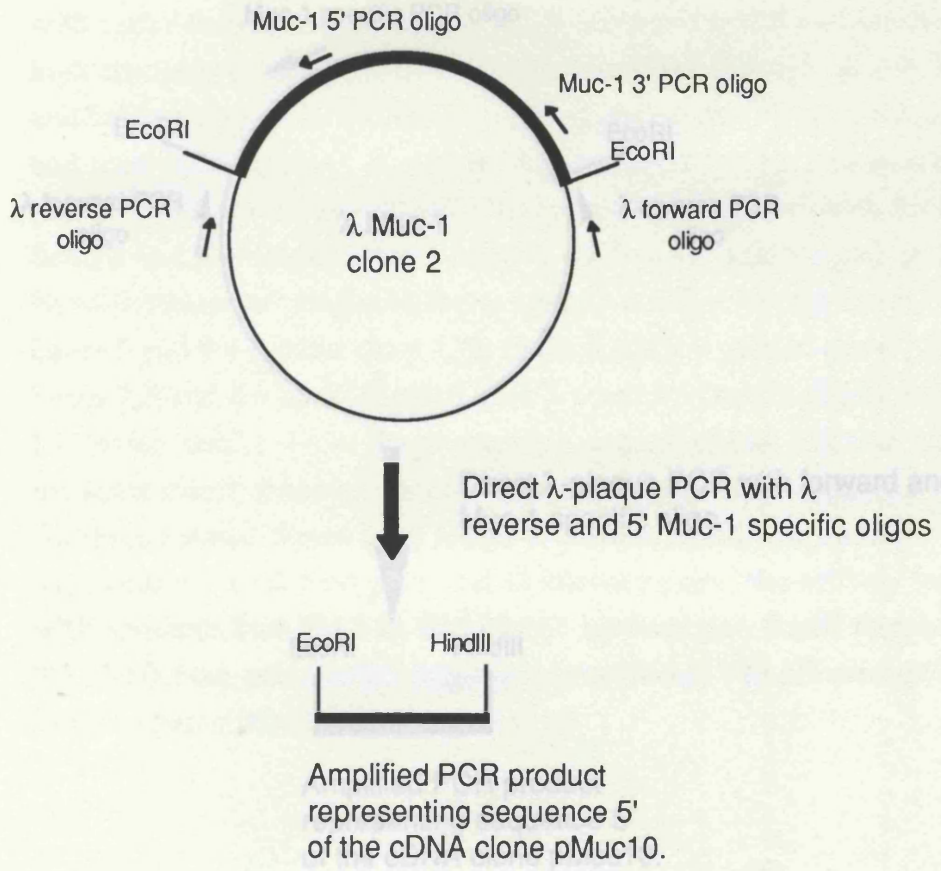

A.
1 2 3
probe a
~2.4kb
1 2 3
probe b
~2.4kb
1 2 3
probe c
~2.4kb
B.
M
T H S H P H Nc H K H H S H P H H H H K H
Nc
P T H M M
probe c
pMuc10 (subcloned)
pMuc2 (subcloned)
probe a
probe b
pMuc2TR (PCR)
pMuc5'
clones (PCR)

Figure 3.32 Direct λ-plaque PCR assay to screen for positive clones containing sequence 5' of that in the mouse Muc-1 3' cDNA clone, pMuc10. Below the diagram, the photograph of an agarose gel depicts the specifically amplified DNA products from one λ Muc-1 clone. Lane 1, λHindIII molecular weight markers; Lane 2, blank. The following combinations of PCR oligos were utilised: Lane 3, λ forward + λ reverse (to amplify the entire insert); Lane 4, λ forward + 5' Muc-1 oligo (see text); Lane 5, λ reverse oligo + 5' Muc-1 oligo; Lane 6, λ forward oligo + 3' Muc-1 oligo; Lane 7, λ reverse oligo + 3' Muc-1 oligo. The amplified product shown in lane 5 was subsequently cloned into pBluescript and designated pMuc2TR. Approximately one-fifth of the total PCR reaction was loaded in each case.


Muc-1 5' PCR oligo
Muc-1 3' PCR oligo
EcoRI
EcoRI
λ reverse PCR oligo
λ Muc-1 clone 2
λ forward PCR oligo
Direct λ-plaque PCR with λ reverse and 5' Muc-1 specific oligos
EcoRI
HindIII
Amplified PCR product representing sequence 5' of the cDNA clone pMuc10.

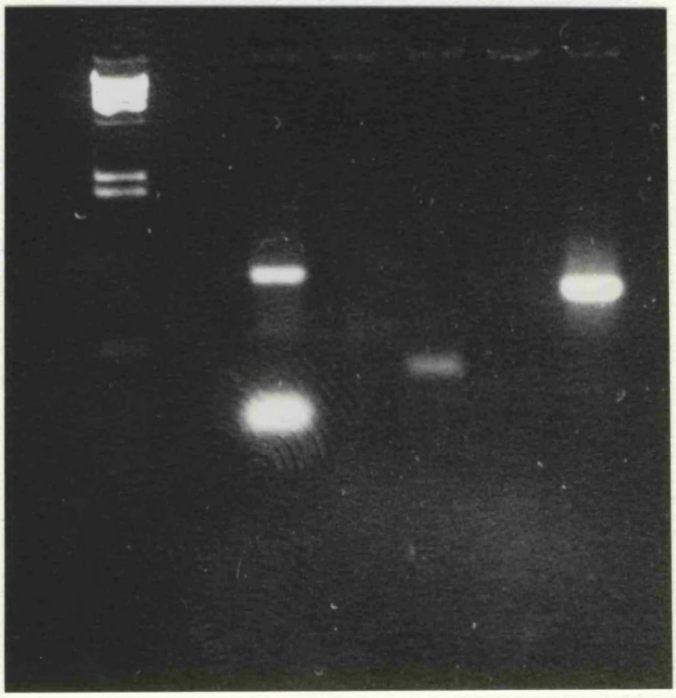


Figure 3.4 A) Autoradiogram displaying representative positive mouse genomic cosmid clones 1.21, 1.22 and 1.23. Membranes were hybridised with radiolabelled mouse Muc-1 cDNA clone pMuc2TR and washed to high stringency. Positive colonies were identified through double lifts and subjected to three rounds of colony purification. **B) Restriction endonuclease digestion of mouse Muc-1 cosmid DNAs.** Five micrograms cosmid DNA were digested with the restriction endonucleases EcoRI and BamHI and size-fractionated through a 0.8% (w/v) agarose gel. Lane 1= λ HindIII molecular weight markers; Lanes 2 and 3 = cosmid clone 1.21; Lanes 5 and 6 = cosmid clone 1.22; Lanes 8 and 9 = cosmid clone 1.23; Lanes 2, 5 and 8 = EcoRI digest; Lanes 3, 6 and 9 = BamHI digest; Lane 10 = 1 kilobase ladder. From the photograph, cosmid clones 1.22 and 1.23 had the same insert, whereas cosmid clone 1.21 appeared to contain a variant. All three cosmid clones were found to possess EcoRI and BamHI fragments of 11 kilobase pairs and 15 kilobase pairs, respectively (compare with Southern blot, Fig 3.2). Both the 11 kilobase pair EcoRI fragment and the 15 kilobase pair BamHI fragment were cloned into pBluescript for further characterisation and sequencing.

A.

B.

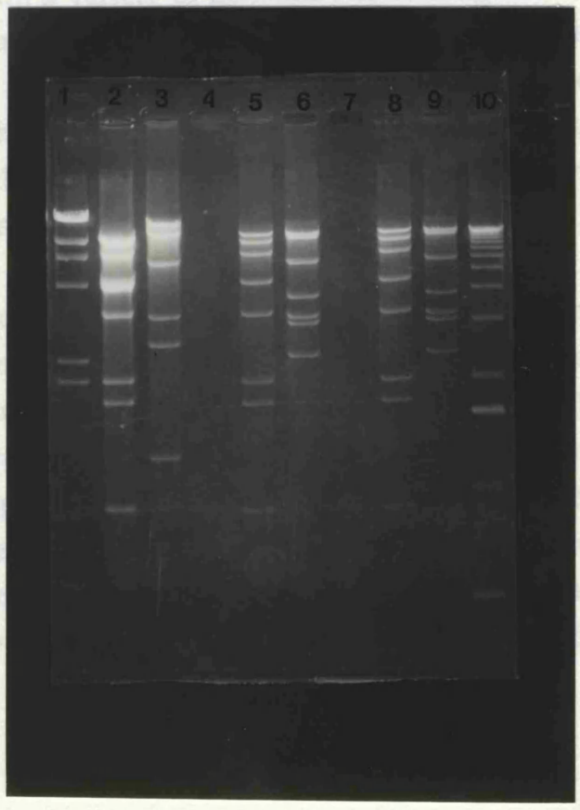


Figure 3.61 Complete nucleotide and predicted protein sequence of the mouse mucin gene Muc-1. Sequence was obtained from both cDNA and genomic clones as follows: The promoter sequence was elucidated from the Muc-1 genomic sub-clone pMucEco through the use of 17 base pair synthetic oligonucleotides directed to sequence on each strand. Throughout, sequencing oligonucleotides were designed such that the sequence determined from them was overlapping. Where ambiguities did arise, due to for instance compression at GC-rich areas, the sequencing reactions were repeated substituting the dGTP labelling mixture and the dGTP termination mixtures with dITP mixtures, as recommended by the manufacturer. The sequence of the Muc-1 cDNA was determined through the use of synthetic oligonucleotides directed both to vector and Muc-1 specific sequences. In order to determine the number and order of tandem repeats within the mouse Muc-1 gene, cDNA clones (pMuc5' and pMuc2TR) were sequenced using synthetic oligonucleotides. In addition, the combination of this sequence information with a variety of genomic sub-cloning and restriction site mapping within and flanking the tandem repeat array allowed the number, order and sequence of the repeats to be unequivocally determined. The remaining sequence was determined from both genomic and cDNA clones from both strands utilising synthetic oligonucleotides as previously described.

Exon sequences are indicated by capital letters, introns, 5' and 3' flanking sequences in lower case. Potential Sp-1 binding sites, the enhancer element identified by Abe, 1993 (-494), the E-MUC1 element, the tata box, translation start site, 16 repeats, stop codon and poly-adenylation signal are underlined. Asterisks denote potential N-linked glycosylation sites. Important cloning restriction endonuclease sites are indicated. Numbering was arbitrarily chosen to exclude the approximately 800 base pairs of undefined sequence within intron 6.

```
   -625                        -600                        -575
cctttttaagtggaaatttcccccaccgcctttctcgggagaaagccaaagagagacc
                         -550                        -525
caggctactggaaagtctggcaggggcggggattgtggttttcagagtaccgggaacg
              -500                        -475
gctagaagggcggggctgttccgggaagtggcgggggggggggggttcggggaggaga
     -450                        -425
ctaaaactagcgacctgtccactattatccagcccccttatgtcctctcaacttgtga
                         -375                        -350
gaagtggtgttcaaggacccgcagagagaataggaaaagggatcggggacctagggat
                -325                        -300
ggacttcccatgtcttgctgtccagggcctttctgtgtttcctttcctcttcttggcg
       -275                        -250
gtcactctgggactaaaaccagagccagggggccccaggtttcctccacatacacacc
                         -200                        -175
tggggtagtcaaggagcggttggtaaaagtgggaggccagatggagaaaaaacaggtt
                 -150                        -125
gtcataaggttgagtgattggagcccaagtaccctacacaggaatgactcctgggggg
        -100                        -75  E-MUC1
aggaggaggaaaagatagggggagggggcggagtttttgttacctgtcacctgctctgac
                         -25                        +1
tgtgcctagggcggggagggggactggtataaagcagcaggcaccagccctgttCCACC
                 +25                        +50
TCACACACGGAGCGCCAGCCTTGAGTTTGTTTTCTAGCCCCTTCCCGCCTGTTCACCA
         +75 SmaI                +100
CCACCATGACCCCGGGCATTCGGGCTCCTTTCTTCCTGCTGCTACTTCTAGCAAGTCT
      M  T  P  G  I  R  A  P  F  F  L  L  L  L  L  A  S  L
+125                        +150                        +175
AAAAGgtgagaggcgcaaggtggggaggggctgcgctgttcaggtgggactcccagtc
  K
                 +200                        +225
tttctgtggagtttgcttactggtagattatgagagtcataaatggccatccatacag
         +250                        +275
gtatggcccagtatgtgccagggggaaaagaggctaaggaaggaggctgagaagagg
  +300                        +325                        +350
tgacccctcaggagagtgggcactatagtaggcaaataatactctttgtggggagtag
                     +375                        +400
agaagagaggctgtgtcaggagaatgagctccggagtggaagaatggggccgtgtgca
                +425                        +450
gtggttccagtctatagttcccacaccctggacgctcaggttcacagagagttcaagg
      +475                        +500
ccagcttaggctacagtatgaatacttgtctcaagtagaggaataaatgagccagaag
                         +550                        +575
cagaggtgagacctctgcttagggagggaggctgtcaggatgaaggcaaaccaggggc
                +600                        +625
tacctgatgctaacatgatcccccttccttccccggttactttgttttcagagtttat
       +650                        +675
atttttcttcttgctctgatttttattcatcaacccagtaggtttttttttttttgtt
                         +725                        +750
gttgttgttttgtttttttttcttcaactttggctctagccggaactaaaacatcac
                 +775                        +850
acatgttatccaaagcgacccagtgtccaactctgttgattagtgctgcgctgctgtg
                         +900                TaqI
agcataaaaacagtttacccctaatcttccactctggttcagGTTTTCTTGCCCTTCC
                                             G  F  L  A  L  P
+925                        +950                        +975
AAGTGAGGAAAACAGTGTCACCTCATCTCAGGACACCAGCAGTTCCTTAGCATCGACT
  S  E  E  N  S  V  T  S  S  Q  D  T  S  S  S  L  A  S  T
```

```
                  +1000                        +1025
ACCACTCCAGTCCACAGCAGCAACTCAGACCCAGCCACCAGACCTCCAGGGGACTCCA
 T  T  P  V  H  S  S  N  S  D  P  A  T  R  P  P  G  D  S
         +1050                      +1075
CCAGCTCTCCAGTCCAGAGTAGCACCTCTTCTCCAGCCACCAGAGCTCCTGAAGACTC
T  S  S  P  V  Q  S  S  T  S  S  P  A  T  R  A  P  E  D  S
+1100   PstI              +1125                      +1150
TACCAGTACTGCAGTCCTCAGTGGCACCTCCTCCCCAGCCACCACAGCTCCAGTGAAC
 T  S  T  A  V  L  S  G  T  S  S  P  A  T  T  A  P  V  N
                   +1175                      +1200
TCCGCCAGCTCTCCAGTAGCCCATGGTGACACCTCTTCCCCAGCCACTAGCCCTTTAA
 S  A  S  S  P  V  A  H  G  D  T  S  S  P  A  T  S  P  L
          +1225                      +1250
AAGACTCCAACAGCTCTCCAGTAGTCCACAGTGGCACCTCTTCAGCTGCCACCACAGC
K  D  S  N* S  S  P  V  V  H  S  G  T  S  S  A  A  T  T  A
 +1275                      +1300                      +1325
TCCAGTGGATTCCACCAGCTCTCCAGTAGTCCACGGTGGTACCTCGTCCCCAGCCACC
 P  V  D  S  T  S  S  P  V  V  H  G  G  T  S  S  P  A  T
                   +1350                      +1375
AGCCCTCCAGGGGACTCCACCAGCTCTCCAGACCATAGTAGCACCTCTTCTCCAGCCA
 S  P  P  G  D  S  T  S  S  P  D  H  S  S  T  S  S  P  A
          +1400             PstI    +1425
CCAGAGCTCCCGAAGACTCTACCAGTACTGCAGTCCTCAGTGGCACCTCCTCCCCAGC
T  R  A  P  E  D  S  T  S  T  A  V  L  S  G  T  S  S  P  A
  +1450                      +1475                      +1500
CACCACAGCTCCAGTGGACTCCACCAGCTCTCCAGTAGCCCATGATGACACCTCTTCC
 T  T  A  P  V  D  S  T  S  S  P  V  A  H  D  D  T  S  S
                   +1525                      +1550
CCAGCCACTAGCCTTTCAGAAGACTCCGCCAGCTCTCCAGTAGCCCACGGTGGCACCT
 P  A  T  S  L  S  E  D  S  A  S  S  P  V  A  H  G  G  T
           +1575                      +1600
CTTCTCCAGCCACCAGCCCTCTAAGGGACTCCACCAGTTCTCCAGTTCACAGTAGTGC
S  S  P  A  T  S  P  L  R  D  S  T  S  S  P  V  H  S  S  A
   +1625                      +1650
CTCCATCCAAAACATCAAGACTACATCAGACTTAGCTAGCACTCCAGACCACAATGGC
 S  I  Q  N  I  K  T  T  S  D  L  A  S  T  P  D  H  N* G
                   +1700                      +1725
ACCTCAGTCACAACTACCAGCTCTGCACTGGGCTCAGCCACCAGTCCAGACCACAGTG
 T  S  V  T  T  T  S  S  A  L  G  S  A  T  S  P  D  H  S
              +1750                      +1775
GTACCTCAACTACAACTAACAGCTCTGAATCAGTCTTGGCCACCACTCCAGTTTACAG
G  T  S  T  T  T  N* S  S  E  S  V  L  A  T  T  P  V  Y  S
     +1800                      +1825
TAGCATGCCATTCTCTACTACCAAAGTGACGTCAGGCTCAGCTATCATTCCAGACCAC
 S  M  P  F  S  T  T  K  V  T  S  G  S  A  I  I  P  D  H
                   +1875                      +1900
AATGGCTCCTCGGTGCTACCTACCAGTTCTGTGTTGGGCTCAGCTACCAGTCTAGTCT
 N* G  S  S  V  L  P  T  S  S  V  L  G  S  A  T  S  L  V
              +1925                      +1950
ATAATACCTCTGCAATAGCTACAACTCCAGTCAGCAATGGCACTCAGCCTTCAGTGCC
Y  N* T  S  A  I  A  T  T  P  V  S  N* G  T  Q  P  S  V  P
      +1975                      +2000
AAGTCAATACCCTGTTTCTCCTACCATGGCCACCACCTCCAGCCACAGCACTATTGCC
 S  Q  Y  P  V  S  P  T  M  A  T  T  S  S  H  S  T  I  A
```

```
                          +2050                      +2075
AGCAGCTCTTACTATAGCACAGTACCATTTTCTACCTTCTCCAGTAACAGTTCACCCC
 S  S  S  Y  Y  S  T  V  P  F  S  T  F  S  S  N* S  S  P
                +2100                    +2125
AGTTGTCTGTTGGGGTCTCCTTCTTCTTCTTGTTCTTTTACATTCAAAACCACCCATT
Q  L  S  V  G  V  S  F  F  F  L  F  F  Y  I  Q  N  H  P  F
       +2150                    +2175
TAATTCTTCTCTGGAAGACCCCAGCTCCAACTACTACCAAGAACTGAAGAGGAACATT
  N* S  S  L  E  D  P  S  S  N  Y  Y  Q  E  L  K  R  N* I
                          +2200                    +2225
TCTGGATTGgtgggtatcagcctagcctctgccatgtgtcccctgacgtagctcttca
 S  G  L
                +2250                    +2275      PstI
ggactgcatggctttcacatcactcctgagtcttctcctcttctcccagTTTCTGCAG
                                                    F  L  Q
        +2300                    +2325
ATTTTTAACGGAGATTTTCTGGGGATCTCTAGCATCAAGTTCAGgtacgttctggatt
 I  F  N  G  D  F  L  G  I  S  S  I  K  F  R
+2350                    +2375                    +2400
tgacttgggggaggaatggtcagtctcgtgactttgtggtgtcgggatgggggtgggg
                  +2425                    +2450
tggggagaggagtgctgagctataagctcagtctatctgagctccctatttcctgtga
     TaqI+2475                    +2500
ccagGTCAGGCTCCGTGGTGGTAGAATCGACTGTGGTTTTCCGGGAGGGTACTTTTAG
      S  G  S  V  V  V  E  S  T  V  V  F  R  E  G  T  F  S
 +2525                    +2550                    +2575
TGCCTCTGACGTGAAGTCACAGCTTATACAGCATAAGAAGGAGGCAGATAGCTATAAT
  A  S  D  V  K  S  Q  L  I  Q  H  K  K  E  A  D  S  Y  N*
                  +2600                    +2625
CTGACTATTTCAGAAGTCAAAGgtgaggtgatagccccagctgcagcctggcaccata
 L  T  I  S  E  V  K
          +2650                    +2675
ctatggggctttaccacctgtttacttctggcgccaggagtgggaaatccacctccct
  +2700                    +2725                    +2750
tggggacttccctgaccaccgctttcccttctagTGAATGAGATGCAGTTCCCTCCCT
                                   V  N  E  M  Q  F  P  P
                    +2775                    +2800
CTGCCCAGTCCCGGCCGGGGGTACCAGGCTGGGGCATTGCCCTGCTGGTGCTGGTCTG
S  A  Q  S  R  P  G  V  P  G  W  G  I  A  L  L  V  L  V  C
           +2825                    +2850
TATTTTGGTTGCTTTGGCTATCGTCTATTTCCTTGCCCTGgtaagtctcaagccttct
  I  L  V  A  L  A  I  V  Y  F  L  A  L
   +2875                    +2900
gcggcgcggtgtgcccttggtaaatggaaggggactggccaatccaatctcctgtctc
                    +2950                    +2975
cctagGCAGTGTGCCAGTGCCGCCGAAAGAGCTATGGGCAGCTGGACATCTTTCCAAC
      A  V  C  Q  C  R  R  K  S  Y  G  Q  L  D  I  F  P  T
            +3000                    +3025
CCAGGACACCTACCATCCTATGAGTGAATACCCTACCTACCACACTCACGGACGCTAC
  Q  D  T  Y  H  P  M  S  E  Y  P  T  Y  H  T  H  G  R  Y
    +3050                    +3075
GTGCCCCCTGGCAGTACCAAGCGTAGCCCCTATGAGGAGgtaaagtgtatcccgaaga
 V  P  P  G  S  T  K  R  S  P  Y  E  E
HindIII              +3125                    +3150
agcttgggccatcgacctgggcagggtggggccttct//agcaggtggcgcactggcc
           +3175                    +3200
gaattggaacccttggagagcttgccagcccttgccccaggtgccatcggggtttta
```

```
   +3225                        +3250                        +3275
cctggaagattctttttccctgaaattatttccccttcactcagGTTTCGGCAGGTAA
                                                V  S  A  G  N
                      +3300                     +3325
TGGCAGTAGCAGTCTCTCTTATACCAACCCAGCTGTGGTGACCACTTCTGCCAACTTG
 G  S  S  S  L  S  Y  T  N  P  A  V  V  T  T  S  A  N  L
             +3350                     +3375
TAGGAGCAAGTCACCCCACCCACTTGGGGCAGCTTTGGCGGTCTGCTCCCTCAGTGGT
 .
    +3400                       +3425
CACTGCCAGACCCCTGCACTCTGATCTGGGCTGGTGAGCCAGGACTTCTGGTAGGCTG
                        +3475                       +3500
TTCATGCCCTTTGTCAAGCGCCTCAACTACGTAAGCCTGGTGAAGCCCAGCCCTGCCC
               +3525                       +3550
TGGGGGACACTGGGGCAGTTAGTGGTGGCTCTCAGAAGGACTGGCCTGGAAAACTGGA
     +3575                       +3600
GACAGGGATGGGAACCCAAACATAGCTGAATAAAAGATGGCCTCCTGTTagttatgcc
                        +3650                       +3675
agctgtggtttttcatctgcaactcttgggtcagagcggccaggactcagctatgaag
               +3700
ggcctggaagcggcctttcggggaggactcccgaa
```

Figure 3.62 Genomic structure of the mouse Muc-1 gene. The seven exons and six introns span approximately 4.4 kilobase pairs. The hatched area represents the 33 base pair signal sequence, the striped area the 16 repeats of 60-63 base pairs in length, and the stippled area the 93 base pair transmembrane sequence. At the 5' end the promoter is shown with EMUC1 and TATA box elements (See text).

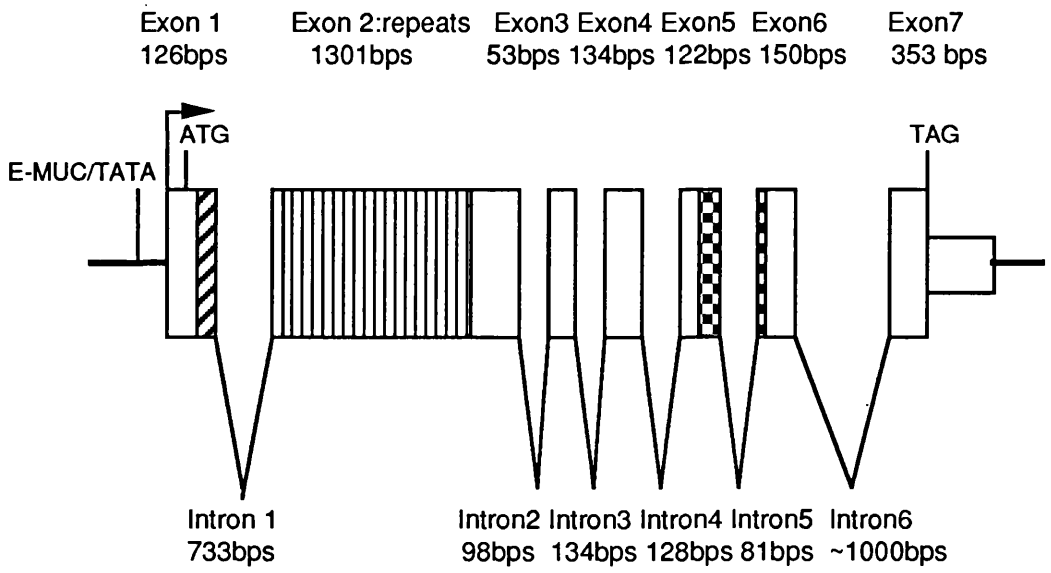

Exon 1
126bps
Exon 2:repeats
1301bps
Exon3
53bps
Exon4
134bps
Exon5
122bps
Exon6
150bps
Exon7
353 bps
ATG
TAG
E-MUC/TATA
Intron 1
733bps
Intron2
98bps
Intron3
134bps
Intron4
128bps
Intron5
81bps
Intron6
~1000bps

Figure 3.63 Dot-matrix plot showing the homology between the human and mouse mucin genomic sequences. A straight line through the diagonal is expected if two sequences are identical. Lines parallel to the main diagonal are indicative of repetitive sequences. Analysis was performed using the Genetics Computer Group (GCG) software and the COMPARE and DOTPLOT programs with a window size of 21 and a stringency of 14. For the purposes of this analysis the last intron of each gene has been deleted. Axes are in base pairs.

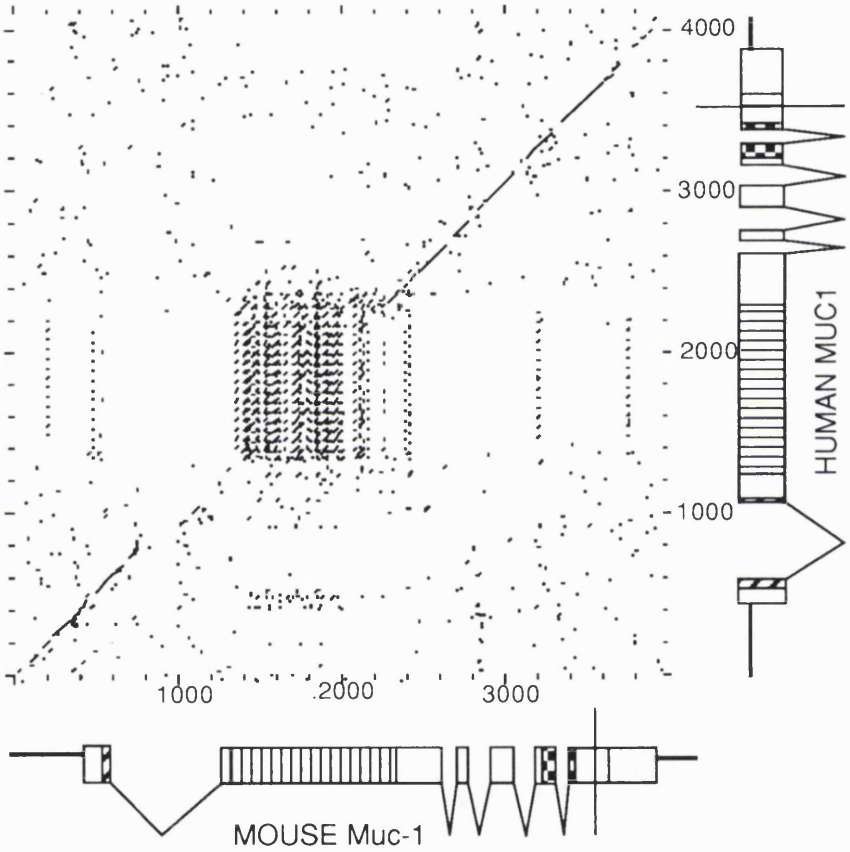
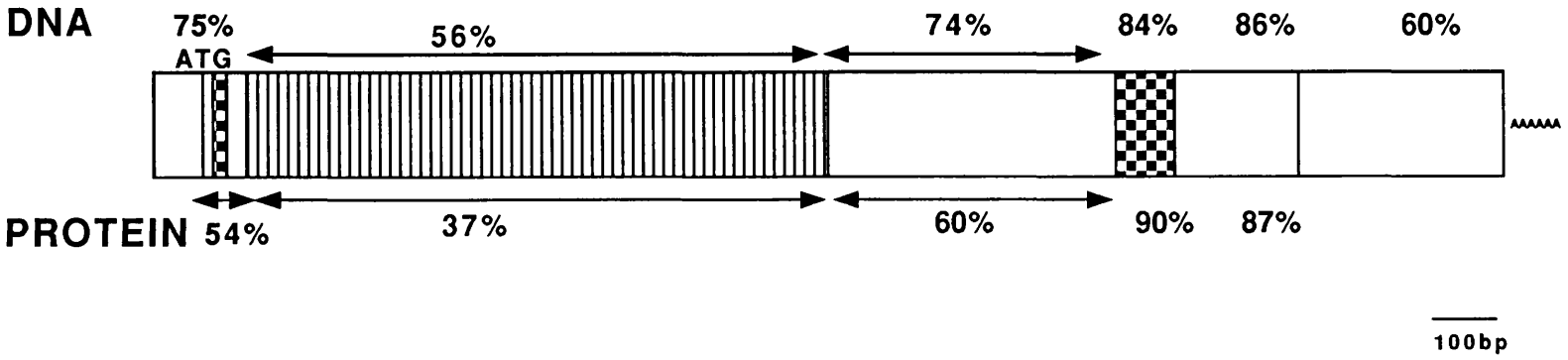

4000
3000
2000
1000
HUMAN MUC1
1000
2000
3000
MOUSE Muc-1

Figure 3.64 A comparison of mouse and human MUC1 amino acid sequence. Conserved amino acids are represented by vertical bars and gaps in the sequence are represented by dots. The signal peptides, potential N-glycosylation sites and the membrane-spanning domain are underlined. Alignment was carried out using the Genetics Computer Group GAP program.

```
MOUSE   1 MTPGIRAPFFLLLLLASLKGF................................ 21
          ||||   ||||||||  |
HUMAN   1 MTPGTQSPFFLLLLLTVLTVVTGSGHASSTPGGEKETSATQRSSVPSSTEKNAV 54

       22 ........................LALPSEENS.........VTSSQDTSSSL 41
                                  ||   |  |         |||   |   |
       55 SMTSSVLSSHSPGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSPVTRPALG 108

       42 ASTTTPVHSSNSDPATRPPGDSTSSPVQSSTSSPATRAPEDSTSTAVLSGTSSP 95
           ||| | |   | |   |   ||  |    || | ||    ||       || |
      109 VSTTPPAHDVTSAPDNKPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAP 162

       96 ATTAPVNSASSPVAHGDTSSPATSPLKDSNSSPVVHSGTSSAATTAPVDSTSSP 149
           |     |   | ||| || | | |   |   |  |  ||   |     ||  |
      163 DTRPAPGSTAPP.AHGVTSAPDTRPAPGSTAPP.AHGVTSAPDTRPAPGSTAPP 214

      150 VVHGGTSSPATSPPGDSTSSPDHSSTSSPATRAPEDSTSTAVLSGTSSPATTAP 203
            || || | | |   ||  | |  || | ||    ||       || | |
      215 .AHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPA 267

      204 VDSTSSPVAHDDTSSPATSLSEDSASSPVAHGGTSSPATSPLRDSTSSPVHSSA 257
            ||  | ||  || | |     |   | ||| || | | |   ||  | |
      268 PGSTAPP.AHGVTSAPDTRPAPGSTAPP.AHGVTSAPDTRPAPGSTAPPAHGVT 319

      258 SIQNIKTTSDLASTPDHNGTSVTTTSSALGSATSPDHSGTSTTTNSSESVLATT 311
          |              | |  ||   |  | ||   | |  ||
      320 SAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAP 373

      312 PVYSSMPFSTTKVTSGSAIIPDHNGSSVLPTSSVLGSATSLVYN.TSAIA.TTP 363
          |               ||   | ||  |     |  |||  || | ||| | |||
      374 PAHGVTSAPDNRPALGSTAPPVHNVTS..ASGSASGSASTLVHNGTSARATTTP 425

      364 VSNGTQPSVPSQYPVSPTMATTSSHSTIASSSYYSTVPFSTFSSNS.SPQLSVG 416
           |  |  | ||     ||     |  | |||   ||||  | |  | ||||| |
      426 ASKSTPFSIPSHHSDTPTTLASHSTKTDASSTHHSTVPPLTSSNHSTSPQLSTG 479

      417 VSFFFLFFYIQNHPFNSSLEDPSSNYYQELKRNISGLFLQIF.NGDFLGISSIK 469
          |||||| | | |  |||||||||  ||||| | ||  ||||   | ||| | ||
      480 VSFFFLSFHISNLQFNSSLEDPSTDYYQELQRDISEMFLQIYKQGGFLGLSNIK 533

      470 FRSGSVVVESTVVFREGTFSASDVKSQLIQHKKEADS.YNLTISEVKVNEMQFP 522
          || |||||  |  |||||    ||  |  | | || | |||||| | |    ||
      534 FRPGSVVVQLTLAFREGTINVHDVETQFNQYKTEAASRYNLTISDVSVSDVPFP 587

      523 PSAQSRPGVPGWGIALLVLVCILVALAIVYFLALAVCQCRRKSYGQLDIFPTQD 576
           ||||  |||||||||||||| ||||||||  |||||||||| ||||||||  |
      588 FSAQSGAGVPGWGIALLVLVCVLVALAIVYLIALAVCQCRRKNYGQLDIFPARD 641

      577 TYHPMSEYPTYHTHGRYVPPGSTKRSPYEEVSAGNGSSSLSYTNPAVVTTSANL 630
          |||||||||||||||||||| || ||||| |||||| ||||||||||  |||||
      642 TYHPMSEYPTYHTHGRYVPPSSTDRSPYEKVSAGNGGSSLSYTNPAVAATSANL 695
```

Figure 3.65 A) Bar diagram to summarise the various levels of homology between the human and mouse Muc-1 mucins at the DNA and protein level for the respective regions; B) Alignment of human and mouse Muc-1 consensus repeats.

A) OVERALL HOMOLOGY



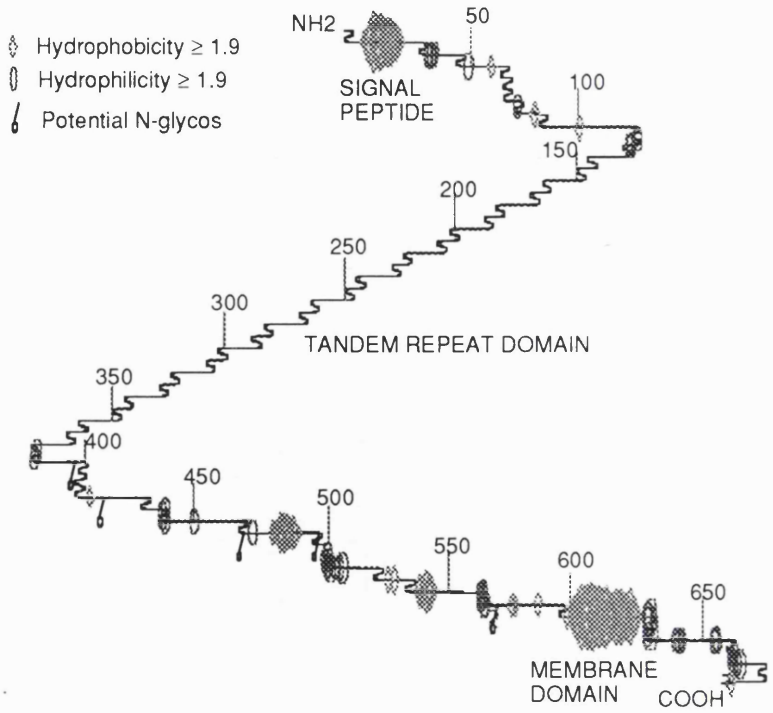
B) REPEAT HOMOLOGY

HUMAN	G	**S**	**T**	A	P	**P**	A	**H**	G	V	**T**	**S**	A	**P**	D	**T**	R	P	A	P
MOUSE	D	**S**	**T**	S	S	**P**	V	**H**	S	G	**T**	**S**	S	**P**	A	**T**	S	A	P	X

(A)

Figure 3.66 Two dimensional structural representation of the human and mouse Muc-1 protein core. The structural prediction was carried out using the Genetics Computer Group (GCG) PEPTIDESTRUCTURE program. Two-dimensional structural prediction was according to the Chou-Fasman algorithim. Areas predicted to be hydrophobic/hydrophilic according to Kyte and Doolittle with a Kd ≥ 1.9 are indicated by diamonds and ovals, respectively. In order to generate core proteins of equivalent size a human MUC1 sequence with 12 consensus repeats was utilised.

HUMAN MUC1 12 REPEAT PROTEIN



MOUSE Muc-1 PROTEIN

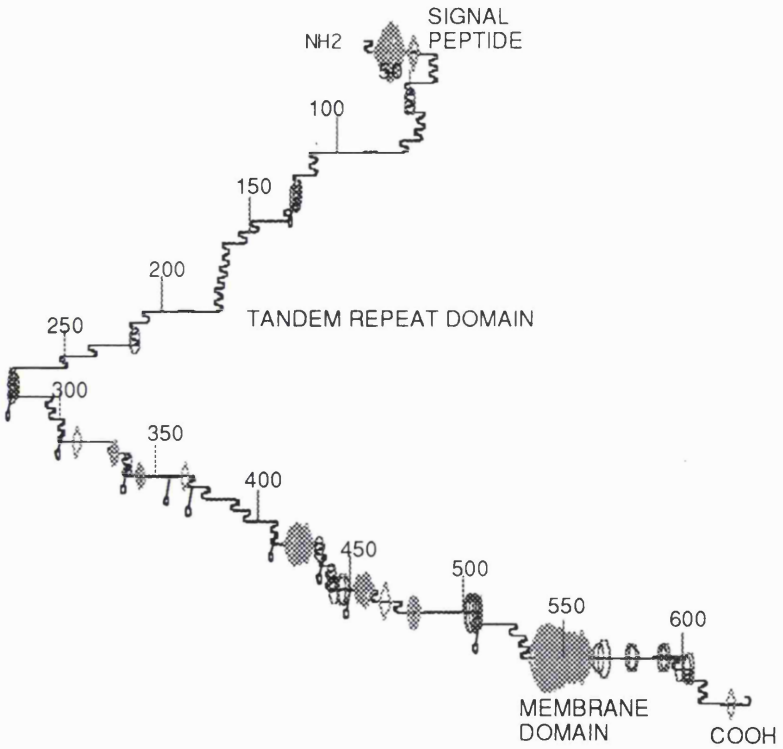


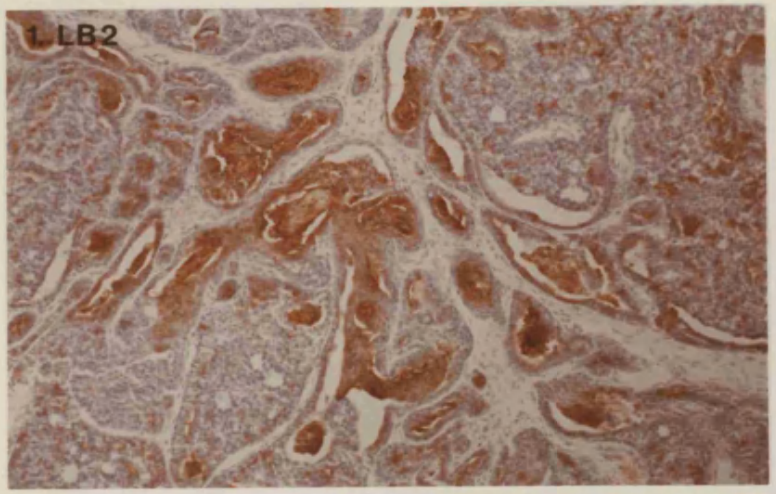
Figure 3.71 Alignment of the promoter sequences of the human and mouse Muc-1 genes. The sequence of the human MUC1 gene from -647 to -15 base pairs is shown. Bases that are conserved between mouse and human are indicated by asterisks. The tata box, Sp1, E-MUC1, potential lactation specific enhancer element, MPBF, sequence and the enhancer element identified by Abe, 1993, are boxed and display a high level of identity. The high overall similarity between the two promoters can be clearly seen.

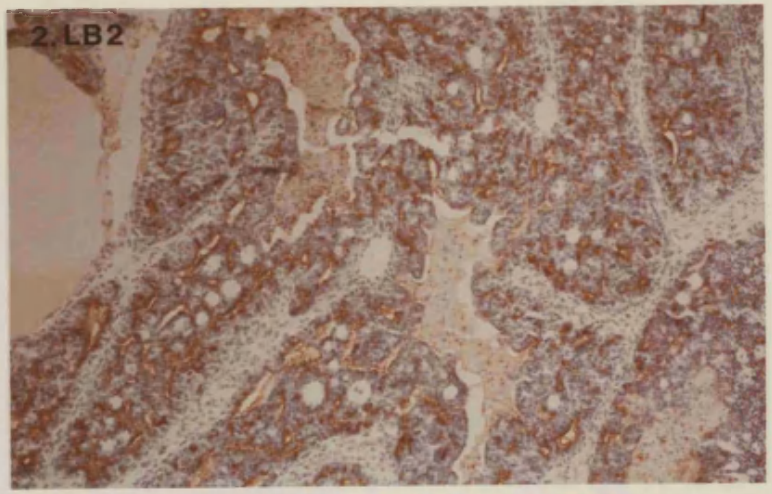
```
        -645
H:    CC--CTTAAGTGGAAATTTCTTCCCCCACTCCTCCTTGGCTTTCTC
M:    **tt***************cc**a**g*ctt**tcc**agaaag*

H:    CAAGGAGGGAACCCAGGCTGCTGGAAAGTCCGGCTGGGGGGGGGAC
M:    ***a***a*-*********a**********t***a****c*****t

                        -550
H:    TGTGGGTTCAGGGGAGAACGGGGTGTGGAACGGGACAGGGAGCGGT
M:    *****t**--------------------c*ga*t**c****a***c

            Sp-1?     MPBF-LIKE    MCF-7 ENHANCER
H:    TAGAAGGGTGGGGCTATTCCGGGAAGTGGTGGGGGGAGGGAG---
M:    ******c*****g*****************c******g***g*gtt

                                                -450
H:    ----------CCCAAAACTAGC-ACCTAGTCCACTCATTATCCAG
M:    cggggaggaga*t*********g****-*******-*********

H:    CCCTCTTATTTCTCGGCCGCTCTGCTTCAG--------------T
M:    ***c*****-----*t**t***aa***gt*agaagtggtgttcaa

H:    GGACCCG--GGGAGGGCGGGGAAGTGGAGTGGGAGACCTAGGGGT
M:    *******ca*a***aata**a**ag***tc***g*********a*

                            -350
H:    GGGCTTCCC-GACCTTGCTGTACAGGACCTCGACCTAGCTGGCTT
M:    **a******atgt********c****g*----------**tt**g

H:    TGTTCCCCATCCCCACGTTAGTTGTTGCCCTGAGGCTAAAACTAG
M:    ****t**tt***t*ttc**g*cg**ca*t***g*a*******c**

                                        -250
H:    AGCCCAGGGGCCCCAAGTTCCAGACTGCCCCTCCCCCCTCCCCCG
M:    ****ag*********g***t*---**c*a*a*a*a*a*-*tggg*

H:    GAGCCAGGGAGTGGTTGGTGAAAGGGGGAGGCCAGCTGGAGAACA
M:    t**t**a****c*******a****t**********a*******a*

                                                -150
H:    AACGGGTAGTCAGGGGGTTGAGCGATTAGAGCCCTTGTACCCTAC
M:    ***a***t****taa*******t****g******aa*********

H:    CCAGGAATGGTCGGGGAGGAGGAGGAAGAGGTAGGAGGTAGGGGA
M:    a********act****g*******---****-*aagaa*******

        Sp-1            E-MUC1
H:    GGGGGCGGGGTTTTGTCACCTGTCACCTGCTCCGGCTGTGCCTAG
M:    ********a*******t************t*a**********

            -50                                -15
H:    GGCGGGCGGGCGGGGAGTGGGGGGACCGGTATAAAGCGGTAGGCG
M:    ******------------a*******t**********a*c****a
```

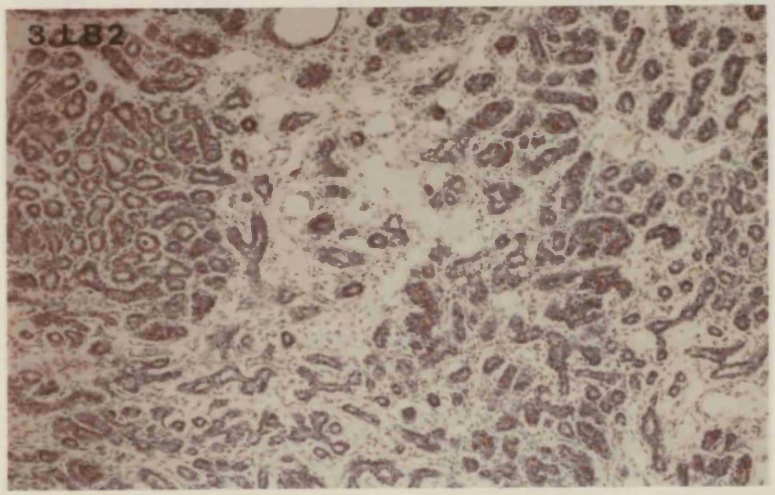
Figure 3.72 Expression of mouse Muc-1 mRNA in spontaneously arising mouse mammary carcinomas. Approximately 10 μg of each RNA sample were size-fractionated through a 1.0 % (w/v) formaldehyde agarose gel, alongside RNA prepared from C57Bl/ICRF mouse 1 day post partum lactating mammary gland (as positive control) and the human embryonic lung fibroblast cell line ICRF-23 (as negative control). After transfer to nylon membrane, the membrane was stained with methylene blue in order to visualise the 28S and 18S ribosomal RNAs and in order to estimate the loading (all samples had approximately the same amount of total RNA, data not shown). The membrane was hybridised overnight to the mouse Muc-1 probe pMuc2TR, and washed to high stringency. In all tracks, except the negative control, a single hybridising transcript, migrating at approximately 2.2 kilobase pairs was observed. The mouse mammary tumours that were utilised in each case were obtained from mice in early pregnancy. The expression of Muc-1 in the mouse mammary gland at this stage of pregnancy has been demonstrated to be low (Parry, 1992). However, a component of this apparent difference in expression levels in the developing mammary gland is the relative proportion of the epithelial component as compared to the total mammary gland. The majority of mammary carcinomas arise through proliferation of the lumenal epithelial cells of the mammary gland, and, therefore, the high level of expression observed in the mammary carcinomas investigated in this study may be due in part to the increased proportion of the epithelial component in the tumour.

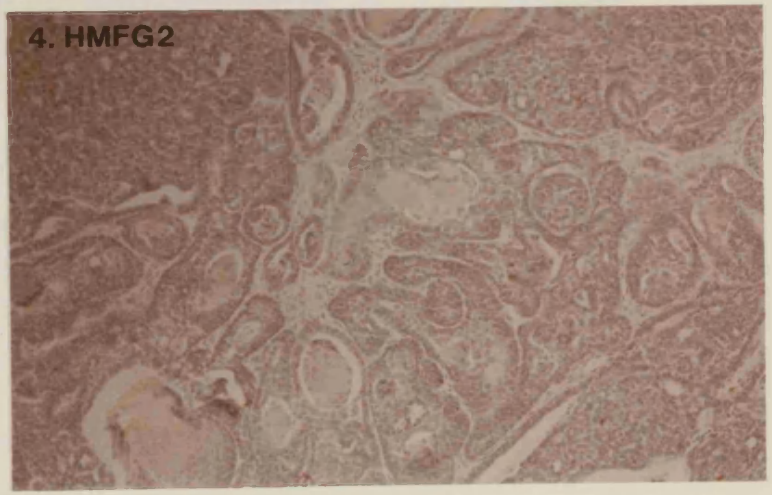

Lact. mouse (10μg)
mammary gland
mouse mammary gland
tumours (10μg)
ICRF-23 (10μg)
2.3kb>

Figure 3.73 **Expression of mouse Muc-1 protein in spontaneously arising mouse mammary carcinomas.** Parrafin embedded formalin fixed mouse mammary tumours were kindly provided by Dr. Clive Dickson, ICRF, London. Sections were prepared and the expression of mouse Muc-1 protein was analysed by immunohistochemistry with two antibodies. Monoclonal antibody LB2, an antibody that recognises epitopes present within the oligosaccharide component of the human mammary mucin (Moss, 1988) was used (Panels 1-3), along with the human MUC1 core protein directed antibody HMFG2 (Burchell, 1983) (Panel 4). All panels are at approximately 180 X magnification. Panels 1-3 are representative fields from within one section stained with the antibody LB2. The marked heterogeneity of staining can be easily seen (compare panels 1 and 2 with panel 3). Panel 4 is a representative field taken from a section stained with HMFG2. The field of view shown corresponds to the same area of the tumour as shown in panel 1. As can be seen, the antibody HMFG2 failed to cross-react with the mouse mammary mucin, Muc-1. An investigation of the protein sequences showed that the peptide epitope for HMFG2, present within the human MUC1 tandem repeat, is not conserved in the mouse Muc-1 repeat sequence. It appears, therefore, that where the core protein sequence of the Muc-1 gene has diverged, the carbohydrate side chains that are attached to the core have been maintained.


1. LB2


2. LB2


3. LB2


4. HMFG2

CHAPTER FOUR:

EVOLUTION OF THE Muc-1 GENE LOCUS.

4.1 Introduction

As early as 1973, it was proposed that a feature of mucin proteins would be the presence of repetitive sequences responsible for providing available attachment sites for O-linked carbohydrate (Pigman, 1973). Through molecular cloning techniques it has subsequently been demonstrated that mucin protein cores are indeed made up of multiple copies of tandem repeats that are rich in the amino acids serine and/or threonine and proline. The first human mucin gene, MUC1, to be isolated and characterised was made up mostly of multiple copies of a 60 base pair 82% GC-rich tandem repeat that encoded a 20 amino acid peptide repeat. Human mucin genes, MUC2-MUC6 have also been found to have a large part of their coding capacity comprised of tandem repeats (Table 1). This repetitive domain has been demonstrated to give mucin genes the characteristics of variable number tandem repeat (VNTR), or minisatellite loci; polymorphisms occur due to expansion and contraction of the number of repeats present within the repetitive domain. The fact that the repetitive portions of mucin genes actually code for an expressed protein product places mucin genes in an interesting class of sequence, the expressed minisatellite sequences.

Hypervariable minisatellite sequences are usually thought of as being non-coding so-called 'junk' DNA. As these sequences do not usually code for a protein product, fluctuations in the number of repeat units that may be present in any one allele is not thought to be subject to strong selection pressure. Fluctuations up or down in repeat unit number are thought to occur randomly, most likely through a process of unequal sister-chromatid exchange and/or DNA slippage (Wolff, 1989). These type of events have been demonstrated to occur in both somatic cells and in germ cells (Armour, 1989a; Kelly, 1991; Jeffreys, 1988a).

Recently, the evolutionary origin of minisatellite loci has begun to be considered (Gray, 1991). Through this investigation it was demonstrated that hypervariability at a VNTR locus can evolve comparatively rapidly and that loci that are observed to be extremely hypervariable in humans are not neccessarily found to be hypervariable in other species. Mucin genes represent a unique opportunity for the investigation of the role of selective processes on the evolution of the tandem repeat sequence and the length of the tandem array in expressed minisatellites. Gendler, 1990a, proposed that the tandem repeat sequence of mucin genes might exist simply as a scaffold for the attachment of O-linked carbohydrate. This would suggest that the most important residues present within a mucin tandem repeat would be the O-linked attachment sites, namely serines and/or threonines, in combination with the α-helix disrupting residue proline. In turn, this would predict that it is these residues that would be conserved through evolution. In this chapter experiments are described that consider how the Muc-1 tandem repeat sequence might have evolved in the Mammalia.

4.2 Investigation of naturally occurring minisatellite polymorphism of the mouse Muc-1 gene

The human MUC1 gene has been demonstrated to exhibit VNTR polymorphism through the presence of variable numbers of a 60 base pair tandem repeat. The sequence of the human MUC1 repeat unit appears to be extremely well conserved within a repeat array. However, an examination of the sequence obtained for the mouse homologue, Muc-1, revealed that the sequence of the tandem repeats was quite highly diverged (Fig 4.21). Homology between repeat units averaged at approximately 75%, the homology getting progressively worse towards the 3' end of the repeat array. In particular, 5 of the 16 repeats were 63 base pairs in length, encoding a 21 amino acid repeat unit, as opposed to the more common 20 amino acid repeat. In an attempt to investigate naturally occurring polymorphisms of the mouse Muc-1 gene, genomic DNA was prepared from tail snips from a large number of wild rodent samples (Diamond, 1990) originating in Europe (kindly provided by Peter King, University College London) and the United States. Screening of these samples through digestion with TaqI and EcoRI restriction endonucleases, and Southern analysis utilising the mouse Muc-1 tandem repeat cDNA probe pMuc2TR revealed the suprising result that the Muc-

1 gene was not variant in rodents (Fig 4.22). Analysis of greater than 50 wild rodent samples, including *Mus musculus domesticus, Mus spretus, Rattus norvegicus, Clethrionomys glareolus* (bank vole) and *Microtus agrestis* (short-tailed vole) revealed no variation in allele size. All samples exhibited a hybridising TaqI restriction fragment of approximately 1.6 kilobase pairs corresponding to a repeat array length of 16 repeats (Fig 4.22, 3.31 and 3.61).

This result, although surprising, correlated with the fact that the sequence of the mouse Muc-1 repeat was so poorly conserved. It is generally accepted that in VNTR loci that exhibit a high level of variation in allele lengths, the sequence of the repeat unit is found to be well maintained (Stephan, 1989). It is thought that this occurs through a process described as crossover fixation. A variant repeat arising through point mutation within a single repeat unit is expected to, through recombination, either spread within the repeat array or, alternatively, be lost altogether from the array (Jarman, 1989). Conversely, as no variation in allele length was detectable in the mouse Muc-1 gene, the sequence of the repeat unit would not be expected to be well maintained and would be subject over time to the gradual accumulation of point mutations.

Although length variation in the mouse Muc-1 gene was undetectable, a detailed pairwise comparison of the sequence of each of the mouse Muc-1 repeats revealed strong evidence for there having been a duplication of 5 repeat units within the repeat array (Fig 4.23). As mentioned previously, 5 of the 16 repeat units possessed an extra codon, in each instance located at the same position within each repeat, and several copies of this length variant could only have been generated by a process of duplication. Therefore, at some time during the evolutionary history of the modern mouse Muc-1 gene, the gene exerienced some expansions and possibly contractions of its repeat domain. One expansion resulted in the duplication of 5 repeat units (Fig 4.23) and the location of the unequal crossover event appears to have occurred between the ends of ancestral repeat numbers 1 and 6, respectively. In addition to this duplication, there appears to have been small fluctuations within the gene, for instance to generate the modern 21 codon repeat number 6.

Alignment of the two sequences that arose as a result of the duplication (Fig 4.23) allowed a calculation of the respective number of

synonymous and non-synonymous substitutions that have occurred since this region was duplicated (Nei, 1986). Synonymous substitutions are classified as nucleotide changes that do not alter the amino acid that the codon specifies, whereas non-synonymous substitutions are classified as nucleotide changes that would be expected to alter the specified amino acid. Through the method of Nei, 1986, this analysis revealed that there have been 7 non-synonymous substitutions and 1 synonymous substitution within the mouse Muc-1 duplication. In an attempt to arrive at a relative estimate for the length of time that has elapsed since the duplication, the duplicated sequences were compared with a similar sequence 201 base pairs in length made up of human MUC1 tandem repeats. A comparison of the mouse and human sequences revealed a total of 28 synonymous and 57 non-synonymous substitutions. Taking these figures into consideration, estimates for the number of synonymous and non-synonymous substitutions per site were derived (Appendix 1). Considering that the divergence of the human and mouse lineages has been estimated to have occurred between 80 and 100 milion years ago (Li, 1990), these estimates implied that the mouse Muc-1 duplication occurred somewhere in the order of 2 to 9 million years ago. Obviously, this figure is subject to a large amount of error, as the rate of non-synonymous nucleotide substitution has been found to be highly variable between different genetic loci (Li, 1985).

Through these calculations it became apparent that the number of non-synonymous substitutions was approximately 3-fold higher than the number of synonymous substitutions over the mouse Muc-1 duplication. In general, the rate of synonymous substitution within a coding sequence is significantly higher than the rate of non-synonymous substitution. The fact that the rate of non-synonymous substitution was found to be elevated within the mouse Muc-1 repeat sequence suggests that different areas within the repeat may be subject to differing selection pressure. The majority of the repeat sequence may not be that important function-wise, and it may be that it is primarily the sequences for the potential O-glycosylation sites that are strongly selected for. The other amino acids may be 'permitted' to fluctuate.

Although an accurate estimate for the mouse Muc-1 duplication could not be derived, the most important piece of evidence that places the duplication to have occurred at least 15 million years ago is the

observation that the allele size, as determined by Southern blots, of both the mouse and rat Muc-1 genes is identical (Spicer, 1991). These two species have been estimated to have diverged somewhere between 10 and 30 million years ago.

The evidence suggests, therefore, that the mouse Muc-1 gene has experienced size fluctuations at some time in the evolutionary past, but has subsequently become fixed in size at 16 repeats. Why the mouse Muc-1 gene lost or, alternatively, failed to develop the minisatellite-like characteristics is not clear. Gray, 1990 showed that non-expressed minisatellite loci can rapidly become hypervariable and just as rapidly become fixed in size. From the data presented thus far it is difficult to say that the mouse Muc-1 gene at one time exhibited hypervariability as marked as that seen for the human MUC1 gene, but it is apparent, however, that the modern mouse Muc-1 gene has evolved through a series of relatively recent repeat duplications.

4.3 Investigation of rodent Muc-1 evolution

The mouse Muc-1 gene was shown not be polymorphic in wild rodent samples from within the sub-order Myomorpha. However, the data presented above suggest that at some time in the evolutionary past the mouse Muc-1 gene may have exhibited a degree of minisatellite-like polymorphism. It was reasoned that if the gene could be demonstrated to have a fixed length in members of the rodent sub-orders, Myomorpha, Scuiromorpha and Caviomorpha, then it would be reasonable to assume that the fixation in the number of repeats may have occurred before the rodent lineage diverged. In addition, there has recently been some debate as to whether the guinea-pig is a true rodent or not (Graur, 1991). It was considered appropriate, therefore, to attempt to investigate the naturally occurring polymorphisms that may occur in members of the Scuiromorpha (squirrel-like rodents) and Caviomorpha (guinea-pig-like rodents) sub-orders.

Genomic DNA was isolated from frozen tissue of the species *Cavia porcellus* (guinea-pig) (kindly provided by Peg Davis, University of Arizona) and *Sciurus carolinensis* (squirrel) (kindly provided by Dr. Pat DeCoursey, University of South Carolina). DNA was restriction digested with the restriction endonuclease EcoRI and hybridised initially with the

mouse Muc-1 cDNA probe pMuc2TR, directed to the mouse Muc-1 tandem repeat. Hybridisation of this probe to mouse genomic DNA restriction digested with EcoRI yielded an 11 kilobase pair invariant hybridising fragment. Initial results indicated that hybridisation to squirrel genomic DNA revealed an invariant fragment of approximately 5.5 kilobase pairs in all six individuals, although the intensity of the hybridisation signal was weak (data not shown). The mouse Muc-1 tandem repeat probe failed to hybridise to the guinea-pig genomic DNA digests under the conditions that were used. In an attempt to obtain a better hybridisation signal, an alternative probe was utilised. This probe, pGMuc3', was a 600 base pair fragment of the guinea-pig Muc-1 cDNA obtained by reverse transcriptase-PCR (described below) which contained sequence encoding the membrane-spanning and cytoplasmic tails of the guinea pig Muc-1 homologue. Hybridisation was carried out with the pGMuc3' probe, against EcoRI digested mouse, squirrel and guinea-pig genomic DNA samples. This probe was observed to hybridise to an invariant 5.5 kilobase pair fragment in squirrel genomic DNA and to an invariant 6.5-7.0 kilobase fragment in guinea-pig genomic DNA (data not shown). These results suggest that the Muc-1 gene is monomorphic in the order Rodentia.

4.4 Chromosomal localisation of the mouse Muc-1 gene

The human MUC1 gene has been localised to chromosome 1q21 and has been shown to be closely linked to the α-spectrin gene (SPTA) (Swallow, 1987b and Middleton-Price, 1988). This area of human chromosome 1 has been shown to be syntenic with a region of mouse chromosome 1 and 3 (Kingsmore, 1989; Moseley, 1989a and 1989b; Dorin, 1990) and contains genes such as the selectin gene family of leukocyte adhesion molecules (Watson, 1990), the involucrin gene (Simon, 1989), the apolipoprotein A2 gene (Oakey, 1992a), the α-spectrin gene, the β-glucocerebrosidase gene (Ginn s, 1985) and the high affinity nerve growth factor receptor gene, Ntrk1 (Kingsmore, in press).

It has been suggested that mutation rates differ among different regions of the mammalian genome (Wolfe, 1989). It is interesting to find, therefore, that human non-expressed minisatellite sequences have been observed to cluster with other tandem repeats, often at the proterminal regions of human chromosomes (Royle, 1988 and Armour, 1989b). The

human MUC1 gene maps to a region of chromosome 1 that is comparatively rich in genes containing repetitive elements, and the MUC1 gene has been demonstrated to be rearranged in a high percentage of human breast tumours (Gendler, 1990b). In addition, immediately upstream of the human MUC1 gene, between MUC1 and the thrombospondin-3 gene there is an Alu repeat (N. Peat and J. Taylor-Papadimitriou, personal communications). The region of the genome to which human MUC1 is localised may, therefore, be a relatively hot-spot for recombination events in general. Taking all these observations into consideration, one possible reason for the potential loss of mouse Muc-1 polymorphism could be that the mouse gene is located within an area of the mouse genome in which the recombination rate is low or reduced in general, or in which the mutation rate is higher. In order to investigate this possibility, the chromosome localisation of the mouse Muc-1 gene was determined.

The chromosomal localisation of the mouse Muc-1 gene was determined with a panel of DNA samples from an interspecific cross that has been characterised for over 600 genetic markers throughout the genome. This study was carried out as part of a formal collaboration with Stephen F. Kingsmore and Michael Seldin of Duke University, North Carolina, USA). Restriction fragment length variants (RFLVs) were determined by Southern analysis of DNA from C3H/HeJ-*gld* and (C3H/HeJ-*gld* x Mus spretus)F_1 parental mice digested with various restriction endonucleases as previously described (Seldin, 1988). *Mus spretus* was chosen as the second parent because of the relative ease of detection of informative RFLV in comparison with crosses of inbred strains. The mouse Muc-1 probe, pMuc2TR, hybridised to 5.1 kilobase pair or 4.7 kilobase pair fragments in PvuII restriction endonuclease digests of C3H/HeJ or *M. spretus* DNA, respectively. Reference loci utilised were the genes encoding mouse calcyclin (*Cacy*), cluster designation 1 (*Cd1*), the high affinity Fcγ receptor (*Fcgr1*), β-glucocerebrosidase (*Gba*), liver red cell pyruvate kinase (*Pklr*), thrombospondin-3 (*Thbs-3*) and an anonymous DNA segment *D3Tu51*, all of which have been previously localised to mouse chromosome 3 (Oakey, 1992b; Kingsmore, in press; Dixit and Seldin, personal communication). One hundred and fourteen [(C3H/HeJ-*gld* x *Mus spretus*)F_1 x C3H/HeJ-*gld*] backcross mice were typed by segregation analysis of unique *M. spretus* RFLV detected with these probes (Green,

1981). At each locus, restriction digested DNA from the backcross mice displayed either a homozygous C3H or heterozygous F_1 pattern.

The RFLV associated with Muc-1 cosegregated with *Cd1*, *Pklr* and *Thbs-3* on chromosome 3. No cross-overs were detected between these four loci, suggesting tight linkage. Indeed, this would fit for the known close linkage of the thrombospondin-3 gene and the Muc-1 gene (Vos, 1992). Comparison of the sequences of the *Pklr* and *Cd1* genes indicated that neither represented the gene that has been identified as being located immediately downstream of the human and mouse Muc-1 genes. Haplotype analysis of the backcross mice is shown in Fig 4.41. From these results the probable gene order ($\pm$ the standard deviation, according to Bishop, 1985) was determined to be centromere- *Gba*-0.9 $\pm$ 0.9 centimorgans (cM) - (*Muc-1*, *Thbs-3*, *Cd1*, *Pklr*) - 0.9 $\pm$ 0.9 cM - (*Cacy*, *D3Tu51*, *Fcgr1*) - telomere (Stephen Kingsmore, personal communication). No cross-overs were detected between the genes grouped in parentheses.

This localisation makes the Muc-1 gene locus an additional member of a large linkage group conserved between mouse chromosome 3 and human chromosome 1q21-1p22 (Moseley, 1989a). This conserved linkage group spans the centromere on human chromosome 1 and includes the human MUC1 gene. The mouse Muc-1 gene was found to be tightly linked to the mouse *Cd1* gene locus, a locus that has been demonstrated to define a conserved linkage group border between human chromosomes 1 and mouse chromosomes 1 and 3 (Moseley, 1989b). Linkage of the mouse Muc-1 gene locus to the *Cd1* gene places it within 5 Mb (megabases) of the chromosome 1-3 break-point that has been identified (Moseley, 1989b; Oakey, 1992b and Stephen Kingsmore, personal communication) (Fig 4.42). Examination of the loci on mouse chromosomes 1 and 3 that have been localised to the regions of synteny to human chromosome 1 revealed that the mouse Muc-1 gene is not physically linked to the α-spectrin gene as is its human counterpart (Middleton-Price, 1988); the mouse α-spectrin gene has been localised to mouse chromosome 1 (Huebner, 1985). It appears, therefore, that the precise genetic environment of the mouse Muc-1 gene is indeed different from that of its human counterpart. It will be interesting to determine if this difference may have played a role in the evolution of the modern mouse Muc-1 gene locus.

4.5 Cloning the Muc-1 gene from diverse mammalian species

The fact that the VNTR portion of the human MUC1 gene is expressed implies that the VNTR polymorphism can not only be observed at the DNA level, but can also be observed at both the RNA and protein levels (Gendler, 1987b; Swallow, 1987a). As described previously, cells of the lactating mammary gland express the Muc-1 protein at one of the highest levels observed (Patton, 1986; Zotter, 1988), and this protein has been detected in the milk of all mammalian species thus far characterised (Patton, 1986; Patton, 1989; Patton, 1990; Welsch, 1990; Spicer, 1991; Campana, 1992; Patton, 1992). The presence of the Muc-1 protein in the milk makes an investigation of the polymorphic status of the molecule in a large number of mammalian lineages more possible, and in many ways easier than DNA analysis. The Muc-1 proteins in milk can be size-fractionated by SDS-PAGE and visualised by silver staining as described. These kind of studies have indicated that the Muc-1 protein is polymorphic in humans, other primates including chimpanzee and rhesus monkey, dogs and cats, cows, horses and goats (Swallow, 1987a; Patton, 1989; Spicer, 1991; Patton, 1990 and Campana, 1992). In each instance, the polymorphism was characterised by the presence of two major protein bands of ≥ 160 kilodaltons (Fig 4.51) that were inherited in a Mendelian fashion. As the two protein bands observed in human samples have been demonstrated to be the result of protein products translated from two alleles containing different numbers of a 60 base pair tandem repeat, it has been assumed that the polymorphism observed in other mammals is similarly based.

Through a comparison of the tandem repeat sequence of the human and mouse Muc-1 genes, it is apparent that the sequence has diverged widely (Fig 3.64 and 3.65). However, although the sequence homology was found to be low overall, the sequences corresponding to the potential sites for O-linked carbohydrate attachment were found to be conserved. Southern analysis of zooblots, utilising the human MUC1 tandem repeat probe pMUC7, indicated that the human probe cross-hybridised with primate genomic DNA only (Pemberton, 1992). Likewise, the mouse Muc-1 tandem repeat probe, pMuc2TR, was observed to cross-hybridise only with rodent genomic DNA samples. At the protein level, the monoclonal antibody HMFG2, which recognises the peptide epitope DTR within the human MUC1 repeat was found to cross-react on

Western blots only with the mucin isolated from human and chimpanzee milk (Fig. 4.52). It appears, therefore, that the sequence of the Muc-1 tandem repeat unit has diverged widely through evolution. These results led us to propose that if the primary function of the Muc-1 tandem repeat domain is to provide a scaffold for carbohydrate attachment, then the sequence of the tandem repeat unit could be 'allowed' to diverge within certain limits, those limits being the maintenance of potential O-glycosylation sites (Spicer, 1991). Thus, it is feasible that the Muc-1 tandem repeat sequence found in one mammalian order may be quite different in sequence from that present in another order. However, both tandem repeats would presumably still be playing the same role within the Muc-1 protein.

In order to investigate the evolution of the sequence of the Muc-1 tandem repeat it was decided to attempt to isolate fragments of the cDNA of the Muc-1 gene from a variety of different mammalian species. In contrast to the lack of sequence conservation of the repetitive domain, the sequence corresponding to the membrane-spanning and cytoplasmic domains of the Muc-1 gene has been well conserved (Spicer, 1991; Pemberton, 1992). Initially, therefore, it was decided to isolate DNA fragments of the 3' end of the Muc-1 gene from a variety of species, and then to use these clones and sequences to 'walk' 5' and isolate fragments corresponding to the tandem repeat domains.

Alignment of the human and mouse Muc-1 sequences permitted the identification of sequences suitable for use as conserved PCR oligonucleotides. Oligonucleotide primers were synthesised corresponding to the end of the cytoplasmic tail sequence of the human MUC1 gene and to a region immediately downstream of the human MUC1 tandem repeat domain. These primers had the following sequence: 5' conserved primer, 5'-AAC CTC CAG TTT AAT TCC TCT CTG GA-3' (corresponding to bases 873 to 908 of the published human MUC1 sequence , Gendler, 1990a); 3' conserved primer, 5'-CTA CAA GTT GGC AGA AGT-3' (corresponding to bases 1500 to 1483 (antisense) of the published human MUC1 sequence. The 5' primer had 22 out of 26 nucleotides conserved between human and mouse, whereas the 3' primer was 100% conserved between the two species. Total RNA was isolated from the following: rabbit kidney, guinea-pig kidney, cow lactating mammary gland, mouse lactating mammary gland, the human

breast carcinoma cell line T47D and the hamster pancreatic carcinoma cell line HP-1. Approximately 10-15 μg of each sample was reverse transcribed, as previously described, with the conserved 3' oligo described above. Following reverse transcription, the first strands were specifically amplified by the polymerase chain reaction, with the two conserved primers, through 45 cycles of 94°C 1 minute, 50°C 30 seconds, and 72°C 1 minute. After amplification, one-quarter of each reaction was analysed by agarose gel electrophoresis, alongside molecular weight markers produced by HaeIII restriction endonuclease digestion of bacteriophage ϕX174 DNA. A specifically amplified PCR product of the expected size (approximately 600 base pairs) was detected in all samples. The fragments were gel-purified on DEAE ion-exchange paper, ligated directly into the PCR T-vector, as described, and transformed into competent *E. coli*. The resultant plasmids were designated pRMuc3' (rabbit), pGMuc3' (guinea-pig), pCMuc3' (cow), pMMuc3' (mouse), pHuMuc3' (human) and pHaMuc3' (hamster).

Sequence was obtained for the inserts from both strands (Fig 4.53) and was translated into protein sequence (Fig 4.54). In addition, sequence for the gibbon homologue of the MUC1 gene was obtained from a genomic clone and kindly provided by Dr. Sandra Gendler and Trevor Duhig. Alignment of the respective protein sequences indicated that the membrane-spanning and cytoplasmic tail domains represented the most highly conserved sequences (Fig 4.54). Indeed, taking conservative changes into account, the sequence of the membrane-spanning domain was deduced to be 100% conserved in all species. This would suggest an important function for the transmembrane domain. It is entirely feasible that this region of the protein may specify the information that is important in directing the sequence to the apical pole of the cell. The presence of conserved cysteine residues opens up the intriguing possibility that monomers of the Muc-1 protein may be associated together. Alternatively, the Muc-1 protein may be linked to an accessory membrane protein through disulphide-bonding to the conserved cysteine residues. (Duwe, 1989).

A comparison of the sequence of the cytoplasmic domains revealed several important features (Fig 4.54). Of the seven tyrosine residues that were present within the sixty-nine residues of the human cytoplasmic tail, six were conserved in all species examined. A consensus

sequence for potential phosphorylation by protein kinase C (Ser Thr X Arg) was also found to be completely conserved and, interestingly, the residue corresponding to the allowed variable amino acid, X, in the consensus site was observed to vary amongst the sequences. Although it has not been shown, it is feasible, therefore, that the Muc-1 cytoplasmic domain is phosphorylated under certain conditions.

The MUC1 protein has been demonstrated to be associated with elements of the actin cytoskeleton (Parry, 1990). This association may be responsible, for instance, for maintaining the polarised distribution of the MUC1 molecule on the apical cell surface (Ojakian, 1988). Association of the MUC1 protein with elements of the actin cytoskeleton may occur through a link protein/s, such as the catenins that associate with the cytoplasmic domain of E-cadherin and link this molecule with the actin microfilaments (Ozawa, 1990). Alignment of the cytoplasmic domain sequences obtained for the different species described herein provides clues as to where potential link proteins may bind. In particular, a region of approximately 27 amino acid residues extending from the tyrosine residue at position 17 to the tyrosine at position 43 appeared to be extremely well conserved. A homology search of the GenBank database yielded no significant homology between this sequence and any other sequences known to bind to cytoskeletal link proteins. It is intriguing to speculate that this region of the protein may be crucial to the function of the Muc-1 molecule through, for instance, trafficking of the protein to the apical cell surface, through association with the actin cytoskeleton and perhaps even through some sort of regulatory mechanism via phosphorylation of tyrosines and/or of the potential protein kinase C site.

Amino-terminal of the membrane-spanning domain, the similarity between the respective sequences rapidly decreased. However, the region that has been implicated in the cleavage-associated secretion of the Muc-1 protein was found to be conserved. The two phenylalanine-arginine doublets that have been identified as possible cleavage targets by the kallikrein serine proteases (Ligtenberg, 1992a) were conserved as was a five residue sequence, Gly Ser Val Val Val that lies between these two potential cleavage sites (Fig 4.54).

The isolation of DNA clones for the 3' ends of the Muc-1 genes, described above, and a determination of their sequences permitted the design of species specific PCR oligonucleotides to be utilised in an attempt to amplify the repeat portion of the various Muc-1 homologues. These oligonucleotides were used in combination with a variety of oligos synthesised according to an alignment between the sequences of the human and mouse Muc-1 5' cDNAs. A variety of oligonucleotides were made and reverse transcriptase-PCR was attempted using a wide range of conditions, buffers and enzymes. Positive controls in each case were the 3' oligos utilised to isolate the 3' cDNA clones described above. However, after repeated attempts it was found that the repetitive portion of the Muc-1 genes could not be amplified efficiently from mRNA. In order to check the integrity of the RNA, oligonucleotides directed to conserved sequence in the 5' end of the mouse Muc-1 gene were utilised to attempt to amplify the 5' Muc-1 cDNA from mouse, hamster and guinea-pig mRNA. In this case, oligonucleotides were observed to efficiently amplify the respective 5' cDNAs, implying that the integrity of the RNA preparations was not the reason for the failure to amplify the repeat domain. Presumably, the presence of multiple copies of the tandem repeat and the GC-content (82% in humans) of the mRNA species made both the first-strand synthesis step and the PCR amplification difficult.

A different approach was required to attempt to isolate clones containing sequence of the repeat domain from other mammalian species. Numerous mucin repeat clones have been previously isolated from cDNA libraries and, therefore, cDNA library screening was considered as the next approach to attempt to isolate Muc-1 repeat clones. Lambda based cDNA libraries constructed from bovine and guinea-pig lactating mammary gland mRNA were kindly provided by Dr. Ian Mather, University of Maryland. The guinea-pig library was a λgt11 based library, whereas the bovine library was constructed in the λZAPII cloning vector. The guinea-pig library had previously been through one round of amplification, whereas the bovine library had been amplified twice. Both libraries were constructed from oligo-dT primed mRNA. Approximately 1.5×10^6 pfus of the guinea-pig library and 3×10^5 pfus of the bovine library where infected into the appropriate bacterial host as described and plated onto 24 x 24 cm LB-Agar plates. After overnight incubation, double lifts were taken from each plate and were hybridised with the guinea-pig 3' cDNA probe, pGMuc3', and the bovine 3' cDNA probe, pCMuc3',

respectively. Hybridisation and stringency washes proceeded according to standard Southern hybridisation procedures. Autoradiography indicated that 5 positive plaques were obtained from the guinea-pig library, whereas the bovine library yielded 23 positively hybridising plaques. All 5 of the guinea-pig plaques and 5 of the 23 bovine plaques were taken through two successive plaque purification steps to arrive at pure clones.

The five guinea-pig clones were rapidly screened by direct λ-plaque PCR, as described, to determine whether or not they contained additional sequence 5' of that present in the clone pGMuc3'. Unfortunately, all five guinea-pig clones were found to be identical and contained only an additional 50-60 base pairs of sequence 5' of that in pGMuc3'. This corresponded to a region approximately 100 base pairs downstream of the repeat domain. pBluescript phagemid DNA was rescued from the five bovine λZAPII clones using the λZAP excision protocol, as described, and used to transform competent *E. coli* cells. Upon plasmid preparation and analysis by restriction endonuclease digestion and agarose gel electrophoresis it became apparent that two of the bovine clones contained inserts of approximately 2 kilobase pairs in size. Sequence determination from these clones revealed the presence of tandem repeats 60 base pairs in length.

4.6 Evolution of the Muc-1 tandem repeat

Through a variety of cloning procedures the sequence of the tandem repeat domain was determined for a number of mammals. In addition to the sequences from the clones described above, the sequence for the rabbit Muc-1 repeat unit was obtained from a rabbit cervical mucin cDNA clone kindly provided by Dr. Beverly Chilton, Texas Tech University. The sequences obtained for the human, gibbon, bovine, rabbit and mouse Muc-1 repeat units, along with the other data presented allowed a consideration of how this portion of the Muc-1 gene may have evolved.

Sequence obtained from cDNA clones of the rabbit and bovine Muc-1 genes and from a genomic clone of the gibbon Muc-1 gene allowed a consensus repeat to be determined for each species. The gibbon and bovine repeats were found to be 60 base pairs (20 amino acids) in length, whereas the rabbit Muc-1 consensus repeat was only 57 base pairs (19

amino acids) in length. Unlike the mouse Muc-1 gene, in these three species each repeat shared close to 100% identity with the next. For reasons discussed previously, in combination with protein polymorphism data, this would suggest that the Muc-1 gene is hypervariable in these three species.

Alignment of the human, gibbon, bovine, rabbit and mouse Muc-1 consensus repeat units at both the nucleic acid and protein levels indicated unequivocally that all the consensus repeat sequences had been derived from a common ancestral Muc-1 repeat sequence (Fig 4.6). At several positions within the repeats the amino acid was found to be conserved but there had been base substitutions at the 'wobble' position. Although the percentage conservation between each consensus repeat was not extremely high (Fig 4.6), several areas were clearly conserved. As predicted, the most highly conserved areas corresponded primarily to potential sites for O-linked attachment of carbohydrate side chains in addition to proline residues. Taking into consideration all five sequences, the conserved residues appeared to be associated into two main regions of the repeat, each region containing potential carbohydrate attachment sites associated with proline residues (Fig 4.6). Other residues within the repeats appeared to be taken from within a restricted group of amino acids, including glutamate (D), alanine (A), valine (V) and glycine (G).

It appears, therefore, that the Muc-1 repeat domain that is present in mammals has evolved from a common ancestral repeat unit that was likely to be 60 base pairs in length. Collectively, primates, artiodactyls (cow, goat etc), lagomorphs (rabbit, hare) and rodents span up to 100 Myr of evolution (Li, 1990). This suggests that the Muc-1 repeat unit is up to 100 Myr old. Therefore, in an attempt to identify the possible ancestral Muc-1 gene, Southern analysis was carried out on various restriction endonuclease digestions of *Xenopus laevis* genomic DNA utilising the mouse Muc-1 cDNA probe, pMuc10. However, even under low stringency conditions this probe failed to hybridise to *Xenopus* DNA. As an alternative approach to the identification and isolation of the Xenopus Muc-1 gene, a probe corresponding to the 3' end of the mouse thrombospondin-3 gene was utilised. As this gene is tightly linked to the Muc-1 gene of both human and mouse, and is situated within only 3 kilobase pairs of the Muc-1 translation start site (Vos, 1992) it was reasoned that these genes might represent a conserved gene cluster. A 2

kilobase pair EcoRI-SmaI restriction fragment, containing the last three exons of the mouse Thbs-3 gene was hybridised against *Xenopus* genomic DNA as described above. At medium stringency (wash conditions of 1 x SSC, 0.1% SDS at 50°C) this probe appeared to detect specific restriction fragments present in EcoRI, HindIII and BamHI digested *Xenopus* genomic DNA (data not shown). These conditions were subsequently utilised to screen a *Xenopus* λFIXII genomic DNA library. Numerous weakly positive plaques were identified on the first screen, yet upon plaque-purification all plaques appeared positive at medium stringency and washing at higher stringency (0.5 x SSC, 0.1% SDS at 50°C) resulted in complete loss of any signal. Although clones for the *Xenopus* homologue of the Muc-1 gene were not isolated, this approach represents a possible strategy for the isolation of *Xenopus* Muc-1 and should not, therefore, be discounted. The fact that the Muc-1 gene is so tightly linked to the thrombospondin-3 gene in both human and mouse makes it likely that they may be an ancient conserved gene cluster. The characterisation of the *Xenopus* Muc-1 gene may reveal additional information regarding how the repeat domain has evolved and in addition may extend the current investigation of the human-mouse conserved autosomal chromosome segment (Kingsmore, 1989; Moseley, 1989a and 1989b) to the amphibians.

4.7 Conclusions

In this chapter several aspects of the evolution of the Muc-1 gene locus have been examined. These studies were initiated through the isolation and sequence determination of the mouse homologue of the human MUC1 gene, when it became apparent that the mouse Muc-1 gene did not exhibit the VNTR polymorphism that has now become characteristic of mucin genes in general. A screen of greater than 50 wild rodent DNA samples, including mice, rats and hamsters detected no variation in the size of the Muc-1 allele (Spicer, 1991). Closer inspection of the sequence of the mouse Muc-1 repeat domain revealed strong evidence for the modern mouse Muc-1 gene being formed through at least one duplication within the repeat region. This duplication may have occurred relatively recently in evolutionary history.

The repetitive portion of the Muc-1 gene forms a large part of the coding domain, and thus, in addition to the hypervariablility being

detectable at the nucleic acid level, the polymorphism can also be visualised at the protein level through SDS-PAGE analysis (Gendler, 1987b; Swallow, 1987a). The polymorphic status of the Muc-1 protein present in the milk of several different mammals, including primates (human, chimpanzee, rhesus monkey), carnivores (dog and cat), artiodactyls (goat and cow), perissodactyls (horse) and rodents (mouse and guinea-pig) was investigated by SDS-PAGE and silver staining. All mammals investigated, apart from mouse and guinea-pig, exhibited the characteristic doublet $\geq$ 160 kDa (Fig 4.51), the protein products of two Muc-1 alleles with differing numbers of tandem repeats. Thus, the rodents represented the only mammalian order in which the Muc-1 gene was not found to be hypervariable. However, as previously described, there was strong evidence that suggested that the modern mouse Muc-1 gene evolved through a series of duplication events within the repetitive domain.

It has been suggested that the mutation rate differs among regions of the mammalian genome and that it may be significantly higher in one region as compared with another (Wolfe, 1989). One factor that may have played a role in the mouse Muc-1 gene being distinct from the mammalian Muc-1 gene is its location within the mammalian genome. The human MUC1 gene has been previously localised to human chromosome 1q21 (Swallow, 1987b; Middleton-Price, 1988) and has been shown to be linked to the locus for α-spectrin. This region of chromosome 1 has been demonstrated to be syntenic with regions of mouse chromosome 1 and 3 (Kingsmore, 1989; Mosely, 1989a and 1989b). In order to determine its chromosomal localisation, the mouse Muc-1 gene was localised through haplotype analysis of interspecific backcrosses. Through this analysis, the mouse gene was localised to chromosome 3 and was found to cosegregate with markers for the genes thrombospondin-3 (Vos, 1992), Cd1 (Balk, 1989) and Pklr (Tani, 1988) (Fig 4.41). This analysis placed the Muc-1 gene within 5 Mb of the chromosome 1/3 breakpoint that has been identified (Moseley, 1989b) and implied that the mouse Muc-1 gene is not linked to the mouse α-spectrin gene (localised to mouse chromosome 1) (Fig 4.42). Thus, one factor that may have contributed to the evolution of the Muc-1 gene could have been either a translocation of the mouse Muc-1 gene to a region of the genome in which the recombination rate was suppressed in general, or alternatively, that after the divergence of the rodent lineage

the ancestral Muc-1 gene of all the other mammals was translocated within a large block of genes to a region of the genome in which the recombination frequency was elevated. Either way, it is apparent that there has been a major chromosome rearrangement in the close vicinity of the Muc-1 gene.

Through a variety of cloning procedures, clones were obtained for the conserved 3' ends and the tandem repeat domain of the Muc-1 gene from a number of mammals. Sequence comparison of the 3' ends of the gene revealed that the membrane-spanning and cytoplasmic domains have been highly conserved through evolution, suggesting an important role for these domains in the function of the Muc-1 protein (Fig 4.54). Several areas were found to be significantly conserved, and these included a consensus sequence for potential phosphorylation by protein kinase C, and six out of seven tyrosine residues in the cytoplasmic domain. The Muc-1 protein has been demonstrated to be associated with elements of the actin cytoskeletal network, and it is possible that a highly conserved area of 27 amino acids that was indentified is involved in this interaction. In addition to conservation of the membrane-spanning and cytoplasmic domains, sequences corresponding to previously identified sites that have been implicated in the protease-mediated cleavage and secretion of the Muc-1 protein were found to be conserved in all species investigated (Fig 4.54).

A comparison of the consensus tandem repeat sequences for human, gibbon, cow, rabbit and mouse indicated that all five sequences have evolved from a common ancestral sequence (Fig 4.6). The consensus repeat of human, gibbon, cow and mouse was 60 base pairs in length (20 amino acids), whereas the rabbit Muc-1 consensus was 57 base pairs in length (19 amino acids). Regions of the repeat that were found to be most highly conserved corresponded to potential O-glycosylation sites and proline residues. This fitted well with the hypothesis that if the primary role of the Muc-1 repeat domain is to act as a scaffold for carbohydrate attachment, then these sequences should be conserved through evolution. Overall, however, the repeats of human, cow, rabbit and mouse were not highly conserved, exhibiting between 30 to 55% identity (Fig 4.6).

Investigation of the molecular phylogeny of the orders Rodentia, Lagomorpha, Primates, Artiodactyla and Carnivora has suggested that the order Rodentia diverged prior to the other four (Li, 1990). With respect to the sequences presented herein this would imply that the oldest Muc-1 sequence obtained to date is the sequence of the mouse Muc-1 gene. At present, therefore, two alternative hypotheses are left regarding the evolution of the Muc-1 gene locus. The first hypothesis considers that the ancestral Muc-1 gene repeat region was hypervariable and that the length of the consensus repeat was set prior to the divergence of the order Rodentia. After the divergence of the order Rodentia, chromosome rearrangements and/or other factors played a role in contributing to the loss of polymorphism and the eventual fixation of the length of the mouse Muc-1 gene. The Muc-1 tandem repeats present in other mammals remained hypervariable and gradually became to be characteristic for species within that order or even sub-order. However, the functionally important residues present within the consensus repeat were conserved through selection. The second hypothesis considers that prior to the divergence of the order Rodentia, the ancestral Muc-1 gene had a repetitive region coding for a protein domain that functioned as a scaffold for carbohydrate attachment, but had not yet become truly hypervariable. Subsequent to the divergence of the rodents, duplication events occurred and/or the gene may have been translocated to a region of the genome where the recombination frequency was elevated. A combination of these factors may have contributed to the mammalian Muc-1 gene eventually becoming hypervariable. Within the rodent lineage, repeat-mediated duplications occurred, but the gene failed to become hypervariable. It may be that the accumulation of five repeats containing an extra codon exerted pairing constraints on any subsequent misaligned duplexes and thus reduced the rate of unequal exchange to such an extent that it allowed base substitutions to accumulate. The fact that the mutation rate in rodents has been demonstrated to be 1.5-2.0 fold higher per generation per year than in other mammals (Wu, 1985) may have also played a role in this process.

The fact that Muc-1 hypervariability can be relatively easily assessed through an investigation of Muc-1 protein present in milk, in addition to the fact that the Muc-1 gene forms part of a conserved linkage group, suggests that these alternative hypotheses could be evaluated

comparatively rapidly. The Muc-1 protein present in the milk of marsupials could be investigated, for instance, and tight linkage of the Muc-1 gene to thromospondin-3 may allow the isolation and characterisation of the Muc-1 gene from species as distant as *Xenopus laevis*.

Figure 4.21 Alignment of the 16 mouse Muc-1 repeats with the derived consensus repeats at both the nucleic acid and protein level. Dashed lines indicate a conserved base or amino acid residue, whereas a differing base or amino acid is as shown. It could be seen that repeats 4-6 and 9-10 possessed an extra codon at the same position and that homology to the consensus repeat decreased markedly in the last 6 of the 16 repeats. Overall, homology to the derived consensus rarely exceeded 90% at the DNA level. Comparison of any one of the 16 repeats with all others gave an average of 15 mismatches over 60 base pairs.

Mouse RepeatDNA Consensus

Codon Number	1	2	3	4	5	6	7		8	9	10	11	12	13	14	15	16	17	18	19	20	%HOMOLOGY TO CONSENSUS
	GAC	TCC	ACC	AGC	A/TCT	CCA	GTC	(GCC)	CAC	AGT	GGC	ACC	TCT	TCC	CCA	GCC	ACC	AGA	GCT	CCA	GA/TG	
								(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	
REPEAT NO. 1	-CA	--G	--T	-C-	---	---	---		---	--C	A--	-A-	--A	GA-	---	---	---	---	C--	---	- G -	78
REPEAT NO. 2	---	---	---	---	---	---	---		--G	---	A--	---	---	--T	---	---	---	---	---	---	--T - - A	92
REPEAT NO. 3	---	--T	---	--T	---	G--	---		-T-	---	---	---	--C	---	---	---	---	-C-	---	---	- - -	90
REPEAT NO. 4	A--	---	G--	---	---	---	--A	---	--T	G--	-A-	---	---	---	---	---	--T	--C	CT-	T--	A - A	79
REPEAT NO. 5	---	---	-A-	---	---	---	--A	-T-	---	---	---	---	---	--A	G-T	---	---	-C-	---	---	- - -	89
REPEAT NO. 6	--T	---	---	---	---	---	--A	-T-	---	G--	--T	---	--G	---	---	---	---	--C	C--	---	- G -	86
REPEAT NO. 7	---	---	---	---	---	---	-A-		--T	---	A--	---	---	--T	---	---	---	---	---	---	--C - - A	90
REPEAT NO. 8	---	--T	---	--T	---	G--	---		-T-	---	---	---	--C	---	---	---	---	-C-	---	---	- - -	90
REPEAT NO. 9	---	---	---	---	---	---	--A	---	--T	GA-	-A-	---	---	---	---	---	--T	--C	CT-	T--	- - A	83
REPEAT NO.10	---	---	G--	---	---	---	--A	---	---	G--	---	---	---	--T	---	---	---	--C	C--	-T-	A G -	84
REPEAT NO.11	---	---	---	--T	---	---	--T		---	---	A-T	G--	--C	AT-	-A-	AA-	-T-	-AG	A--	A--	T C A	67
REPEAT NO.12	---	-TA	G-T	---	---	---	-A-		---	-A-	---	---	--A	GT-	A--	A-T	---	--C	T--	G--	C - -	73
REPEAT NO.13	-G-	--A	G--	-C-	-G-	---	-A-		---	---	--T	---	--A	A-T	A--	A-T	-A-	--C	T--	GA-	T C A	65
REPEAT NO.14	-T-	-TG	G--	-C-	---	---	--T		T--	---	A--	-TG	C-A	-T-	T-T	A-T	---	-A-	-TG	A-G	T C A	58
REPEAT NO.15	-G-	--A	G-T	-T-	-T-	---	-A-		---	-A-	---	T--	--G	GTG	-T-	C-T	---	--T	T--	GTG	T - -	63
REPEAT NO.16	-G-	--A	G-T	-C-	-G-	-T-	---		T-T	-A-	AC-	T-T	G-A	ATA	G-T	A-A	--T	CCA	-TC	AGC	A - T	43

Mouse Repeat AA Consensus

Residue Number	1	2	3	4	5	6	7		8	9	10	11	12	13	14	15	16	17	18	19	20	
	D	S	T	S	S	P	V	(A)	H	S	G	T	S	S	P	A	T	S	A	P	V/E/S	
								(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	
REPEAT NO. 1	A	-	-	T	T	-	-		-	-	S	N	-	D	-	-	-	R	P	-	G	55
REPEAT NO. 2	-	-	-	-	-	-	-		Q	-	S	-	-	-	-	-	-	R	-	-	-	85
REPEAT NO. 3	-	-	-	-	T	A	-		L	-	-	-	-	-	-	-	-	T	-	-	-	80
REPEAT NO. 4	N	-	A	-	-	-	-	-	-	G	D	-	-	-	-	-	-	-	L	S	K	67
REPEAT NO. 5	-	-	N*	-	-	-	-	V	-	-	-	-	-	-	A	-	-	T	-	-	-	81
REPEAT NO. 6	-	-	-	-	-	-	-	V	-	G	-	-	-	-	-	-	-	-	P	-	G	81
REPEAT NO. 7	-	-	-	-	-	-	D		-	-	S	-	-	-	-	-	-	R	-	-	-	85
REPEAT NO. 8	-	-	-	-	T	A	-		L	-	-	-	-	-	-	-	-	T	-	-	-	80
REPEAT NO. 9	-	-	-	-	-	-	-	-	-	D	D	-	-	-	-	-	-	-	L	S	-	81
REPEAT NO.10	-	-	A	-	-	-	-	-	-	G	-	-	-	-	-	-	-	-	P	L	R	76
REPEAT NO.11	-	-	-	-	-	-	-		-	-	S	A	-	I	Q	N	I	K	T	T	-	50
REPEAT NO.12	-	L	A	-	T	-	D		-	N*	-	-	-	V	T	T	-	-	S	A	L	45
REPEAT NO.13	G	-	A	T	-	-	D		-	-	-	-	-	T	T	T	N*	-	S	E	-	45
REPEAT NO.14	V	L	A	T	T	-	-		Y	-	S	M	P	F	S	T	-	K	V	T	-	20
REPEAT NO.15	G	-	A	I	I	-	D		-	N*	-	S	-	V	L	P	-	-	S	V	L	35
REPEAT NO.16	G	-	A	T	-	L	-		Y	N*	T	S	A	I	A	T	-	P	V	S	N*	20

Figure 4.22 Variation at the mouse Muc-1 locus. Ten to fifteen micrograms of mouse tail genomic DNA were digested with the restriction endonuclease TaqI and size-fractionated through a 1.2% (w/v) agarose gel alongside λHindIII molecular weight markers. After electrophoresis, the DNA was transferred to nylon membrane and subjected to standard Southern analysis utilising the probe pMuc2TR. In all cases a single hybridising fragment of approximately 1.6 kilobase pairs was observed. This fragment was also detected in a screen of greater than 50 wild mouse samples indicating that the mouse Muc-1 gene is not polymorphic.

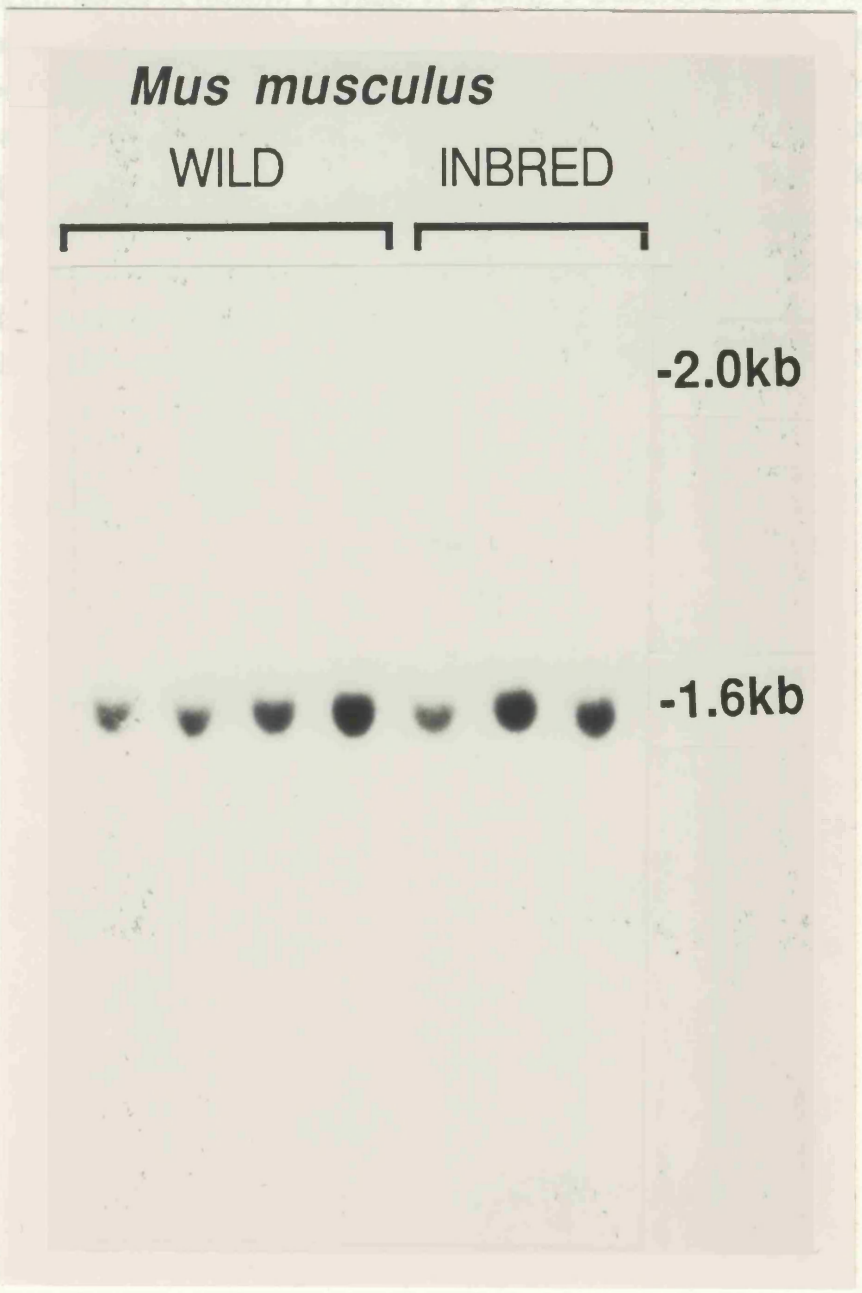
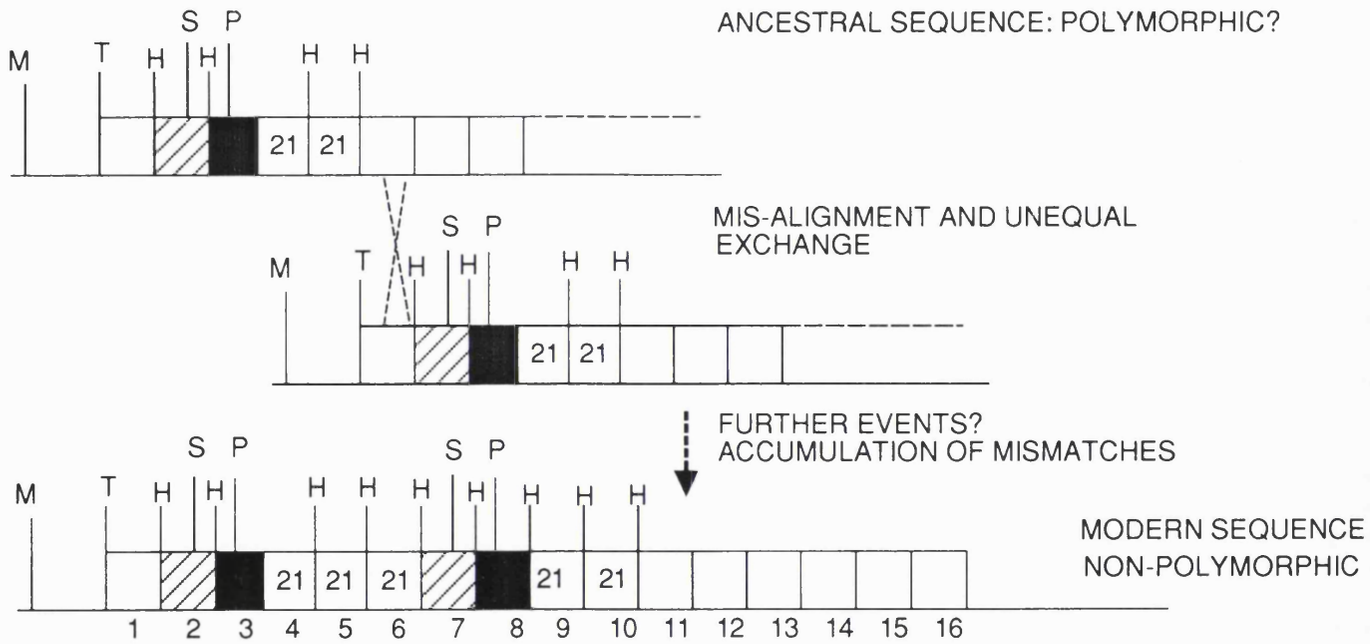

Mus musculus
WILD
INBRED
-2.0kb
-1.6kb

Figure 4.23 Diagrammatic representation of the unequal exchange event that is proposed to have taken place to generate the modern mouse Muc-1 gene. The misalignment is proposed to have been followed by a recombination event between ancestral repeat numbers 1 and 6 such that repeats 2-5 were tandemly duplicated. All sites for the restriction endonucleases M (MspI); T (TaqI); H (HinfI); S (SacI); and P (PstI) are shown. Solid boxes indicate identical repeats, striped boxes represent repeats that share close to 100% identity, 21s indicate those repeats with 21 amino acid residues as opposed to 20. Alignment between the modern sequences stretching from the end of repeat number 1 to the end of repeat number 4, and from the end of repeat number 6 to the end of repeat number 9, indicated very high homology with only 8 mismatches over 204 base pairs. In contrast, alignment of any one repeat with another yielded on average 15 mismatches over 60 base pairs.




```
REPEAT NO. 1                REPEAT NO. 2                                                    REPEAT NO.3
CCAGCCACCAGACCTCCAGGGGACTCCACCAGCTCTCCAGTCCAGAGTAGCACCTCTTCTCCAGCCACCAGAGCTCCTGAAGACTCTACCAGTACTGCAGTC
||||||||||| |||||||||||||||||||||||||| ||| |||||||||||||||||||||||||||||||||||| |||||||||||||||||||||||
CCAGCCACCAGCCCTCCAGGGGACTCCACCAGCTCTCCAGACCATAGTAGCACCTCTTCTCCAGCCACCAGAGCTCCCGAAGACTCTACCAGTACTGCAGTC
REPEAT NO. 6                REPEAT NO. 7                                                    REPEAT NO. 8

                                                REPEAT NO. 4
CTCAGTGGCACCTCCTCCCCAGCCACCACAGCTCCAGTGAACTCCGCCAGCTCTCCAGTAGCCCATGGTGACACCTCTTCCCCAGCCACTAGCCTTTCAAAA
|||||||||||||||||||||||||||||||||||||||||||| ||||| |||||||||||||||||||||| |||||||||||||||||||||||||||||||||||| ||
CTCAGTGGCACCTCCTCCCCAGCCACCACAGCTCCAGTGGACTCCACCAGCTCTCCAGTAGCCCATGATGACACCTCTTCCCCAGCCACTAGCCTTTCAGAA
                                                REPEAT NO. 9
```

Figure 4.41 Chromosomal localisation of the mouse Muc-1 gene through haplotype analysis of 114 [C3H/HeJ-*gld* x *Mus spretus*)F_1 x C3H/HeJ-*gld*] interspecific backcross mice. Segregation of *Cacy, Cd1, D3Tu51, Fcgr1, Gba, Muc-1, Pklr* and *Thbs-3* on mouse chromosome 3 in 114 interspecific backcross mice. Closed boxes represent the homozygous C3H pattern and open boxes the heterozygous F_1 pattern. Each column represents a chromosomal haplotype identified in the backcross progeny. No crossovers were detected between those loci that are grouped. It can be seen that in 112 of the 114 mice all the markers on chromosome 3 that were utilised cosegregated, suggesting fairly tight linkage. However, the Gba gene and the cluster, Fcgr1, Cacy, D3Tu51, were both observed to segregate from the Muc-1, Thbs-3, Cd1, Pklr cluster in one mouse, indicating that they were separated by a crossover event. This study localised the mouse Muc-1 gene to chromosome 3 and was carried out as part of a formal collaboration with Dr. Stephen F. Kingsmore and Dr. Michael Seldin of Duke University, North Carolina, USA.

Gba	■	□	■	□	■	□
Muc-1, Thbs-3, Cd1, Pklr	■	□	□	■	■	□
Fcgr1, Cacy, D3Tu51	■	□	□	■	□	■
No. of backcross mice	53	59	1	0	1	0

Figure 4.42 Approximate localisation of the mouse Muc-1 gene with reference to a breakpoint in homology that has been identified between mouse chromosomes 1 and 3 and human chromosome 1. Through haplotype analysis, the Muc-1 gene was found to cosegregate with the Cd1, Pklr and Thbs-3 genes. This placed the Muc-1 gene within 5 Mb of the previously identified position of a breakpoint in an extended region of synteny between human chromosome 1 and mouse chromosomes 1 and 3. The approximate position of the mouse Muc-1 gene is indicated here in relation to the breakpoint. Asterisk indicates the approximate map position of the human MUC1 gene with respect to the positions of genes that have been shown to be localised to this region of chromosome 1. Modified from Oakey, 1992b.

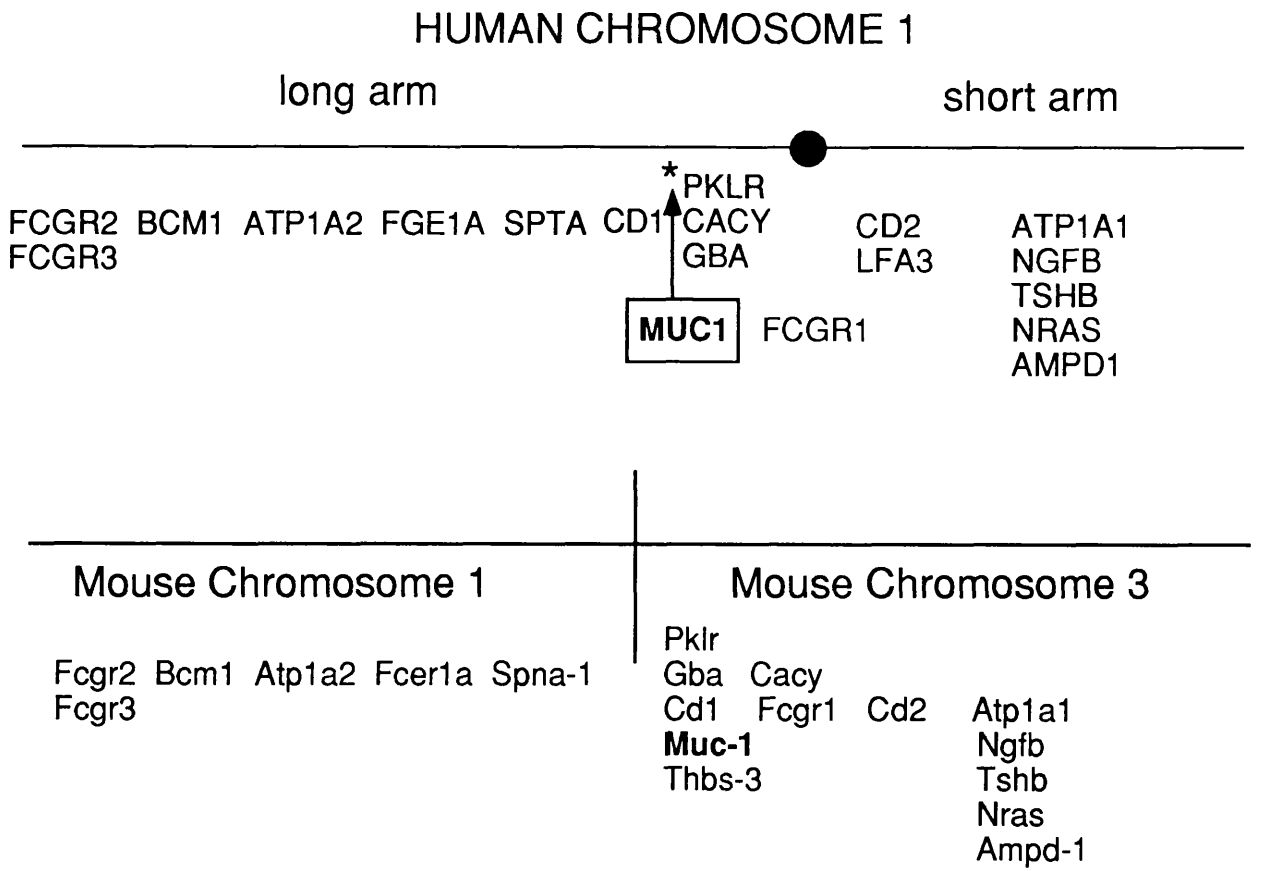

HUMAN CHROMOSOME 1
long arm
short arm
FCGR2 BCM1 ATP1A2 FGE1A SPTA CD1
FCGR3
*PKLR
CACY
GBA
MUC1
FCGR1
CD2
LFA3
ATP1A1
NGFB
TSHB
NRAS
AMPD1
Mouse Chromosome 1
Fcgr2 Bcm1 Atp1a2 Fcer1a Spna-1
Fcgr3
Mouse Chromosome 3
Pklr
Gba Cacy
Cd1 Fcgr1 Cd2
Muc-1
Thbs-3
Atp1a1
Ngfb
Tshb
Nras
Ampd-1

Figure 4.51 Milk-fat-globule polymorphism of Muc-1 in a variety of mammalian species. Approximately 5 μg milk-fat-globule proteins isolated from milk of a variety of mammalian species were analysed as described. All lanes revealed a high molecular mass milk-fat-globule protein ≥ 160 kDa. In all lanes, except mouse and guinea-pig, the protein was characterised by the existence of two bands, presumably from two respective different length alleles.

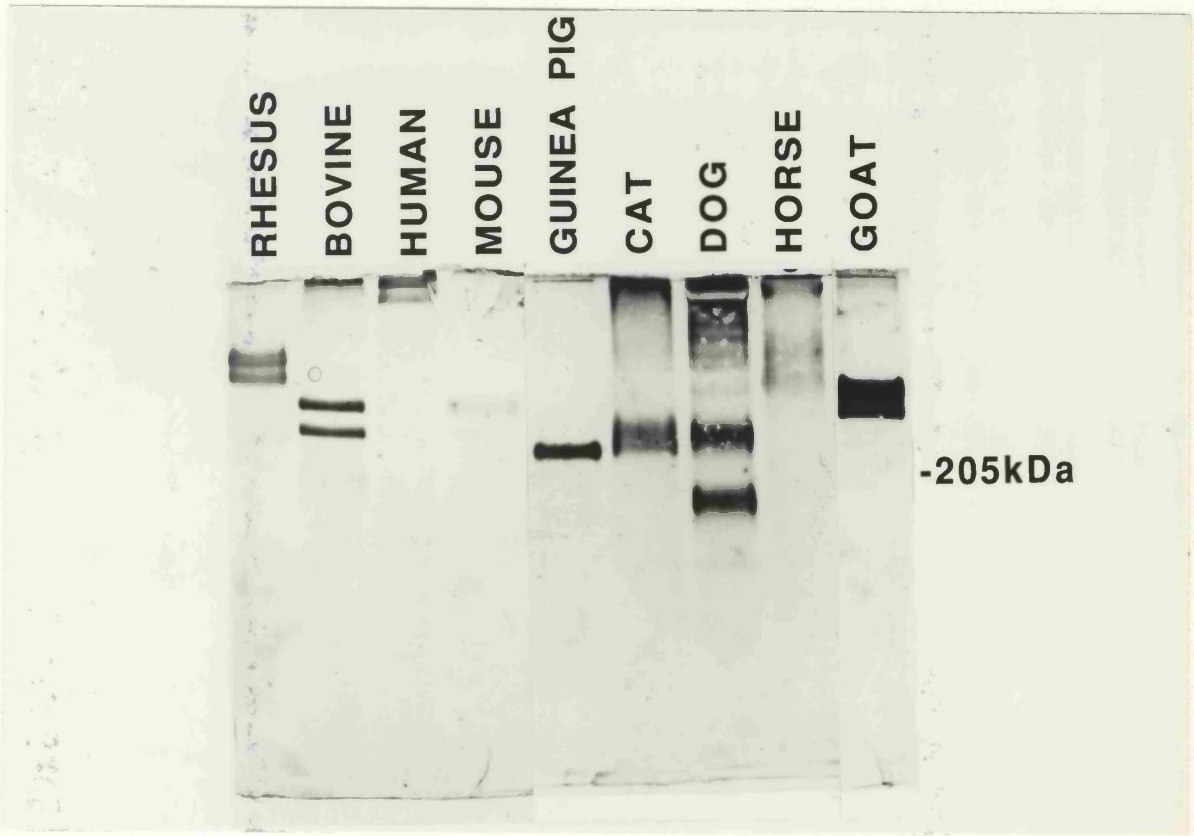

RHESUS
BOVINE
HUMAN
MOUSE
GUINEA PIG
CAT
DOG
HORSE
GOAT
-205kDa

Figure 4.52 Western blot of Muc-1 milk-fat-globule proteins screened with the monoclonal antibody HMFG2. Approximately 5-10 μg milk-fat-globule proteins were size-fractionated by SDS-PAGE through a 5% running gel and 3% stacking gel. Proteins were transferred to nylon membrane and subjected to Western analysis utilising the monoclonal antibody, HMFG2, that recognises the peptide epitope DTR, within the human MUC1 repeat. This antibody cross-reacted only with the human and chimpanzee Muc-1 proteins suggesting that the sequence of the Muc-1 tandem repeat has diverged through evolution. The proteins remaining in the gel after transfer were visualised by silver staining as described. Data kindly provided by Dr. Stu Patton, University of California at San Diego.

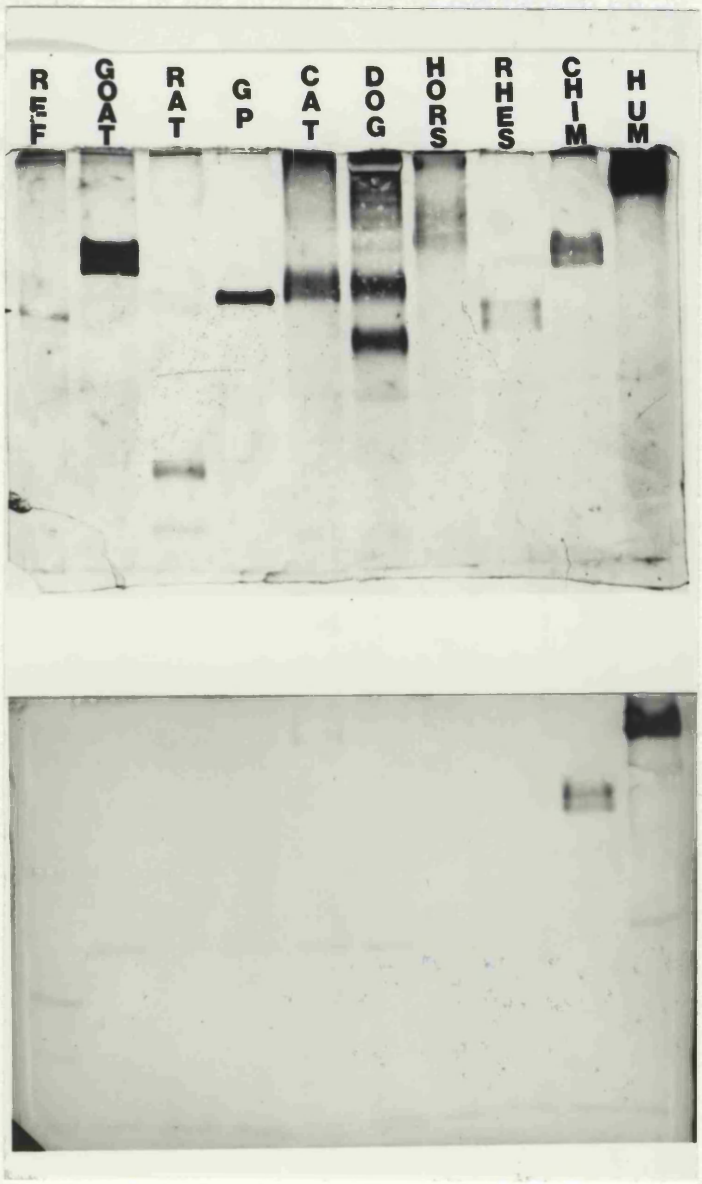
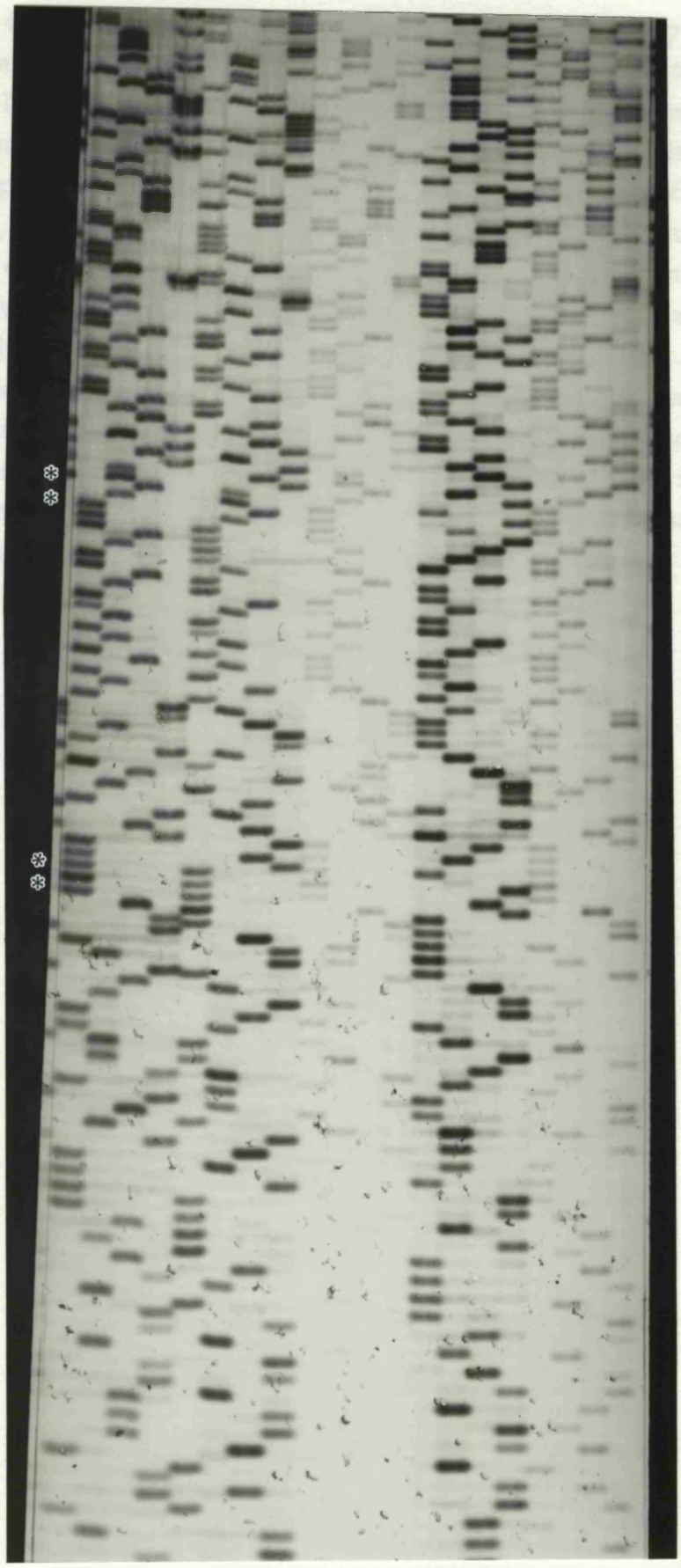

REF
GOAT
RAT
GP
CAT
DOG
HORS
RHES
CHIM
HUM

Figure 4.53 Sequencing gel displaying equivalent sequence of the mouse, hamster, guinea-pig, cow and rabbit Muc-1 cytoplasmic tail domains. Approximately 5 μg of double-stranded plasmid DNA were sequenced by the di-deoxy chain termination method as described. Equivalent Muc-1 cDNA clones were sequenced in each instance using the same oligonucleotide primer. The sequence shown represents the sequence of the cytoplasmic tail of the various species (sequenced on the antisense strand). Close inspection revealed regions of high sequence conservation. Asterisks (**) indicate areas noted to be particularly well conserved. The bovine and hamster cDNA clones were found to lack 2 and 1 codons, respectively, within the cytoplasmic tail sequence, and this resulted in a shift downwards of the homologous sequence observed on the gel. The reactions were loaded in the order GTAC.



MOUSE HAMSTER G.PIG BOVINE RABBIT

Figure 4.54 Conservation of the membrane-spanning and cytoplasmic tail domains of Muc-1. Sequence was determined from both DNA strands of Muc-1 DNA clones for all the species listed. Utilising known sequence for the human and mouse Muc-1 genes, the sequences were translated into the respective amino acid sequences. Shown is a comparison of the gibbon, bovine, rabbit, mouse and hamster Muc-1 carboxy-terminal region with the equivalent region of the human MUC1 protein. Conserved amino acid residues are indicated by a dash. Spaces are indicative of the codon being absent from that sequence. Boxed areas represent the potential proteolytic cleavage sites amino-terminal of the membrane-spanning domain, and the membrane-spanning and cytoplasmic domains, respectively. Conserved tyrosines (Y) are italicised. The consensus site for potential phosphorylation by protein kinase C is underlined.

```
HUMAN   EDPSTDYYQELQRDISEMFLQIYKQGGFLGLSNIKFRPGSVVVQLTLAFREGTINVHDVETQFNQYKTEA
GIBBON  -----N-----------LI-------D---V-----------------S---------T------A----H----
COW     -N-Q-S-------S-WGLI------RD-----E-----------E---------- TAEW-KA--S-LEAH-
RABBIT  -V--SK-------NV-ALIS-M-G-KN-----G-R---------D-I---------T-EVT-QS--I-KIPQ-
MOUSE   ----SN-----K-N--GL----FN -D---I-S-----S-----ES-VV------FSAS--KS-LI-H-K--
HAMSTER ----SN-----K-NV-GL---VFS RA---I-T-E--S-----DS-VI-----RV-ASE-KS-LL-HEQ--
G. PIG  -N---R-----E-N-TRL-----Q QD----L-----------A-ES-VI--KNAV-EEF-KS-LT-RK -D
```

TRANSMEMBRANE

```
HUMAN   ASRYNLTISDVSVSDVPFPF SAQSGAGVPGWGIALLVLVCVLVALAIVYLIALAVCQCRRKNYGQLDIFP
GIBBON  -------------------- ------------------------------------------------
COW     -- ------G---YSA---S ---A-T---------------------I-----V----G--KCE---V--
RABBIT  -- ---N--R-T-R---S-S ----  ---------------L----I-F----------S----L-R
MOUSE   D- ------E-K-NEMQ--P ----RP-------------I--------FL----------S--------
HAMSTER EE ---A--KIN-GEMQ--S ----WP-------------I------------------------------
G. PIG  -T----V--ED-ARE-QVTSST---  ---------------L----------M------S--H------
```

CYTOPLASMIC TAIL

```
HUMAN   ARDTYHPMSEYPTYHTHGRYVPPSSTDRSPYEKVSAGNGGSSLSYTNPAVAATSANL.
GIBBON  ---A----------------------N--------E---------------------.
COW     TL-A-------S-----------P--K-----E--------N-----  L-------.
RABBIT  PG----------------------G--R-----E--------G---------------.
MOUSE   TQ----------------------G--K-----E------S----------VT-----.
HAMSTER IQ----------------------G--K-----E------ --------V--T-----.
G.PIG   T-----------------------G--K-----E--T----G----------V-----.
```

4.6 Alignment and comparison of the consensus Muc-1 tandem repeats obtained for human (H), gibbon (G), cow (B), rabbit (R) and mouse (M). Through sequence analysis of a variety of cDNA and genomic clones the consensus sequences for the gibbon, cow, rabbit and mouse Muc-1 tandem repeats were determined at both the nucleic acid and protein levels. Shown is a comparison of the consensus repeats with the human MUC1 consensus repeat. Areas of highest sequence conservation (boxed) were found to correspond to potential attachment sites for O-linked carbohydrate and proline residues. Below is a table summarising the homology between the respective consensus repeat units.

	1	2	3	4	5	6	7		8	9	10	11	12	13	14	15	16	17	18	19	20
H:	GGC	TCC	ACC	GCC	CCC	CCA	GCC		CAC	GGT	GTC	ACC	TCG	GCC	CCG	GAC	ACC	AGG	CCG	GCC	CCG
G:	--G	---	---	---	---	---	-T-		---	---	---	---	---	---	---	---	---	---	--C	CA-	-TT
B:	-A-	C-A	--T	A--	T-T	--G	AA-		-G-	A-C	AAG	---	--A	--T	---	-C-	GA-	GAC	--T	A--	---
R:	-T-	--A	G--	A--	AG-	--T	-T-		---	-AA	---	---	--T	---	---	-C-	---	--C	---		T--
M:	-A-	---	---	AG-	T-T	---	-T-	GCC	---	A--	-G-	---	--T	T--	--A	-C-	---	--A	G-T	C-A	GA-

H:	G	S	T	A	P	P	A		H	G	V	T	S	A	P	D	T	R	P	A	P
G:	-	-	-	A	-	-	V		-	-	-	-	-	-	-	-	-	-	-	H	L
B:	D	P	-	T	S	-	N		R	S	K	-	-	-	-	A	D	D	-	T	-
R:	V	-	A	T	S	-	V		-	E	-	-	-	V	-	A	-	S	-		S
M:	D	-	-	S	S	-	V	A	-	S	G	-	-	S	-	A	-	S	A	P	X

PERCENT HOMOLOGY BETWEEN REPEAT UNITS AT THE AMINO ACID LEVEL

	M	R	B	G	H
H	40	45	40	85	
G	45	50	35		
B	40	40			
R	55				

CHAPTER FIVE:

TARGETED INACTIVATION OF THE MOUSE Muc-1 GENE.

5.1 Introduction

Mucin proteins are associated with the apical or lumenal surface of epithelial cells, and as such they have been supposed to play primarily a protective role, protecting the epithelia for instance from bacterial attack or dessication. Traditionally, mucin glycoproteins have been assumed to make up the characteristic viscous mucous found in the tracheobronchial tract, the digestive system and the female reproductive tract. Indeed, of the mucin genes that have been isolated it appears that the majority code for extremely large secreted glycoproteins that have the potential to form huge extended complexes through disulphide bonding. However, the first gene isolated for a human mucin glycoprotein, MUC1, was found to encode a protein product that was not of the typical secretory mucin type, but was, rather, a membrane-associated mucin-like glycoprotein.

The MUC1 glycoprotein has been demonstrated to be expressed by the majority of simple secretory epithelial organs (Zotter, 1988; Braga, 1992; Peat, 1992). This appears to be in contrast to the other human mucin genes that have been isolated, which have been shown to have a much more restricted spatial expression pattern (Gum, 1989; Gum, 1990; Porchet, 1991; Aubert, 1991; Toribara, 1993). In all the normal tissues where MUC1 is expressed it is localised to the apical or lumenal surface. In carcinoma, however, epithelial cells progress towards a non-polarised state and in these cells the MUC1 protein can be detected on all surfaces.

Since the initial isolation and characterisation of the human MUC1 gene, numerous functions have been attributed to its protein product. Functions that have been attributed to the MUC1 glycoprotein

are in general a reflection of its large size. It has been estimated that the fully glycosylated human MUC1 protein may extend as far as 300 to 500 nm above the cell surface. As such, this molecule is likely to be one of the largest molecules present on the surface of MUC1 expressing cells. Transfection of human MUC1 expression constructs into cells in tissue culture demonstrated that the human MUC1 protein product is capable of blocking the homotypic adhesion molecule E-cadherin (Wesseling, 1992). It is possible, therefore, that one of the major biological roles of the MUC1 glycoprotein is in affecting the adhesive properties of molecules such as E-cadherin.

Possible functions that have been proposed for the MUC1 membrane glycoprotein range from a role in the metastatic spread of carcinomas to a possible role in the organogenesis of epithelial lumens. However, the majority of functions that have been attributed to this molecule have been attributed through inference and association. Recently, through the advent of mouse embryonic stem cell technology and the development of methods to mutate specific genes through gene targeting, it has become possible to create mice deficient in a particular protein product. Thus, the effect of a lack of one specific protein can now be analysed *in vivo*. This is a powerful technique for an investigation into the function of a particular protein, and it has been used successfully in the functional analysis of a large number of gene products. The production of mice deficient in Muc-1 protein may allow us to better define the true role of Muc-1 both in the normal situation and in carcinomas. This chapter describes experiments designed to specifically mutate the mouse Muc-1 gene and generate mice deficient in the Muc-1 membrane-associated mucin glycoprotein.

5.2 Targeting vector design and construction

Genomic clones containing the entire mouse Muc-1 gene were isolated from a Balb/c mouse cosmid library as described previously. Extensive sequencing and restriction digest analysis allowed a detailed restriction map to be drawn up of the mouse Muc-1 gene. The gene spans approximately 4.5 kilobase pairs and was found to be centrally located within an 11 kilobase pair EcoRI restriction endonuclease fragment (Fig. 5.21). Approximately 2.6 kilobase pairs upstream of the transcription start site of both the human and mouse Muc-1 genes, the last exon and

polyadenylation signal of a second gene, the thrombospondin-3 (Thbs-3) gene, was identified (Vos, 1992). In addition, approximately 3.0 kilobase pairs downstream of the human and mouse Muc-1 polyadenylation sequences there is evidence for the presence of a third gene encoding a large (>10 kilobase pairs) transcript that appears to be ubiquitously expressed (S. Gendler, T. Duhig, and H. Vos personal communications). All three genes are in the same transcriptional orientation and appear to make up a tightly linked cluster of genes. However, there is no evidence for a common evolutionary origin of the three genes and their expression profiles appear to be independently regulated (Zotter, 1988; Braga, 1992, Peat, 1992; Vos, 1992; Kovarik, 1993, S. Gendler and T. Duhig, personal communication).

In an effort to mutate the mouse Muc-1 gene, a targeting replacement vector was designed and constructed (Fig. 5.22). The vector was designed such that a selectable marker gene for the neomycin phosphotransferase gene product, driven by the mouse phosphoglycerate kinase-1 (Pgk-1) promoter and polyadenylation signals (kindly provided by Dr. Mike McBurney, University of Ottawa) (Tybulewicz, 1991) was inserted into the first exon of the Muc-1 gene, replacing the sequence corresponding to introns 1 through 5 and exons 2 through 6 of the Muc-1 gene.

In order to make subsequent cloning manipulations easier, the pgkneo cassette, in pKJ-1, was subcloned into pBluescriptSKII+ as an EcoRI-HindIII restriction endonuclease fragment (Fig 5.22). This plasmid was designated pBS-pgkneo. An approximately 4.0 kilobase pair EcoRI-SmaI restriction fragment of the mouse Muc-1 gene was ligated and subcloned into pBluescriptKSII+ and designated pBS-5'Muc-1. The pgkneo cassette, in pBluescript, was isolated as an XbaI-ClaI restriction fragment, and the 5' Muc-1 arm of homology was isolated as a HindIII-SpeI restriction fragment. Both fragments were ligated into the HindIII-ClaI sites of the vector pBluescriptSKII+ (SpeI and XbaI sites are complementary) to yield the plasmid pBS-5'pgkneo (Fig 5.22).

To create the 3' arm of homology for the targeting vector, several factors were taken into consideration. Screening for homologous recombinants using the polymerase chain reaction (PCR) requires that one of the arms of homology be generally less than 2 kilobase pairs in

size (Kim, 1988). In addition, it was reasoned that as the Muc-1 gene is relatively compact it would be possible to delete most of the coding domain of the gene by deleting as little as 3 kilobase pairs of genomic sequence. We reasoned that deletion of most of the coding domain would be certain to result in the creation of a null allele. A third reason that influenced the choice of the 3' arm of homology came from work that demonstrated that the presence of multiple copies of the human minisatellite consensus core sequence (5'-AGAGGTGGGCAGGTGG-3') resulted in an approximately 14-fold increase in intrachromosomal homologous recombination between plasmids in mammalian cells (Wahls, 1990). As the mouse Muc-1 gene was found to contain a repetitive domain with sequences sharing some homology to the minisatellite consensus core sequence (5'-AGAGGTGCCACTGTGG-3', Muc-1 consensus repeat antisense), we decided to attempt to investigate the effect of the presence of Muc-1 repeats on the frequency of homologous recombination in ES cells. We decided, therefore, to create two different targeting vectors (Fig 5.22). In each vector, the 5' 4.0 kilobase pair EcoRI-SmaI restriction fragment was identical. However, the 3' arms differed. In one vector, designated pMuc-1GT Type I, the 3' arm was approximately 1.0 kilobase pair in size and included sequence that extended from the HindIII restriction endonuclease site present at the start of the last intron, to a region immediately upstream of the Muc-1 polyadenylation sequence. The 3' arm of the second targeting vector, designated pMuc-1GT Type II, was made up of the repetitive domain of the mouse Muc-1 gene and was approximately 950 base pairs in size.

The sequence for each of the 3' arms of homology was amplified using the polymerase chain reaction (PCR) with specific synthetic oligonucleotides designed to the known sequence for the mouse Muc-1 gene (Appendix 2). PCR was carried out on pMucEco plasmid DNA using standard conditions. In each case, oligonucleotides were designed such that the specific PCR product could be restriction digested with the restriction endonucleases HindIII (at the 5' end) and NotI (at the 3' end) (Fig 5.22).

In an attempt to enrich for homologous recombinants, the Muc-1 targeting vectors were designed with the Herpes Simplex virus thymidine kinase (HSV-tk) gene flanking the 5' arm of homology (the HSV-tk gene was kindly provided by Dr Randall Johnson, Dana-Farber

Cancer Institute). The HSV-tk gene was provided as a 3.4 kilobase pair BamHI fragment of HSV-1 cloned into the unique NaeI restriction endonuclease site of the vector pBluescriptSK+ (Johnson, 1989). The 3' arms of homology were ligated into the HindIII-NotI site of the pBluescriptSK-HSV-tk plasmid. These plasmids were designated 3' Type I and 3' Type II, respectively (Fig 5.22). To create the final targeting vectors, the 5'Muc-1/pgkneo cassette was sub-cloned as an approximately 6.0 kilobase pair HindIII restriction fragment into the HindIII restriction sites of the two 3'-HSV-tk vectors. The PCR amplified 3' arms of homology were sequenced to ensure that no polymerase errors had occurred during the PCR amplification. In addition, the sequences of all ligation junctions were confirmed. Both targeting vectors were designed such that they could be linearised prior to electroporation at a unique NotI restriction endonuclease site (Fig 5.22).

5.3 Establishment of conditions for use in the gene targeting experiments

Prior to the electroporation of the targeting vectors into the mouse embryonic stem cell lines E14TG2a and GK129, the optimal concentration of G418 for selection purposes was determined by growing ES cells at normal cell densities for ten days on a feeder layer of G418 resistant STO-neos at concentrations of G418 ranging from 50 μg/ml (active constituent) up to 500 μg/ml (active constituent). The optimal concentration for use with both ES cell lines was found to be 400 μg/ml G418 and this concentration was used throughout. The electroporation conditions were also optimised, using published values for reference purposes. Approximately 4 X 10^7 ES cells were electroporated in a total volume of 800 μl PBS with a luciferase reporter construct (Andreason, 1989) at an approximate concentration of 5 nM (nanomolar). Cells were electroporated using a BioRad Gene-Pulser at a constant capacitance of 250 μF (microFarads) with voltages ranging from 150 to 400 volts. The efficiency of electroporation was determined 48 hours later by a calculation of the luciferase activity, expressed as a percentage of total cellular proteins (data not shown). In this way it was found that the optimal electroporation conditions for use with the ES cell line E14TG2a were 350 V, 250 μF.

Identification of homologous recombinants using the polymerase chain reaction allows the rapid screening of potentially thousands of

antibiotic resistant ES colonies (Kim, 1988). In order to optimise conditions for use in the PCR screening of Muc-1 mutated ES clones we designed and constructed two additional vectors. PCR identification of homologous recombinants relies on the use of a primer specific to the targeting vector (usually within the neo gene) in combination with a primer specific for sequence present within the endogenous flanking chromosomal DNA. Homologous recombination between the targeting vector and its chromosomal target results in the juxtaposition of the two PCR primers (Kim, 1988). A PCR amplified product of the expected size should, therefore, only be amplified from cells in which the targeting vector has integrated by homologous recombination. We designed two vectors that consisted of the 5' Muc-1 and pgkneo cassette ligated to a 3' Muc-1 derived sequence. The 3' Muc-1 sequence utilised was in each case similar to those present in the targeting vectors pMuc-1GT Type I and II, but contained more 3' Muc-1 sequence (Appendix 2). These constructs were electroporated into E14TG2a cells, as described, and G418 resistant colonies were picked for optimisation of PCR conditions. Briefly, a single G418 resistant colony was picked and pooled with either 2, 5 or 10 colony equivalents (we estimated that a single colony was made up of ≤ 500 cells) into 100 µl trypsin/EDTA as described. After inactivation of the trypsin with an equal volume of prewarmed complete medium, 100 µl of the cell suspension was transferred to a microcentrifuge tube. Cells were recovered by centrifugation at approximately 2,000 rpm for 5 minutes at room temperature. Medium was aspirated carefully, leaving a visible cell pellet behind. To each cell pellet, 25 µl sterile distiled water containing 20 µg Proteinase K were added. The tubes were mixed and incubated in a PCR thermal cycler (Hybaid, United Kingdom) at 55°C for 30 minutes followed by a 10 minute incubation at 95°C, in order to inactivate the Proteinase K. After inactivation of the enzyme, tubes were spun briefly and placed on ice. Twenty five microlitres of a 2X PCR buffer of the following composition were added to each tube: 20 mM Tris-HCl pH 8.3, 100 mM KCl, 3.0 mM $MgCl_2$, 4% (v/v) deionised formamide, 2 µM each oligonucleotide, 400 µM dNTPs and 2.5 units Taq polymerase (Boehringer Manneheim). Mineral oil was overlaid on the reactions which were allowed to proceed through 40 cycles of 95°C, 1 minute; 57°C, 30 seconds; and 72°C, 2 minutes. As a positive control for the PCR amplification, 1 ng of plasmid DNA was utilised. The negative control that was used was parental ES cell DNA with no G418 resistant ES DNA.

At the completion of the PCR reactions, 25 µl of each PCR reaction were loaded onto a 1% (w/v) agarose gel and electrophoresed alongside λHindIII molecular weight markers. It was found that in our hands, after 40 cycles, only the positive plasmid control DNA produced an amplified product of the expected size. Therefore, 5 µl was taken from each PCR reaction and seeded into a fresh PCR reaction containing PCR oligonucleotides that were internal (nested) to the first ones used. PCR amplification continued for a further 30 cycles as described. At the completion of the PCR reactions, 25 µl were analysed by agarose gel electrophoresis as described. Using this type of analysis we found that although a visible PCR amplified product could be seen in most lanes after the second round of amplification, the results were not consistent enough that they could be used to efficiently identify any positive homologous recombinant that might be present. This was found to be particularly true for the Type II construct which possessed the mouse Muc-1 repetitive domain as its 3' arm of homology. Therefore, at this stage in the analysis it was decided that screening G418 resistant ES colonies by Southern blotting using a flanking probe would represent a more efficient way of identifying Muc-1 homologous recombinants.

5.4 Targeted inactivation of the Muc-1 gene in E14TG2a cells: identification and analysis of targeted clones

The Muc-1 replacement vectors, pMuc-1GT Type I and Type II, were electroporated into the mouse ES cell line E14TG2a as described. Cells were at passage 20 when they were electroporated. After electroporation, cells were carefully removed from the electroporation cuvette and added to 20 ml prewarmed medium. The cuvette was washed twice with complete medium to ensure that the majority of electroporated cells were recovered. Cells were distributed over ten 9 cm plates of STO-neo feeders and allowed to recover. As a control for the electroporation and selection, an equivalent number of ES cells were electroporated, under the same conditions, in the absence of plasmid DNA. Twenty four hours after the electroporation, medium was aspirated from the plates and selective medium was added. Eight plates received ES medium supplemented with 400 µg/ml G418 and 2 µM gancyclovir, whereas two plates received ES medium supplemented with 400 µg/ml G418 alone. The negative control plates received medium with both selective agents. Selection was continued for 10 days as

described. The number of resistant colonies was estimated on each plate after 10 days of selection, and used to calculate an approximate enrichment factor acheived using G418 with gancyclovir, as opposed to using G418 only. A general observation was that colonies grown under double selection were generally much smaller than those grown in the presence of G418 only. For the Type I construct, approximately 950 double resistant colonies were obtained and the respective gancyclovir enrichment factor was found to be approximately 2.4 fold. The Type II construct gave rise to 399 double resistant colonies, and the respective gancyclovir enrichment factor for this vector was 3.4 fold.

Double resistant colonies were picked, as described, and expanded for analysis by Southern blotting. A 3' flanking probe was selected for the identification of homologous recombinants. This probe was an approximately 800 base pair PstI-EcoRI fragment located at the 3' end of the 11 kilobase pair EcoRI fragment. It was reasoned that EcoRI restriction endonuclease digestion of mouse genomic DNA, and Southern analysis utilising this probe would normally identify an 11 kilobase pair hybridising restriction fragment. However, the design of the two targeting vectors resulted in the incorporation of novel EcoRI sites into the Muc-1 gene locus. It was calculated that a Muc-1 allele that had been targeted with the Type I construct would yield an approximately 5.5 kilobase pair hybridising restriction fragment, whereas an allele targeted with the Type II construct would yield an approximately 8.0 kilobase pair fragment (Fig 5.41 and 5.42). Genomic DNA prepared from 172 Type I and > 100 Type II double resistant ES clones was digested with EcoRI and screened with the 3' flanking probe in an attempt to detect Muc-1 homologous recombinants. Positive diagnosis of homologous recombination depended upon the presence of the novel Muc-1 hybridising fragment, in addition to the endogenous 11 kilobase pair hybridising fragment. By autoradiography, one out of 172 Type I colonies, designated #32.1, was found to display the expected pattern diagnostic of a targeted replacement event at the Muc-1 gene locus (Fig 5.41). In addition, a second Type I colony, designated #23.2, displayed an aberrant pattern of 11 kilobase pairs and 8 kilobase pairs (Fig 5.43). Of the Type II resistant colonies that were screened, all were found to be indistinguishable from the parental ES cell line DNA.

A frequency of homologous recombination of 1/172 double resistant clones, or 1/413 G418 resistant clones (taking the 2.4 fold enrichment factor into consideration), was obtained with the pMuc-1GT Type I targeting vector. However, as only a single targeted clone was obtained, the accuracy of this figure was considered to be subject to a large random-sampling error. Clone #32.1 was analysed further using a variety of DNA probes to ensure that there were no additional insertions of the vector DNA at other locations within the genome. As the Muc-1 gene is flanked closely on either side by other expressed sequences, probes were used to determine that the thrombospondin-3 gene upstream of the Muc-1 gene, and the uncharacterised gene downstream of the Muc-1 gene had not been inactivated (Fig 5.41). In addition, the normal chromosome constitution of clone #32.1 was confirmed by karyotyping (Fig 5.44). G-banding analysis detected no major chromosome rearrangements (data not shown). The *in vitro* differentiation capacity was also investigated by embryoid body formation (Fig 5.44). This clone was observed to progress through all the documented stages of *in vitro* differentiation.

The enrichment factors that were achieved under the double selection regimen were disappointing and, therefore, the initial Southern blots were stripped and rehybridised with a radiolabelled probe specific for the HSV-tk gene. Greater than 25% of the double resistant clones were found to possess the expected hybridising restriction fragments characteristic of the incorporation of the HSV-tk gene into the genome (data not shown). This suggested that in at least 25% of the double resistant colonies the HSV-tk gene had been inactivated in some manner.

Clone #23.2 was extensively analysed using a variety of DNA probes. It was deduced that in this clone 3 copies of the targeting vector had concatemerised in a head-to-tail-head-to-tail-tail-to-head fashion, prior to insertion into the long-arm (5' arm of the Muc-1 gene) (Fig 5.43). It was reasoned that if the concatemerised construct had inserted by a single homologous integration event it would be expected that the diagnostic 3' flanking probe would not detect any shifted hybridising band. Therefore, in addition to the insertion event there must have been some kind of vector mediated rearrangement or deletion event resulting in the generation of an altered Muc-1 allele with an approximately 8.0 kilobase pair hybridising EcoRI fragment. This rearrangement may have

been induced by the presence of a large inverted repeat that was created by the insertion (Fig 5.43).

Aberrant targeting events similar to that identified in clone #23.2 have been described previously (Hasty, 1991a; Thomas, 1992), and several factors are thought to play a role in influencing the frequency of these type of events. It has been demonstrated that aberrant recombination events occur at highest frequency when the length of one of the arms of homology of a replacement targeting vector is reduced below 1 kilobase pair in size and the other arm of homology is disproportionately larger (Thomas, 1992). The Type I targeting vector that was utilised had a distribution of homology such that the 5' arm of homology was four times longer than the 3' arm, which was approximately 1 kilobase pair in length. If recombination frequency was equally distributed over each vector arm, the 5' arm would, therefore, have been four times as likely as the 3' arm to pair and recombine with its chromosomal target. In addition, it has been suggested that the depression of recombinational fidelity caused by a reduced arm length may be compounded if there is sequence heterology between the vector arm and its target (Thomas, 1992). The vector arms utilised in the Muc-1 targeting vectors described above were Balb/c in origin, whereas the ES cell line, E14TG2a, was 129 derived.

Clone #32.1 was microinjected into C57Bl blastocysts as described. Cells were at passages 27-30 at microinjection. All chimaeras derived from this clone had an ES contribution to coat colour of less than 30%. Sixteen males and fourteen females were derived and back-crossed against C57Bl mice to test for transmission of the agouti coat colour. None of the #32.1 chimaeric mice transmitted the agouti coat colour to their offspring (Table 5).

5.5 Re-isolation of the Muc-1 gene from 129 genomic DNA library: 129 Muc-1 targeting vector design and construction

It was reasoned that one of the limiting factors encountered in the experiments described above was the comparatively low frequency of gene targeting. Only one correctly targeted clone was obtained and this clone failed to contribute to the germline formation of the chimaeras that were derived from it. Recently, it has been demonstrated that the use of

targeting vectors derived from isogenic DNA sequences resulted in an increase in the targeting efficiency in mouse embryonic stem cells (te Riele, 1992). Therefore, in an attempt to increase the targeting efficiency of the Muc-1 gene, a cosmid library constructed from 129Sv mouse genomic DNA (Stratagene, USA) was screened utilising the mouse Muc-1 probe pMuc2TR. Approximately 5 X 10^5 colonies were screened using standard Southern hybridisation conditions as described. Greater than 200 double positive clones were obtained, five of which (Fig. 5.51A) were taken through two further rounds of colony purification to arrive at pure clones. Cosmid DNA was prepared and restriction fragment analysis revealed that all five cosmids were identical, presumably because the library had been amplified. The 11 kilobase pair EcoRI fragment containing the mouse Muc-1 gene was sub-cloned into pBluescriptSKII+ in the same orientation as the equivalent 11 kilobase pair fragment that had been previously isolated from the Balb/c genomic library. This plasmid was designated 129 Muc-1 E2. In an effort to investigate the existence of strain-specific restriction fragment length polymorphisms, both the 129 and the Balb/c-derived 11 kilobase pair EcoRI fragments in pBluescriptSKII+ were digested with a panel of 17 frequent cutting restriction endonucleases. Four out of the 17 enzymes yielded a restriction fragment length polymorphism (Fig 5.51B), indicating the presence of strain-specific point mutations. te Riele, 1992, reported that within the mouse retinoblastoma gene locus, a comparison of the sequence of both a 129 and Balb/c derived genomic clone revealed that the longest stretch of perfect homology was only 278 base pairs. To investigate strain homology within the Muc-1 gene, SacI, PstI and KpnI restriction fragments containing portions of the 129Sv Muc-1 repetitive domain were sub-cloned into pBluescript and sequenced. Sequence analysis revealed that the 129 and Balb/c-derived sequences shared greater than 600 base pairs of continuous homology within the repeat region of the gene (data not shown).

An additional way to increase targeting efficiency is to increase the absolute amount of homology within the targeting vector (Deng, 1992). A new 129 derived targeting vector was designed, 129Muc-1GT, in which 9 kilobase pairs of homology were incorporated into the construct. This was in contrast to the 5 kilobase pairs of homology that were utilised in the Balb/c targeting vectors. The 9 kilobase pairs were made up of a 5' SmaI fragment of 2 kilobase pairs and a 3' SmaI-EcoRI fragment of

approximately 7 kilobase pairs (Fig 5.21 and Fig 5.52). An *E. coli* β-galactosidase gene (Kalnins, 1983) lacking the first 7 codons, including the translation initiation site, was ligated in-frame with the mouse Muc-1 gene to create a Muc-1/LacZ fusion protein (Fig. 5.52). This fusion protein was designed to be under the control of the Muc-1 transcription and translational machinery and incorporated the promoter and first three codons of the mouse Muc-1 gene ligated in-frame, at the SmaI site in the first exon, to the eighth codon of the LacZ gene (Fig. 5.52). The LacZ gene was kindly provided by Dr. Rob Krummlauf, National Institute for Medical Research, Mill Hill, London.

In order to create the fusion, the LacZ gene (#839) was sub-cloned as an approximately 3 kilobase pair restriction fragment into the BamHI site of pBluescriptKSII+. The gene was then sub-cloned from pBluescriptKSII+ as a SmaI-EcoRI restriction fragment into pBluescriptSKII+. This cloning step resulted in the loss of the β-galactosidase stop codon and was necessary in order to place the correct cloning sites at the 5' end of the LacZ gene to allow an in-frame ligation between Muc-1 and LacZ to be made. A 2 kilobase pair SmaI restriction fragment containing the mouse Muc-1 promoter and part of the first exon was ligated into the SmaI site that had been created at the 5' end of the LacZ gene. This plasmid was designated pBS-5'Muc-1LacZ (Fig 5.52). The ligation junction was sequenced to ascertain that the splice had resulted in an in-frame fusion between Muc-1 and LacZ (Fig 5.53). The cloning strategy was designed such that unique sites for the restriction endonucleases EcoRI and HindIII would be present at the 3' end of the LacZ gene (Fig 5.52). An approximately 2 kilobase pair EcoRI-HindIII restriction fragment, containing the pgkneo gene, was cloned directly into these sites to create the intermediate plasmid pBS-5'Muc-1LacZneo. To restore the LacZ stop codon and also to place the polyadenylation signal of SV40 at the 3' end of the LacZ gene, an approximately 1 kilobase pair EcoRI fragment was sub-cloned from the plasmid pPGK (E/T)LacZ (kindly provided by Dr. Mike McBurney, University of Ottawa) into the EcoRI site between the LacZ and pgkneo genes (Fig 5.52). This plasmid was designated pBS-5'Muc-1LacZSVpAneo. All ligation junctions were confirmed throughout by double-stranded DNA sequencing. To test the expression of the Muc-1/LacZ fusion protein, large-scale CsCl preparation DNA was prepared and transfected into three mammalian cell lines using calcium-phosphate precipitation (Graham, 1973). Briefly, 5 μg

supercoiled plasmid DNA was transfected into the hamster pancreatic adenocarcinoma cell line, HP-1, the human pancreatic carcinoma cell line, ZR-75-1, and the human fibrosarcoma cell line, HT1080, as described. Both HP-1 and ZR-75-1 have been demonstrated to express high levels of the Muc-1 protein (Pemberton, 1992; Kovarik, 1993), whereas HT1080 cells have been shown not to express MUC1 (Kovarik, 1993). Twenty four to forty eight hours after transfection, cells were washed with PBS and assayed for β-galactosidase activity as described. Staining revealed LacZ positive cells only in transfected cultures of ZR-75-1 and HP-1 (Fig 5.53) indicating that the Muc-1/LacZ fusion protein was being correctly expressed under the control of the Muc-1 regulatory sequences.

The final 129 Muc-1 targeting vector was created as follows (Fig. 5.52): the plasmid 129 Muc-1 E2 was digested with the blunt-ended restriction endonuclease SmaI and a 10.0 kilobase pair fragment was gel-purified by DEAE paper purification. The 8.0 kilobase pair 5'Muc-1-LacZ-SVpA-pgkneo cassette that had been previously created was removed from the vector as a single NotI-HindIII restriction fragment. Three prime overhanging ends were filled in using Klenow DNA polymerase and the blunted fragment was gel-purified. The 10 kilobase and 8 kilobase pair fragments were ligated overnight at 16°C then transformed into competent cells. As the 8.0 kilobase pair fragment had been blunt-ended, two alternative orientations were possible for the products of the ligation. Therefore, in order to select for plasmids in which the insert had been ligated in the correct orientation, mini-prep plasmid DNA was restriction digested separately with EcoRI and HindIII. Plasmids that were diagnosed as possessing the insert in the correct orientation were sequenced to confirm all ligation junctions. The final 129 Muc-1/LacZ targeting plasmid was designated 129 Muc-1GT.

5.6 Targeted inactivation of the mouse Muc-1 gene in E14TG2a and GK129 cells with an isogenic DNA construct

129 Muc-1GT plasmid DNA was linearised with NotI prior to electroporation into the ES cell line E14TG2a, as described. Cells were electroporated at passage 20. Cells were plated on 9 cm plates and selection proceeded 24 hours after the electroporation. Colonies were selected in ES medium plus 400 μg/ml G418. Due to the disappointing

enrichment factors achieved using the HSV-tk gene, this gene was omitted from the 129 Muc-1GT construct. Ten days after the start of selection, colonies were picked as described. Although 4 X 10^7 ES cells were electroporated, only 67 G418 resistant colonies were obtained in this experiment. The reason for such a large discrepancy between this experiment and the previous in the respective number of resistant ES colonies was unknown. Possible reasons that were suggested were that the position of the neomycin cassette within the 129 construct and the presence of the LacZSVpA gene 5' of the neo gene could influence its expression and, therefore, influence the number of resistant colonies obtained. Alternatively, the ES cells may not have been in the optimal growth phase at the time of electroporation in the second experiment.

Forty seven G418 resistant clones were analysed by EcoRI restriction endonuclease digestion and Southern analysis with a 5' flanking probe. To detect homologous recombinants, a 2 kilobase pair EcoRI-SmaI restriction fragment was utilised. This probe would detect an 11 kilobase pair endogenous EcoRI restriction fragment, but would hybridise to an additional novel 7 kilobase pair restriction fragment in clones that had been correctly targeted (Fig 5.61). Of the 47 clones screened, 4 displayed the expected hybridisation pattern (Fig 5.62). A frequency of 4/47 G418 resistant colonies represented a 1/12 replacement frequency, a 34 fold enhancement in targeted replacement frequency over the Balb/c Type I construct. This enhancement in targeting efficiency was thought to be due to both the use of isogenic DNA and an increase in the absolute amount of homology in the targeting vector. In addition to the clones that displayed the expected pattern for a targeted replacement event, 4 clones displayed aberrant hybridising bands (Fig 5.63). Despite the fact that the minimum arm length in the 129 derived targeting vector was 2 kilobase pairs in size, these clones were deduced to be the result of insertion of at least 10 concatemerised copies of the targeting vector DNA into the Muc-1 locus. When both the correctly targeted and the aberrantly targeted clones were taken into consideration, a total of 8/47, this implied that the 129 Muc-1GT construct targeted the Muc-1 locus at a frequency of 17%.

The ES cell line E14TG2a was at passage 20 in the two electroporations described above and was microinjected at passages 27-30. Clone #32.1 appeared to be undifferentiated in culture and went through

all the normal developmental stages when induced to differentiate. However, the chimaeras that were derived from this clone failed to transmit the agouti coat colour to their offspring. It is generally accepted that as the passage number of an ES cell line increases, the relative number of cells that may be capable of colonising the germline of a developing embryo decreases. For this reason we obtained a second ES cell line, GK129 (kindly provided by Dr. Graham Kay, Clinical Research Centre, Medical Research Council, Harrow, UK) (Philpott, 1992), at passage 7. The 129 Muc-1GT vector was electroporated into this cell line at passage 10, utilising conditions recommended by Dr. Graham Kay. G418 resistant colonies were selected for in complete ES medium supplemented with 400 μg/ml G418. After 10 days of selection, resistant colonies were picked. In contrast to the number of G418 resistant colonies obtained with this targeting vector in E14TG2a cells, greater than 2×10^3 G418 resistant GK129 cells were obtained.

G418 resistant GK129 colonies were picked, expanded and screened for the presence of the expected diagnostic hybridising restriction fragment, as described above. The frequency of targeted replacement of the Muc-1 gene in this experiment was found to be approximately 1 in 18 G418 resistant colonies (Fig 5.64). In addition, there were numerous colonies that appeared to be targeted aberrantly, as described above.

5.7 Analysis of targeted ES cell clones

All the targeted ES clones from both the E14TG2a and GK129 cell lines were analysed extensively prior to their microinjection into mouse blastocysts to generate chimaeras. Clones were analysed by Southern blotting with a variety of internal probes, to ensure that no additional insertions of the targeting vector or rearrangements of the flanking genes had occurred (Fig 5.62). It was found that, in clones that demonstrated the expected pattern for a targeted replacement event, this event was the only integration of the targeting vector DNA that had occurred. In addition to the analysis of targeted clones by Southern blotting, the expression pattern of the Muc-1/LacZ fusion protein was analysed through *in vitro* and *in vivo* differentiation. Cells were allowed to differentiate in suspension culture to form embryoid bodies. Embryoid bodies were also replated onto gelatinised tissue culture plates, allowed to attach and differentiate further. At each stage of development, embryoid bodies or

the cells that had been allowed to attach and differentiate were stained for the presence of Muc-1 driven LacZ expression. It was found that cells present within both simple and cystic embryoid bodies stained positive for LacZ. Cells within the outer endodermal ring of simple embryoid bodies did not stain for LacZ, whereas there appeared to be a high level of heterogeneity of staining of cells within the core of the embryoid bodies. In an attempt to identify the type of cells that were staining positive for LacZ protein, embryoid bodies of both the simple and cystic type were replated onto gelatinised dishes and allowed to attach and spread. Cells were subsequently stained at several time points, but it was found that when cells were allowed to attach and spread on plastic, none of the cells that were present, including a large number of epithelial cells, were observed to stain for LacZ protein. Parry, 1992, demonstrated that mouse mammary carcinoma cell lines expressed Muc-1 at highest levels when grown in the presence of an extracellular matrix. Perhaps, therefore, the transfer and growth of differentiating ES cells on tissue-culture plastic resulted in the down-modulation or loss of expression of the Muc-1 driven LacZ gene.

Sub-cutaneous injection of ES cells into athymic nude mice results in the relatively rapid generation of well differentiated teratocarcinomas containing a wide range of tissue types. In an attempt to analyse the expression pattern of the Muc-1/LacZ fusion protein *in vivo*, we injected several of the targeted clones into nude mice as described, along with the parental ES cell line as a negative control. All clones and the parental cell lines formed well differentiated teratocarcinomas within two to three weeks. Mice were sacrificed and tumours were taken for histological analysis as described. Half of each tumour was fixed and stained overnight to detect the presence of LacZ, and the other half was fixed in methacarn to be used for immunohistochemistry. We found that the fixation protocol used in the preparation of tissues for LacZ staining destroyed the epitope of the CT1 polyclonal antiserum used to detect Muc-1 protein. In addition, fixation of tumours in methacarn (60% (v/v) methanol, 30% (v/v) chloroform, 10% (v/v) acetic acid) was found to destroy LacZ activity. Sections were prepared from all tumours. LacZ stained tumour sections were counterstained with eosin only in order that the blue stain, characteristic of LacZ activity, would be observable. In addition, for an investigation of morphology, some sections were stained with Gomori's trichrome (Fig 5.71). All sections were found to contain

diverse tissue types including branching epithelial lumens, cartilage and bone, smooth and striated muscle, and blood islands (Fig 5.71). LacZ positive staining was observed only in cells within lumenal structures (Fig 5.71), suggesting that the Muc-1 promoter and fusion was driving the tissue-specific expression of the LacZ gene. This was expected as, through gene targeting, the Muc-1 promoter, driving expression of the fusion protein, was placed in its endogenous site within the genome. Sections were also analysed by immunohistochemistry, to detect the presence of the mouse Muc-1 protein, with the polyclonal antiserum CT1 (Pemberton, 1992). CT1 was found to stain the apical surface of epithelial lumens in tumours derived from both the parental cell lines and the targeted cell lines (Fig 5.71). There was no observable difference in the intensity of the staining pattern between Muc-1 targeted and parentally derived tumours (data not shown).

Three of the Muc-1/LacZ targeted E14TG2a clones were injected into C57Bl/6 blastocysts and chimaeras were generated as described. All three clones gave rise to chimaeras, the majority of which had less than 20% ES contribution to coat colour (Table 5). All chimaeric animals were back-crossed against C57Bl/6 mice to screen for germline transmission. However, all three E14TG2a derived targeted ES clones failed to transmit the ES coat colour to their offspring (Table 5).

Three of the Muc-1/lacZ targeted GK129 clones were injected into C57Bl/6 blastocysts, as described. Clone #56 consistently gave rise to chimaeras that showed greater than 75% ES contribution to coat colour, with a significant proportion showing greater than 95% ES coat colour contribution and the majority appearing to have 100% ES-derived coat colour (Fig 5.72). In addition, of eight mice born from an injection of clone #22, three showed greater than 95% ES coat colour contribution. Clone #132 also gave rise to chimaeric mice with a significant ES contribution to coat colour (Table 5). All chimaeric animals were back-crossed against C57Bl/6 mice to screen for germline transmission of the ES-derived coat colour.

It appeared that the use of a second ES cell line, obtained at an earlier passage number, greatly increased both the frequency of chimaeras obtained and the level of ES contribution to coat colour of each chimaera (Table 5). In turn, a higher frequency of chimaera formation and a more

extensive contribution of the ES cell to the development of the mouse would be expected to increase the probablility of obtaining a mouse in which the germline had been colonised by the ES cells.

5.8 Germline transmission analysis of a specific Muc-1/LacZ mutation

All GK129/Muc-1 chimaeras were back-crossed with C57Bl/6 mice to screen for germline transmission. Males were paired with three female C57Bl/6 mice 4-6 weeks old, whereas female chimaeras were paired with a single male C57Bl/6 mouse. Germline transmission was assessed by the presence of agouti offspring. Several of the male and female chimaeras successfully transmitted the agouti coat colour to their offspring (Table 5). As the mutation that had been created at the Muc-1 locus was only present in one of the two alleles, on average 50% of agouti offspring were expected to be heterozygous for the Muc-1/LacZ mutation. The remaining 50% would be homozygous wild-type.

In order to screen the agouti offspring for the presence or absence of the disrupted Muc-1 allele, a PCR-based approach was adopted. Oligonucleotides were synthesised that would be specific for the mouse Muc-1 gene and the *E. coli* β-galactosidase gene. Two Muc-1 oligos were synthesised and these had the sequence, 5'-ACC TCA CAC ACG GAG CGC CAG-3' and 5'- CAG CAG GAA GAA AGG AGC CCG- 3', respectively. These oligonucleotides corresponded to nucleotides +3 to +23 and +104 to +84 (on the antisense strand) (Fig 3.61). These oligos were used in conjunction with an oligo specific to the β-gal gene, 5'- TTC TGG TGC CGG AAA CCA GGC-3'. To screen for the presence of the disrupted Muc-1 allele, tail DNA was made from agouti offspring as described, and was subjected to PCR amplification with the oligos described above. The oligos were designed such that the two Muc-1 oligos together would specifically amplify a fragment of approximately 120 base pairs from the wild-type allele. Inclusion of one of the β-gal oligonucleotides in the PCR reaction enabled the detection of the targeted allele. PCR amplification of genomic DNA from mice heterozygous for the Muc-1 mutation would result in the specific amplification of two different DNA fragments. The two Muc-1 oligos would amplify an 120 base pair fragment from the wild-type allele, and the 5' Muc-1 oligo, in combination with one of the β-gal oligos would amplify an approximately 250 base pair fragment from the targeted allele. PCR amplification was carried out on between 500 ng and

1 μg total genomic DNA in a total volume of 100 μl using standard buffer conditions. The cycles utilised were 40 cycles of 94°C, 1 minute; 60°C, 1 minute; 72°C 1 minute. The final cycle was followed by a five minute extend at 72°C in order to ensure that all products were fully extended. Twenty microlitres of each PCR reaction were analysed by agarose gel electrophoresis through a 2.5% agarose gel. Products were visualised by ethidium bromide staining (Fig 5.8).

5.9 Conclusions

In this chapter, experiments designed to specifically mutate the mouse Muc-1 gene in mouse embryonic stem cells have been described. The gene was successfully targeted through two different replacement targeting vectors, one of which was designed to incorporate a Muc-1 driven Muc-1/LacZ fusion protein sequence into the endogenous Muc-1 locus.

Initially, a replacement vector constructed from a Balb/c derived Muc-1 clone was utilised to target the Muc-1 gene in the ES cell line E14TG2a. This vector, pMuc-1GT TypeI, incorporated approximately 5 kilobase pairs of homology and contained the selectable gene neo flanked by the mouse phosphoglycerate kinase-1 (Pgk-1) gene promoter and polyadenylation signals. At the 5' end of the construct an HSV-tk gene was included, such that the Positive-Negative-Selection (PNS) system, described by Mansour, 1988, could be utilised (Fig 5.22). After electroporation and double selection with G418 and gancyclovir, one correctly targeted clone, designated #32.1, was obtained (Fig 5.41). The targeting frequency obtained using this vector was calculated to be 1 in 413 G418-resistant colonies. In addition to the single correctly targeted clone, one clone, #23.2, was observed to have the targeting vector incorporated into the correct site but through an insertion event rather than a replacement event (Fig 5.43). Through Southern analysis it became apparent that in this clone 3 copies of the targeting vector had concatemerised prior to insertion into the 5' region of the Muc-1 gene by a modified single recombination event. Previously, it has been shown that the relative distribution of homology within a targeting vector can influence the frequency at which this type of aberrant targeting event may occur (Hasty, 1991a). Subsequently, it has been demonstrated that when the length of one of the arms of homology falls below 1 kilobase

Alternatively, the concentration of G418 used in our experiments appeared to be significantly higher than that reportedly used by other laboratories and this may also have contributed to selection of clones carrying more than one copy of the neo gene.

pair, the frequency of aberrant targeting events increases (Thomas, 1992). The distribution of homology in the Muc-1 targeting vector, pMuc-1GT TypeI, was such that the 5' arm was approximately 4.0 kilobase pairs in length and the 3' arm was approximately 1.0 kilobase pair in length. It is possible, therefore, that the unequal distribution of homology present in the targeting vector may have been responsible for the aberrant targeting event demonstrated in clone #23.2.

Clone #32.1 was found to have the correct chromosome constitution through karytotype analysis, and appeared to proceed normally through all the described *in vitro* differentiation steps (Fig 5.44). Microinjection into C57Bl/6 blastocysts yielded chimaeras with a relatively low percentage of ES contribution to coat colour, as judged by the extent of agouti pigment in the coat (Table 5). All mice were test-bred to determine whether or not the ES cells had contributed to the formation of the germline. Unfortunately, none of the mice were found to transmit the ES coat colour to their offspring.

It has recently been shown that the use of isogenic DNA derived constructs greatly increased the targeting frequency in mouse embryonic stem cells (te Riele, 1992). In an attempt to target the mouse Muc-1 locus at a higher frequency, the mouse Muc-1 gene was re-isolated from a 129Sv cosmid library (Fig 5.51). A second replacement targeting vector, designated 129Muc-1GT, was designed and constructed (Fig 5.52). This construct incorporated approximately 9 kilobase pairs of homology, in contrast to the Balb/c construct that incorporated only 5 kilobase pairs. In addition, the *E. coli* β-galactosidase gene was included in the construct, ligated in-frame with the Muc-1 gene (Fig 5.53). The enrichment factors previously obtained utilising the negatively selectable gene, HSV-tk, were disappointingly low in the initial experiments; greater than 25% of the double resistant ES clones retained the HSV-tk gene. For this reason, the HSV-tk gene was omitted from the 129-derived targeting vector.

Electroporation of the 129Muc-1GT vector into the ES cell line E14TG2a and subsequent Southern analysis indicated that this vector correctly targeted the Muc-1 gene at a frequency of 1 in 12 G418-resistant colonies (Fig 5.61 and 5.62). This was calculated to be an approximately 34-fold increase in targeting frequency over the frequency obtained with the pMuc-1GT TypeI replacement vector. Presumably, both the increase in

total homology and the use of isogenic DNA contributed to this dramatic increase in targeting efficiency (Deng, 1992).

Targeted clones were analysed extensively by Southern blotting with a variety of internal and flanking probes in order to confirm that no additional rearrangements of flanking DNA (containing the 5' thrombospondin-3 gene and the 3' uncharacterised gene) had occurred, and that the vector had not integrated elsewhere in the genome. In every instance, clones that had been initially identified as being correctly targeted were found to possess only a single copy of the targeting vector, correctly integrated into the Muc-1 locus.

To investigate the expression pattern of the Muc-1/LacZ fusion, targeted ES cells were injected sub-cutaneously into athymic nude mice. All cells gave rise to well-differentiated teratocarcinomas comprised of a large number of tissue types, including epithelial lumens. Staining revealed that LacZ positive areas were confined to cells present within these epithelial structures (Fig 5.71). Sections were also analysed immunohistochemically with the polyclonal antiserum, CT1, in order to detect the mucin protein. Muc-1 protein was detected on the apical surfaces of cells present within the majority of lumenal structures.

As described previously, all clones were microinjected into C57Bl/6 blastocysts and chimaeras were derived. Again, chimaeras derived from targeted E14TG2a clones were found to have a relatively low ES contribution to coat colour and were subsequently found not to transmit the ES coat colour to their offspring (Table 5). It appeared, therefore, that the ES cell line, E14TG2a, that had been acquired was unable to give rise to germline chimaeras. Indeed, microinjection of the parental cell line at passages 18-20 yielded chimaeras with an ES coat colour contribution of up to 40-50%, yet none of these mice transmitted the agouti coat colour to their offspring. For this reason, a second ES cell line, GK129 (Philpott, 1992), was acquired at passage 7. This cell line had been demonstrated to give rise to germline chimaeras with a high percentage of ES-derived coat colour (Philpott, 1992). GK129 cells were targeted with the 129Muc-1GT vector and targeted clones were obtained at an approximate frequency of 1 in 18 G418-resistant colonies (Table 5). Microinjection of these clones into C57Bl/6 blastocysts yielded chimaeras with as much as 100% ES coat colour contribution (Table 5 and Fig 5.72).

Upon back-crossing these mice with C57Bl/6 mice, it was found that several of the chimaeras, including males and females, transmitted the agouti coat colour indicative of germline transmission to their offspring. PCR analysis of tail DNA from the agouti offspring indicated that the expected 50% were heterozygous for the disrupted Muc-1 allele (Fig 5.8). In order to obtain mice deficient in Muc-1 protein, mice identified as being heterozygous for the mutation will be crossed. These mice will be fundamental to the functional analysis of the Muc-1 protein during embryonic development, adult development of, for instance, the mammary gland, and during tumorigenesis.

It cannot be specified at this time whether or not mice deficient in Muc-1 protein will be viable. The fact that the Muc-1 protein is present early in epithelial organogenesis suggests that it may be playing a role in this process. However, if a lethal phenotype is expected then mice would presumably die soon after birth, due to for instance lung and/or digestion related problems. The use of the Muc-1/LacZ fusion will help greatly in the histochemical analysis of such mice. In addition, utilising the fact that the Muc-1 protein is made up of three major domains, external repeat, membrane-spanning, and cytoplasmic, the effect of replacing the various domains and/or combinations of them as transgenes in the homozygous mice can be analysed. In addition, this approach, in combination with site-directed mutagenesis of specific residues within the cytoplasmic tail, will allow questions to be addressed regarding the function of this region of the protein.

Alternatively, if homozygous mice are observed to be healthy after birth a large number of functional experiments have been designed to ask specific questions regarding the possible role of the protein in the various tissues in which it is expressed. Mice could be challenged, for instance, with pathogenic bacteria species to investigate the possible role of Muc-1 in the protection from bacterial infection. Muc-1 deficient mice could also be crossed with mice deficient for the cystic fibrosis transmembrane conductance regulator (CFTR) protein (Snouwaert, 1992). The major phenotype in CFTR patients appears to be related to mucous production and, therefore, Muc-1 deficient mice could allow this aspect of cystic fibrosis to be addressed. In addition, crossing Muc-1 deficient mice with the various 'Onco-mice' that have been shown to develop tumours

at high frequency will permit an investigation of the role of Muc-1 during tumorigenesis.

The actual biological function of the Muc-1 protein in both the normal tissues and tumours in which it is expressed has remained unclear. The creation of mice deficient in this protein is expected to be crucial in defining the potential role of Muc-1 during a wide range of biological processes.

Figure 5.21 Structure of the mouse Muc-1 gene locus. A detailed restriction map for the Muc-1 gene locus was obtained through DNA sequencing and extensive restriction digestion analysis of plasmid DNA. Open boxes represent the exons of the adjacent thrombospondin-3 gene (Vos, 1992); filled boxes represent the seven Muc-1 exons; arrow indicates the approximate position of the start site for the uncharacterised gene that is located downstream of Muc-1. E, EcoRI; B, BamHI; S, SmaI; H, HindIII.

Structure of the mouse Muc-1 gene locus

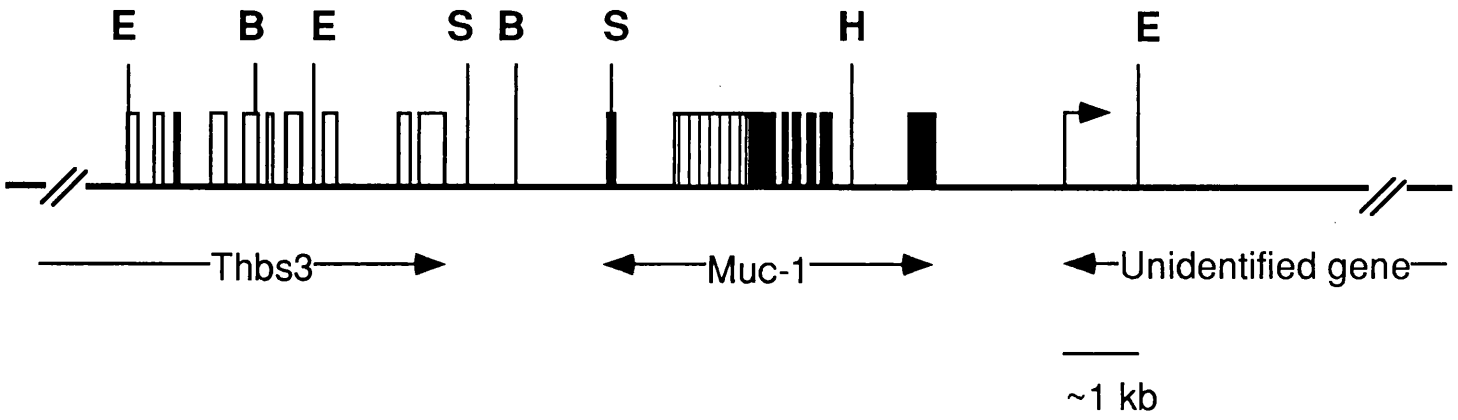
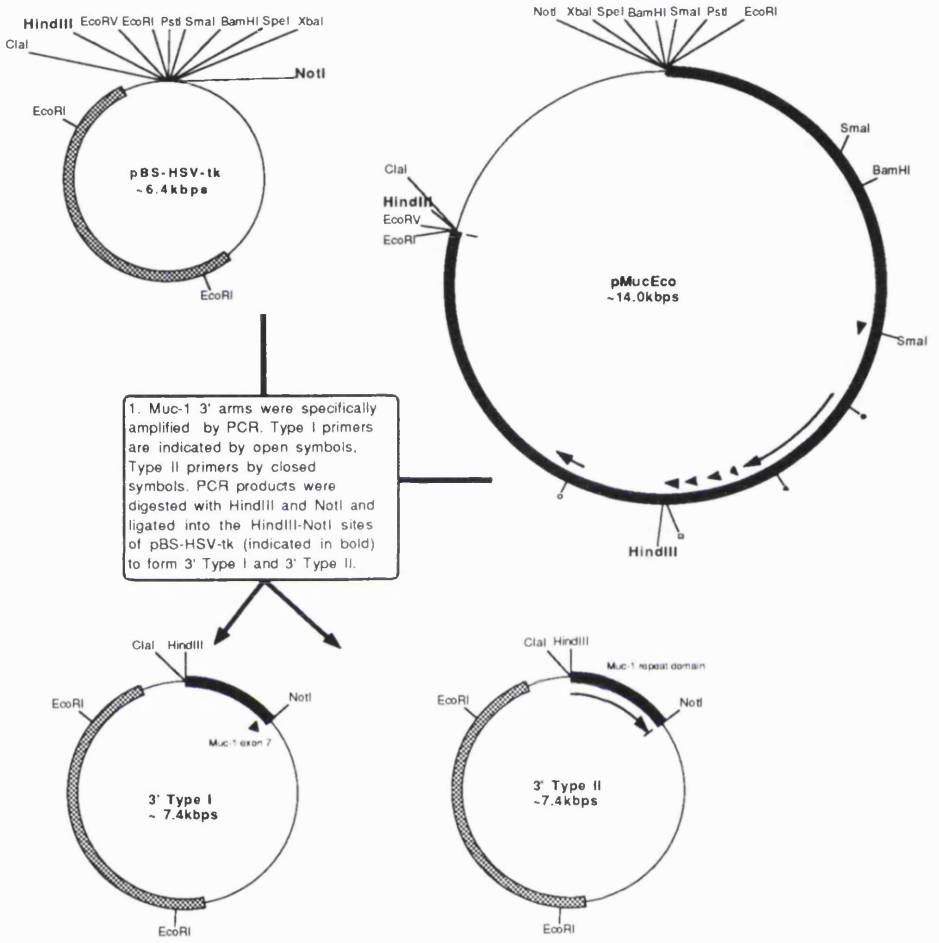
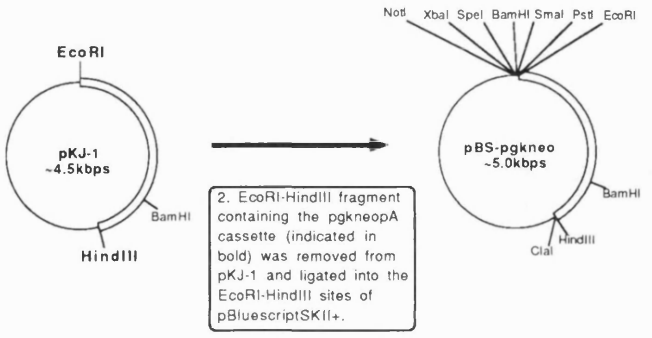
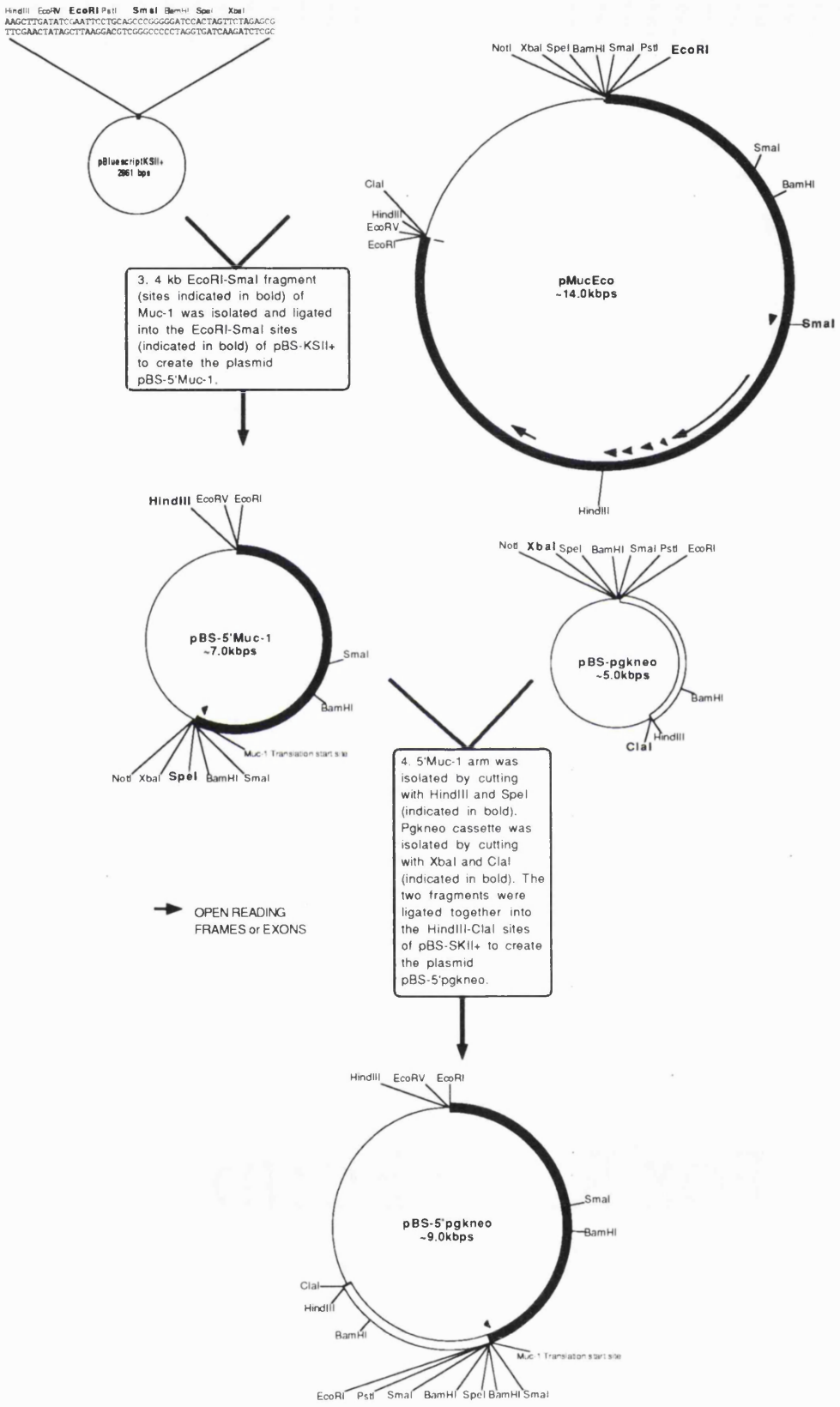


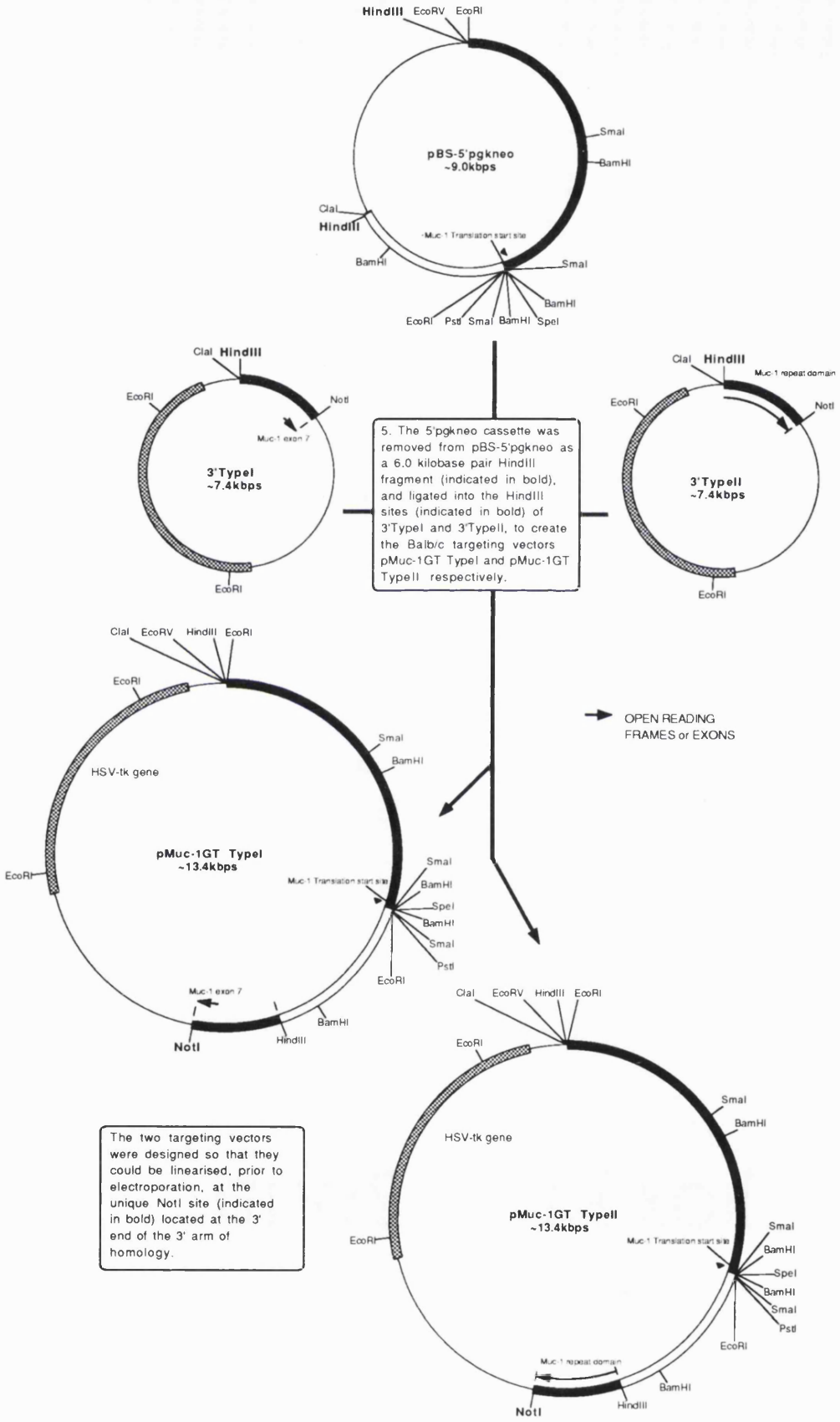
Figure 5.22 Cloning strategy for the construction of the Balb/c Muc-1 targeting vectors pMuc-1GT Type I and Type II. All vectors were drawn to scale, 1200 base pairs/cm. Arrow heads represent open reading frames and/or exons of the Muc-1 gene. Important restriction endonuclease sites used in the construction of the plasmids are highlighted in bold type.



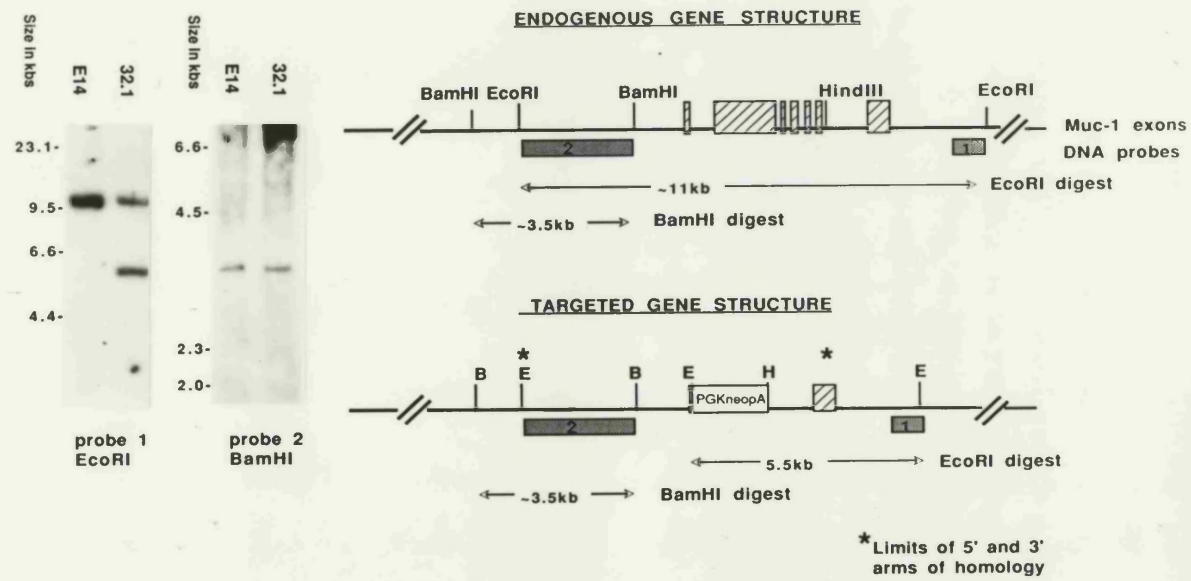


→ OPEN READING FRAMES or EXONS


HindIII EcoRV EcoRI PstI SmaI BamHI SpeI XbaI
AAGCTTGATATCGAATTCCTGCAGCCCGGGGGATCCACTAGTTCTAGAGCG
TTCGAACTATAGCTTAAGGACGTCGGGCCCCCTAGGTGATCAAGATCTCGC
pBluescriptKSII+
2961 bps
NotI XbaI SpeI BamHI SmaI PstI EcoRI
SmaI
BamHI
ClaI
HindIII
EcoRV
EcoRI
pMucEco
~14.0kbps
SmaI
HindIII
3. 4 kb EcoRI-SmaI fragment (sites indicated in bold) of Muc-1 was isolated and ligated into the EcoRI-SmaI sites (indicated in bold) of pBS-KSII+ to create the plasmid pBS-5'Muc-1.
HindIII EcoRV EcoRI
pBS-5'Muc-1
~7.0kbps
SmaI
BamHI
Muc-1 Translation start site
NotI XbaI SpeI BamHI SmaI
NotI XbaI SpeI BamHI SmaI PstI EcoRI
pBS-pgkneo
~5.0kbps
BamHI
ClaI
HindIII
4. 5'Muc-1 arm was isolated by cutting with HindIII and SpeI (indicated in bold). Pgkneo cassette was isolated by cutting with XbaI and ClaI (indicated in bold). The two fragments were ligated together into the HindIII-ClaI sites of pBS-SKII+ to create the plasmid pBS-5'pgkneo.
OPEN READING FRAMES or EXONS
HindIII EcoRV EcoRI
SmaI
BamHI
pBS-5'pgkneo
~9.0kbps
ClaI
HindIII
BamHI
Muc-1 Translation start site
EcoRI PstI SmaI BamHI SpeI BamHI SmaI

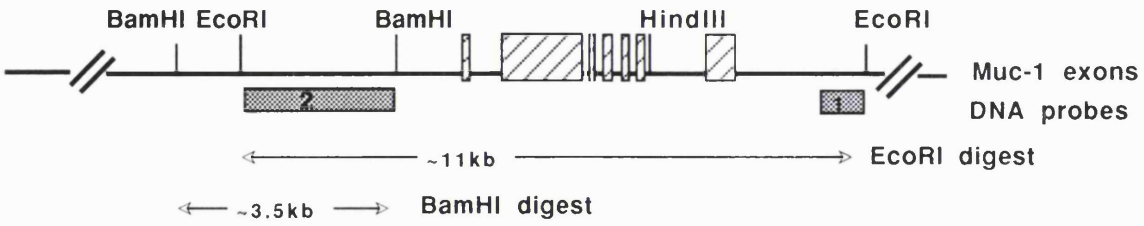

HindIII EcoRV EcoRI
SmaI
BamHI
pBS-5'pgkneo
~9.0kbps
ClaI
HindIII
Muc-1 Translation start site
BamHI
SmaI
BamHI
EcoRI PstI SmaI BamHI SpeI
ClaI HindIII
EcoRI
NotI
Muc-1 exon 7
3'TypeI
~7.4kbps
EcoRI
5. The 5'pgkneo cassette was removed from pBS-5'pgkneo as a 6.0 kilobase pair HindIII fragment (indicated in bold), and ligated into the HindIII sites (indicated in bold) of 3'TypeI and 3'TypeII, to create the Balb/c targeting vectors pMuc-1GT TypeI and pMuc-1GT TypeII respectively.
ClaI HindIII
Muc-1 repeat domain
EcoRI
NotI
3'TypeII
~7.4kbps
EcoRI
ClaI EcoRV HindIII EcoRI
EcoRI
SmaI
BamHI
HSV-tk gene
pMuc-1GT TypeI
~13.4kbps
EcoRI
SmaI
BamHI
Muc-1 Translation start site
SpeI
BamHI
SmaI
PstI
EcoRI
Muc-1 exon 7
BamHI
NotI
HindIII
OPEN READING FRAMES or EXONS
The two targeting vectors were designed so that they could be linearised, prior to electroporation, at the unique NotI site (indicated in bold) located at the 3' end of the 3' arm of homology.
ClaI EcoRV HindIII EcoRI
EcoRI
SmaI
BamHI
HSV-tk gene
pMuc-1GT TypeII
~13.4kbps
EcoRI
SmaI
BamHI
Muc-1 Translation start site
SpeI
BamHI
SmaI
PstI
EcoRI
Muc-1 repeat domain
BamHI
NotI
HindIII

5.41 Targeted inactivation of the Muc-1 gene with the replacement vector, pMuc-1GT TypeI. The mouse Muc-1 gene is contained within an 11 kilobase pair EcoRI restriction fragment and is made up of six introns and seven exons (hatched boxes). Probe 1, an 800 base pair PstI-EcoRI restriction fragment of the 11 kilobase pair EcoRI fragment, detected an 11 kilobase EcoRI fragment in mouse genomic DNA. Targeted replacement of the endogenous mouse gene with the targeting vector, pMuc-1GT Type I resulted in the deletion of exons 2-6 and the insertion of the pgkneopA cassette into exon 1, immediately downstream of the translation initiation site. In addition, insertion of the pgkneopA cassette resulted in the incorporation of a novel EcoRI site into the locus such that Probe 1 would now hybridise to an approximately 5.5 kilobase pair EcoRI fragment. As probe 1 flanked the arms of homology utilised in the targeting vector (see asterisks), this probe would only detect a 5.5 kilobase EcoRI fragment if the construct had recombined correctly, through a double-crossover event, into the endogenous Muc-1 locus. A targeted clone would be diagnosed as being correctly targeted if probe 1 hybridised to a 5.5 kilobase pair EcoRI fragment in addition to the 11 kilobase pair EcoRI fragment; clone #32.1 displayed this pattern, the result of targeted disruption of one Muc-1 allele. A second probe, probe 2, was utilised to determine whether or not there had been any rearrangements at the 5' end of the construct, and/or whether there had been any additional insertions of the vector DNA at other locations within the genome; BamHI restriction digest of mouse genomic DNA and hybridisation to this probe yielded a 3.5 kilobase pair hybridising fragment. This fragment was observed in both the parental E14 genomic DNA and the targeted clone #32.1 genomic DNA. Approximately 15 μg total genomic DNA were digested overnight at 37°C with the appropriate restriction endonuclease. DNA was size-fractionated through a 0.7% (w/v) agarose gel overnight at 1 V/cm then transferred onto nylon membrane as described. The membrane was hybridised to the respective probe and stringency washes were carried out using standard Southern conditions. E, EcoRI; B, BamHI; H, HindIII. * indicates the 5' and 3' ends of the two respective arms of homology.

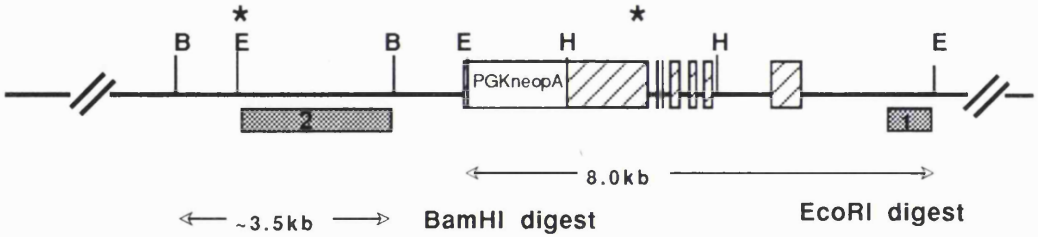

Size in kbs
E14
32-1
23.1-
9.5-
6.6-
4.4-
probe 1
EcoRI
Size in kbs
E14
32-1
6.6-
4.5-
2.3-
2.0-
probe 2
BamHI
ENDOGENOUS GENE STRUCTURE
BamHI EcoRI
BamHI
HindIII
EcoRI
Muc-1 exons
DNA probes
2
1
~11kb
EcoRI digest
~3.5kb
BamHI digest
TARGETED GENE STRUCTURE
B
E
B
E
H
E
PGKneopA
2
1
5.5kb
EcoRI digest
~3.5kb
BamHI digest
*Limits of 5' and 3' arms of homology

5.42 Predicted structure of the Muc-1 locus after targeted inactivation with the replacement vector, pMuc-1GT TypeII. The Type II targeting vector contained the Muc-1 repeat domain as its 3' arm of homology. Targeted replacement of the endogenous Muc-1 gene by this vector was predicted to result in the disruption of the gene by the insertion of the pgkneopA cassette. As described previously, insertion of the pgkneopA cassette into the gene incorporated a novel recognition site for the restriction endonuclease EcoRI. The incorporation of the novel EcoRI site into the gene would result in a shift in the size of the restriction fragment hybridising to probe 1. The normal restriction fragment was 11 kilobase pairs, whereas the additional fragment present in cells correctly targeted by the Type II vector was predicted to be 8 kilobase pairs. An additional probe, probe 2, was utilised in order to ensure that there had been no rearrangements at the target site and/or no insertions of the vector DNA at ectopic sites within the genome. A screen of greater than 100 Type II double resistant colonies failed to detect any clone that displayed anything other than the normal 11 kilobase pair hybridising fragment. E, EcoRI; B, BamHI; H, HindIII. * indicates the 5' and 3' ends of the two respective arms of homology.

ENDOGENOUS GENE STRUCTURE



TYPEII TARGETED GENE STRUCTURE



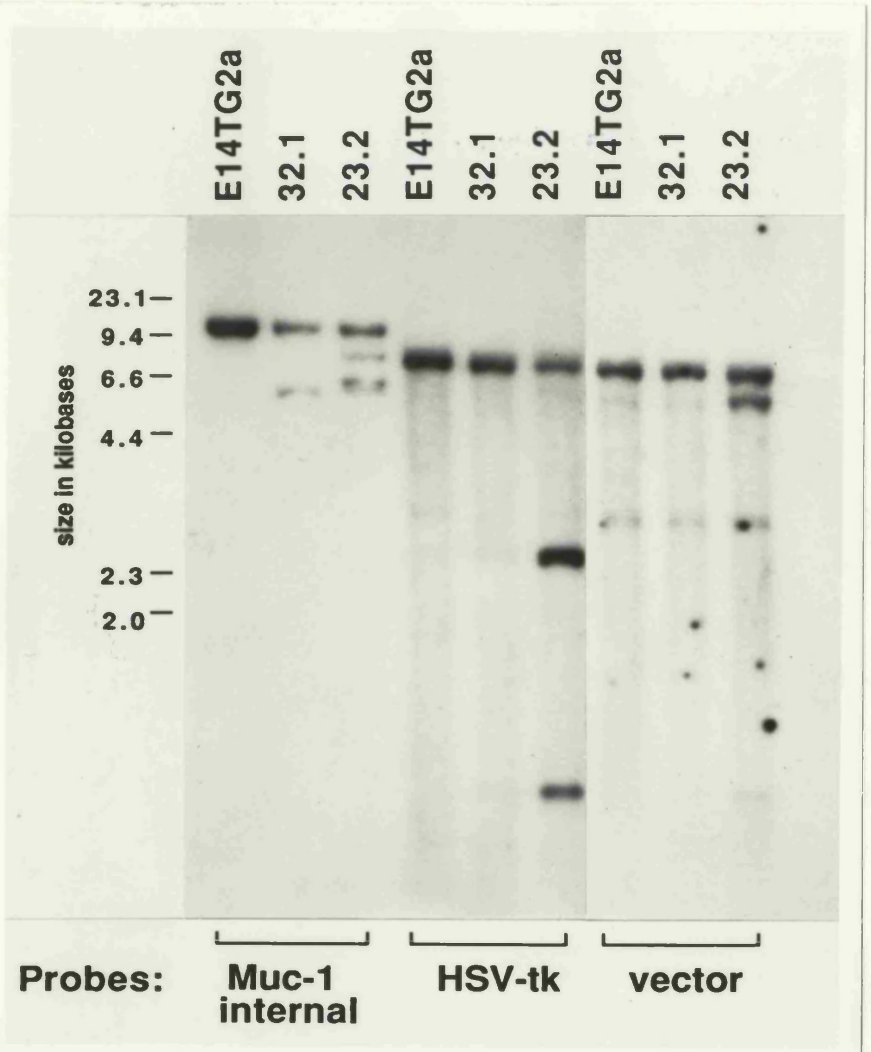
*Limits of 5' and 3' arms of homology

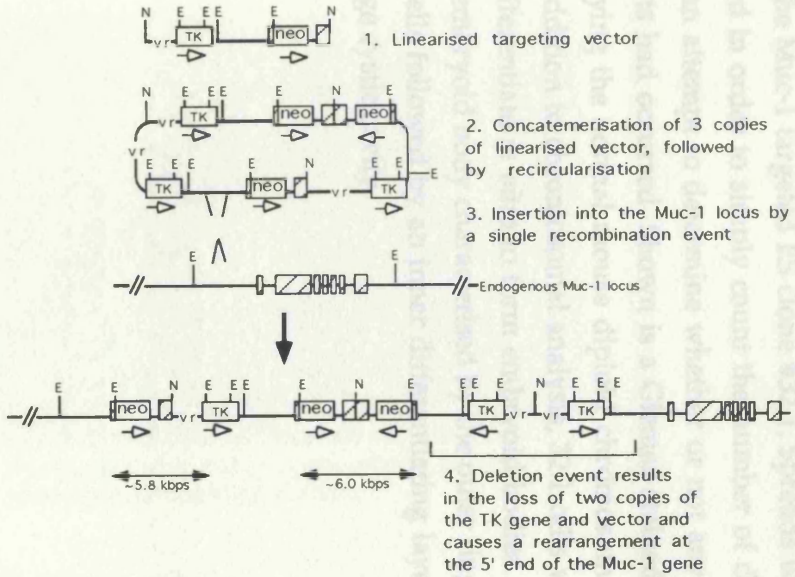
5.43 Southern analysis of an aberrantly targeted ES clone, #23.2.
Approximately 15 μg total genomic DNA from parental (E14TG2a), correctly targeted clone #32.1, and aberrantly targeted clone #23.2, were digested overnight with EcoRI, size-fractionated through a 0.8% (w/v) agarose gel, transferred to nylon membrane and hybridised with the following probes: a) Muc-1 internal (3' cDNA clone pMuc10); this probe was expected to hybridise to the endogenous Muc-1 sequence in addition to the vector 3' arm sequence. b) HSV-tk; the entire HSV-tk cassette was removed from the plasmid as two EcoRI fragments of approximately 2.5 and 0.9 kilobase pairs, labelled and utilised together as the HSV-tk probe; this probe was predicted to hybridise to specific restriction fragments of 2.5 and 0.9 kilobase pairs if the entire HSV-tk cassette had been incorporated into the genome. c) Vector; pBluescript plasmid DNA was linearised with EcoRI, labelled and used as the vector probe.

The Muc-1 internal probe was found to hybridise to the predicted 11 kilobase pair endogenous restriction fragment in all three samples. In addition, the targeted clone #32.1 was found to display the expected 5.5 kilobase pair fragment diagnostic for the correct targeted replacement event. The aberrantly targeted clone #23.2 displayed the endogenous 11 kilobase fragment and the shifted 8 kilobase fragment, previously detected with the 3' flanking probe, probe 1. However, the internal probe also hybridised to two additional fragments of approximately 6 kilobase pairs in size. The smaller of these two hybridising fragments appeared to hybridise with the same signal intensity as the 8 kilobase pair fragment, which was known to represent 1 copy (as this fragment was also detected with the 3' flanking probe). (It should be pointed out that the reason for the difference in signal intensity between the endogenous bands and the targeted bands was due to the contribution of feeder cell DNA to the endogenous hybridising band). A densitometric scan of the autoradiogram indicated that the larger of the two 6 kilobase pair bands hybridised with twice the signal intensity of the smaller band. It was deduced that the reason for the difference in signal intensity could be due to a tail-to-tail organisation of an inserted concatemer. A tail-to-tail arrangement would juxtapose two 3' arms of homology as inverted repeats, and this would result in two copies of the 3' arm of homology being located on a single EcoRI restriction fragment of approximately 6 kilobase pairs. As there was also a second fragment of approximately 6

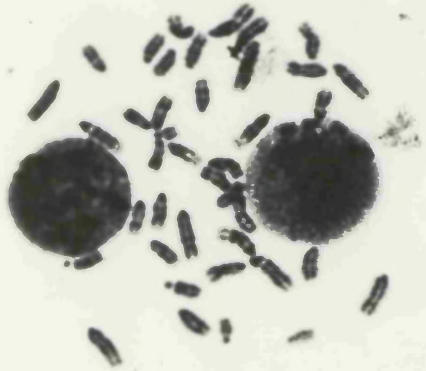
kilobase pairs that hybridised to the Muc-1 internal probe, it was reasoned that there was a third copy of the vector integrated.

Hybridisation of the three samples with the HSV-tk and vector probes confirmed that clone #23.2 was the result of an insertion of the vector DNA into the genome. Through these Southern blots and others it was deduced that the aberrantly targeted Muc-1 locus in clone #23.2 resulted from the insertion, by a single recombination event, of 3 concatemerised copies, in a head-to-tail-head-to tail-tail-to head arrangement, of the targeting vector DNA into the long arm of homology at the 5' end of the Muc-1 gene. This insertion event was subsequently modified through the deletion of the two copies of the HSV-tk gene and the two copies of the vector DNA that were present as an inverted repeat at the 5' end of the endogenous Muc-1 gene. This event may have occurred extrachromosomally prior to insertion due to the presence of inverted repeats or, alternatively, may have been selected for after insertion, by culture in the presence of gancyclovir. E, EcoRI; N, NotI; TK, Herpes Simplex virus thymidine kinase gene; vr, pBluescript vector; Hatched box, exon 7 of the Muc-1 gene; open arrows indicate the direction of transcription.

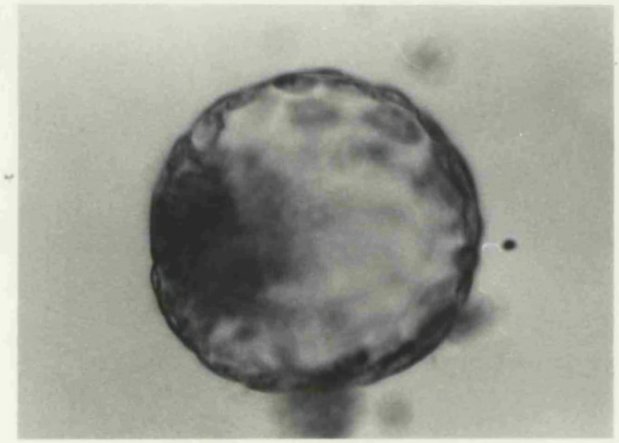

E14TG2a
32.1
23.2
E14TG2a
32.1
23.2
E14TG2a
32.1
23.2
size in kilobases
23.1
9.4
6.6
4.4
2.3
2.0
Probes:
Muc-1 internal
HSV-tk
vector


1. Linearised targeting vector
2. Concatemerisation of 3 copies of linearised vector, followed by recircularisation
3. Insertion into the Muc-1 locus by a single recombination event
Endogenous Muc-1 locus
~5.8 kbps
~6.0 kbps
4. Deletion event results in the loss of two copies of the TK gene and vector and causes a rearrangement at the 5' end of the Muc-1 gene

5.44 Chromosome analysis and embryoid body formation assay of Muc-1 targeted clone #32.1. Chromosome spreads were prepared, as described, from cells of the Muc-1 targeted ES clone #32.1. Spreads were either Giemsa stained in order to simply count the number of chromosomes, or G-banded in an attempt to determine whether or not any chromosome rearrangements had occurred. Shown is a Giemsa stained chromosome spread displaying the normal mouse diploid chromosome constitution of 2n=40. In addition to chromosomal analysis, 32.1 cells were also allowed to differentiate *in vitro* to form embryoid bodies. Shown is a typical cystic embryoid body characterised by the outer ring of endodermal cells followed by an inner differentiating layer of ectodermal cells and a large cystic cavity.



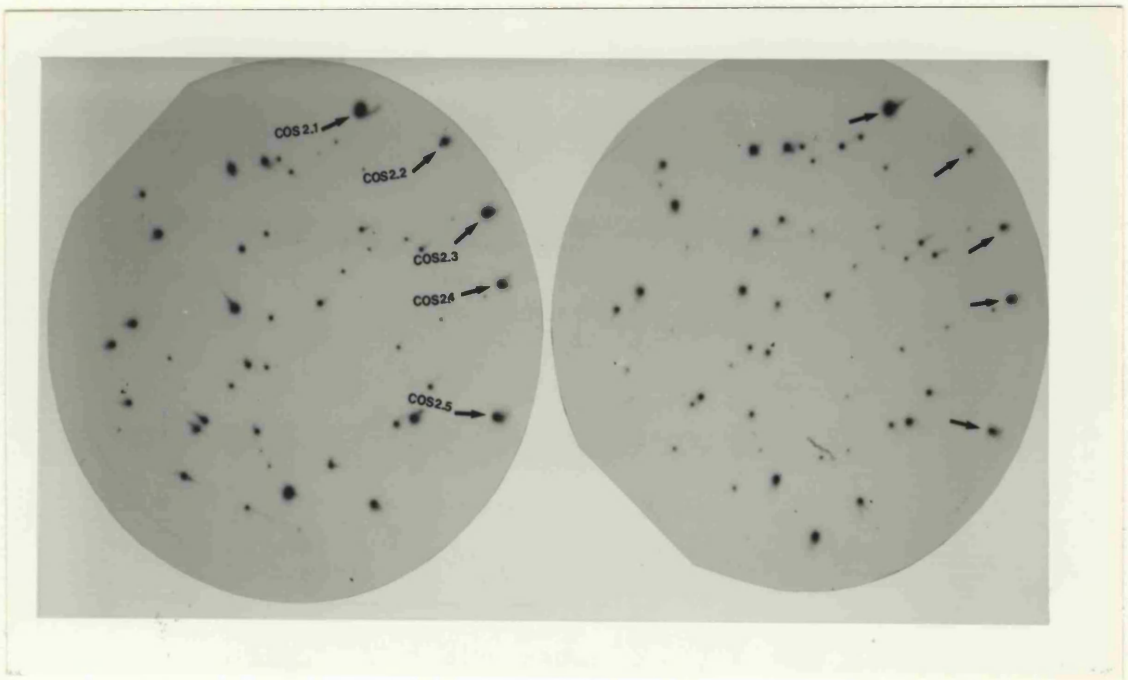
Chromosome analysis by karyotyping: Giemsa stained disrupted metaphase showing 2n=40.

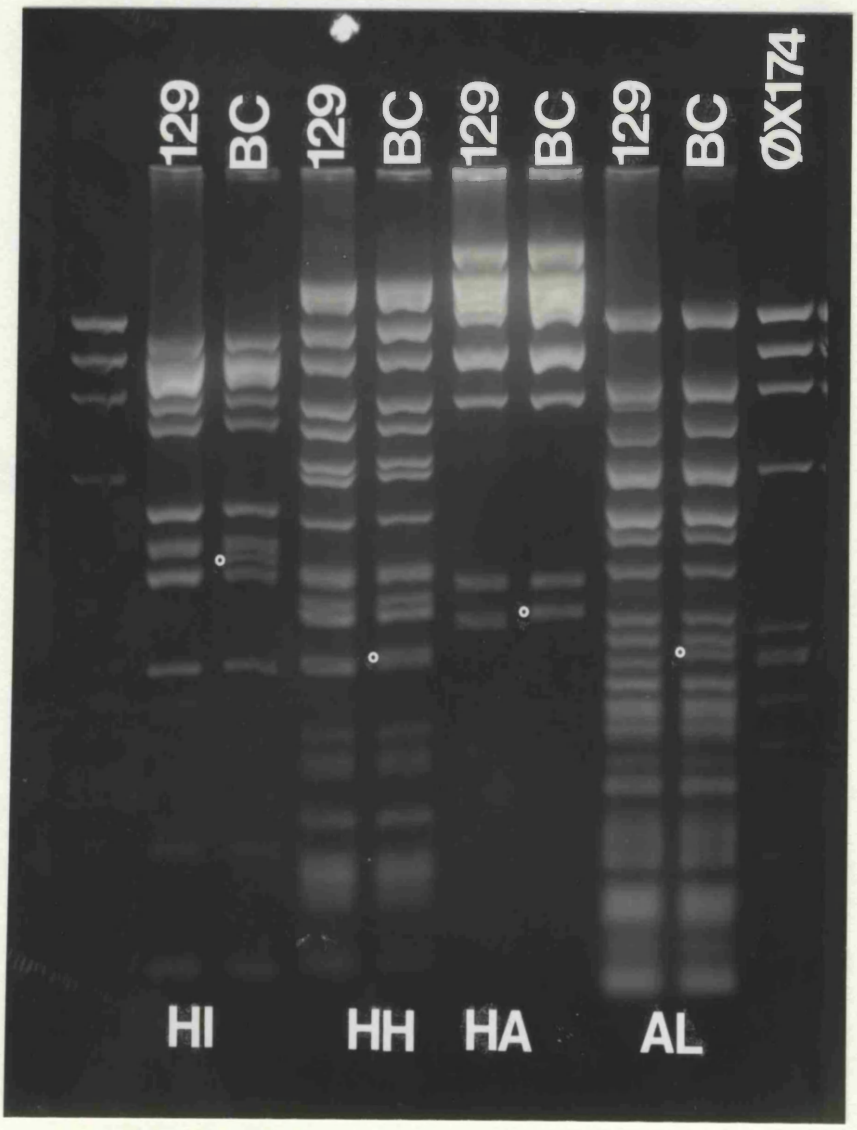


Pluripotentiality: 8 day cystic embryoid body growing in suspension culture.

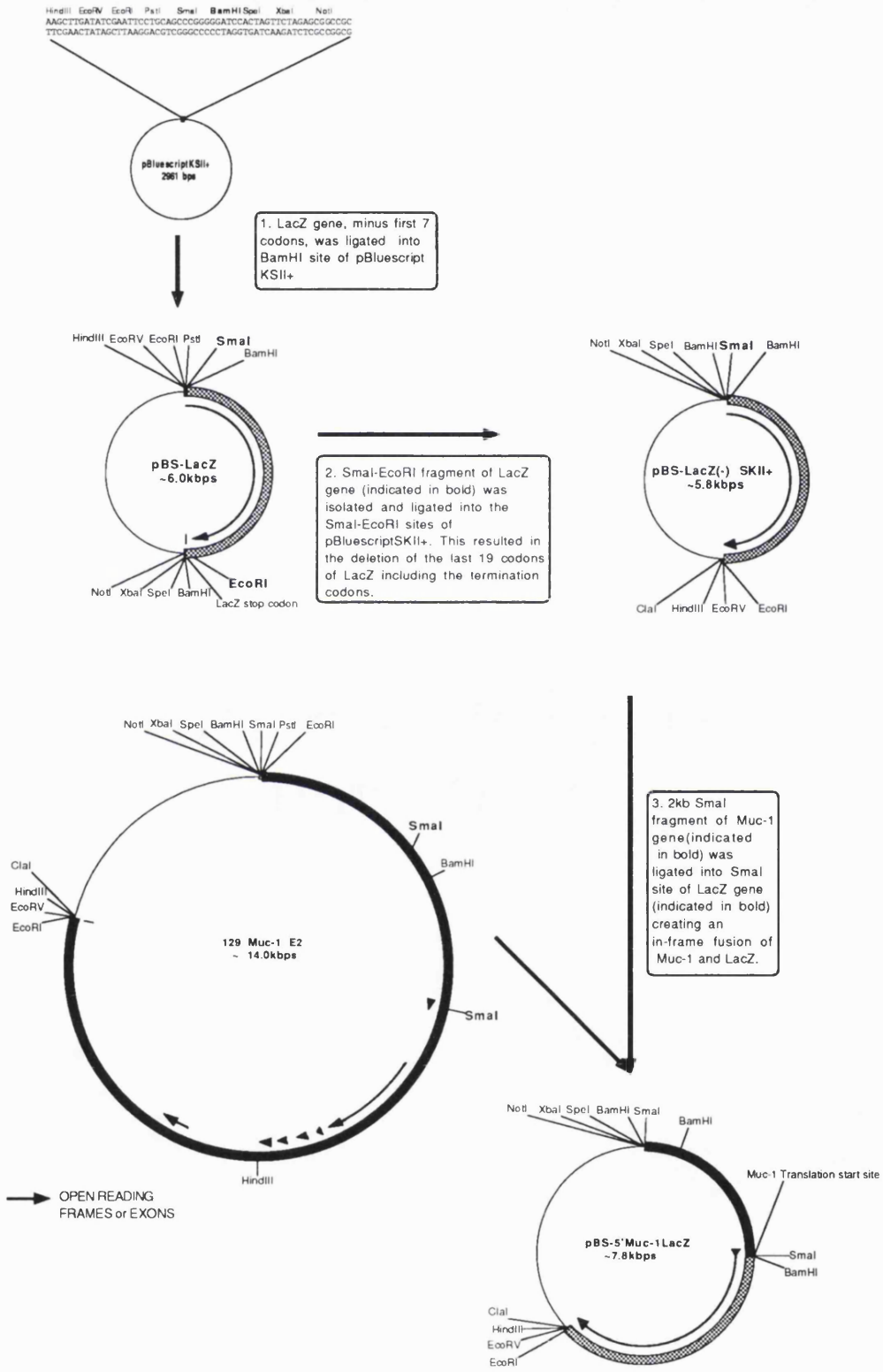
Muc-1 targeted ES clone #32.1

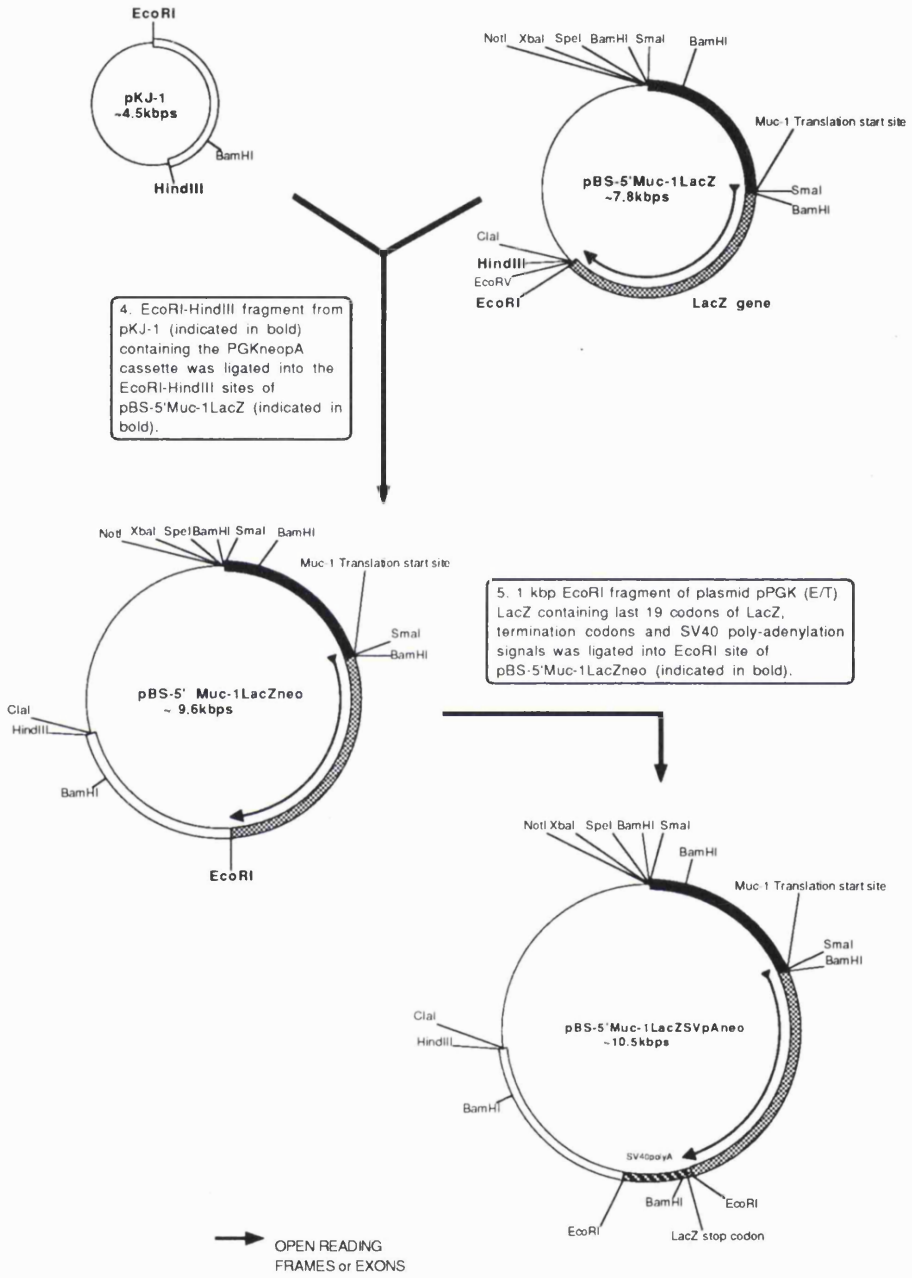
5.51 A) Double positive colonies obtained through screening a 129Sv cosmid library with the mouse Muc-1 cDNA probe pMuc2TR. Duplicate lifts were taken from a 129Sv cosmid library growing on nylon membranes on LB-Agar supplemented with ampicillin. Lifts were screened through hybridisation with the radiolabelled Muc-1 cDNA probe pMuc2TR. Shown are autoradiograms from two representative duplicate lifts displaying numerous positive colonies. Muc-1 positive colonies were found to occur at a frequency of 4×10^{-4} (1 in 2500) and, therefore, appeared to have been preferentially amplified. 5 positive colonies, designated 2.1-2.5, were picked and replated for further colony purification. **B) RFLP investigation of the Muc-1 gene isolated from two mouse strains.** The 11 kilobase pair EcoRI fragment, previously identified as containing the mouse Muc-1 gene, was cloned from the 129Sv Muc-1 cosmids into pBluescriptSKII+ in the same orientation as the Balb/c derived 11 kilobase pair Muc-1 EcoRI fragment. Approximately 4 μg plasmid DNA were digested to completion with a panel of 17 frequent cutting restriction endonucleases. Digested DNA was size-fractionated through a 2.5% agarose gel at 5V/cm, alongside ϕX174 HaeIII digested DNA molecular weight markers, and visualised over UV illumination. 4 of the 17 restriction endonucleases displayed an RFLP, indicated by a white circle, characterised by the presence of a novel restriction fragment in either the Balb/c (BC) or 129Sv (129) lane. HI, HinfI; HH, HhaI; HA, HaeII; AL, AluI.

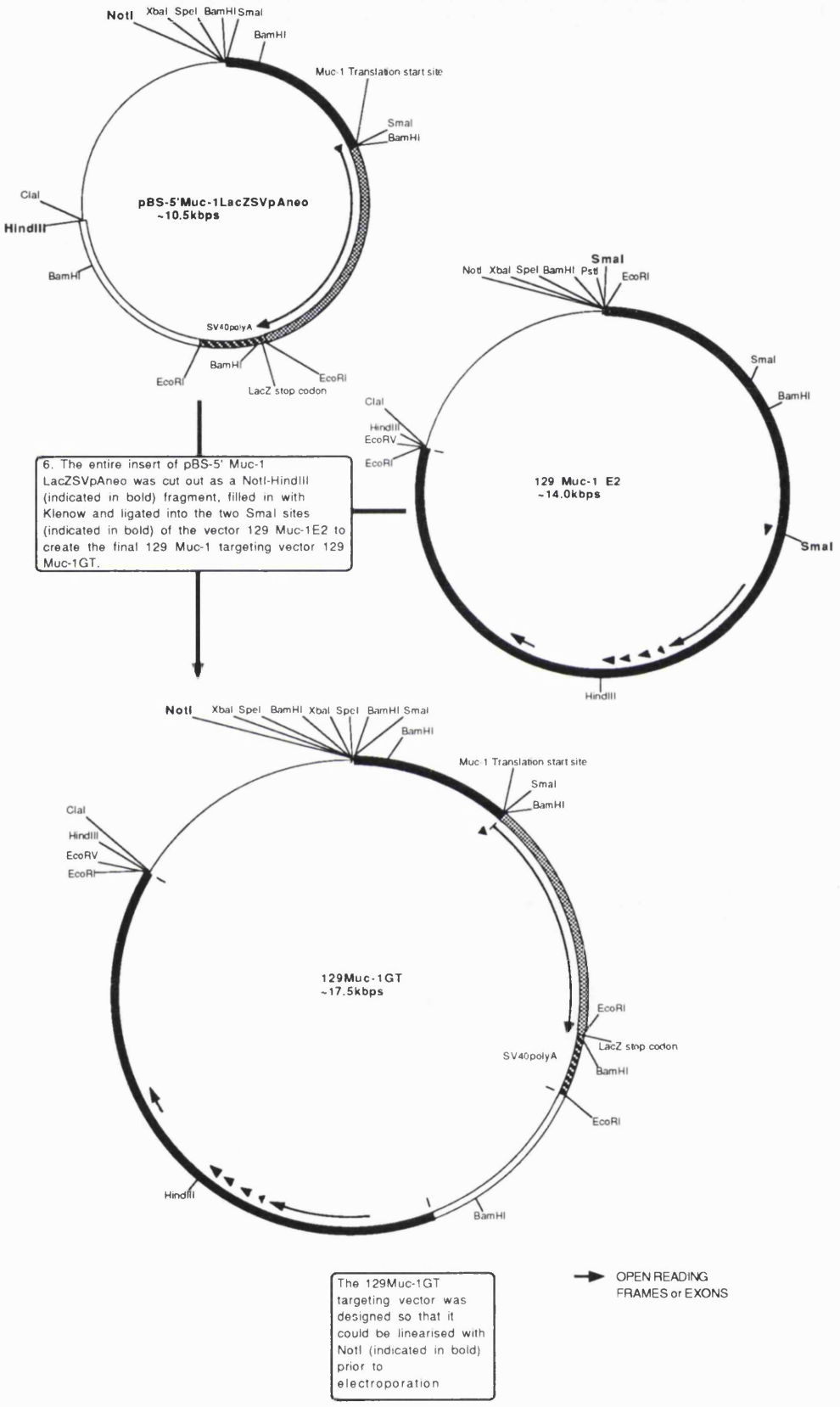

COS 2.1
COS2.2
COS2.3
COS24
COS2.5


129
BC
129
BC
129
BC
129
BC
ØX174
HI
HH
HA
AL

5.52 Cloning strategy for the construction of the 129 Muc-1 replacement vector, 129Muc-1GT. All vectors were drawn to scale, 1200 base pairs/cm. Arrow heads represent open reading frames and/or exons of the Muc-1 and LacZ genes. Important restriction endonuclease sites used in the construction of the plasmids are highlighted in bold type.

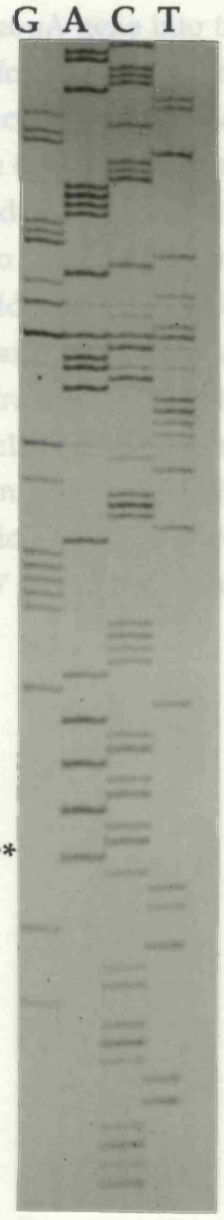

HindIII EcoRV EcoRI PstI SmaI BamHI SpeI XbaI NotI
AAGCTTGATATCGAATTCCTGCAGCCCGGGGGATCCACTAGTTCTAGAGCGGCCGC
TTCGAACTATAGCTTAAGGACGTCGGGCCCCCTAGGTGATCAAGATCTCGCCGGCG
pBluescriptKSII+
2961 bps
1. LacZ gene, minus first 7 codons, was ligated into BamHI site of pBluescript KSII+
HindIII EcoRV EcoRI PstI SmaI
BamHI
pBS-LacZ
~6.0kbps
NotI XbaI SpeI BamHI
EcoRI
LacZ stop codon
2. SmaI-EcoRI fragment of LacZ gene (indicated in bold) was isolated and ligated into the SmaI-EcoRI sites of pBluescriptSKII+. This resulted in the deletion of the last 19 codons of LacZ including the termination codons.
NotI XbaI SpeI BamHI SmaI BamHI
pBS-LacZ(-) SKII+
~5.8kbps
ClaI HindIII EcoRV EcoRI
NotI XbaI SpeI BamHI SmaI PstI EcoRI
SmaI
BamHI
ClaI
HindIII
EcoRV
EcoRI
129 Muc-1 E2
~ 14.0kbps
SmaI
HindIII
OPEN READING FRAMES or EXONS
3. 2kb SmaI fragment of Muc-1 gene(indicated in bold) was ligated into SmaI site of LacZ gene (indicated in bold) creating an in-frame fusion of Muc-1 and LacZ.
NotI XbaI SpeI BamHI SmaI
BamHI
Muc-1 Translation start site
pBS-5'Muc-1LacZ
~7.8kbps
SmaI
BamHI
ClaI
HindIII
EcoRV
EcoRI


EcoRI
pKJ-1
~4.5kbps
BamHI
HindIII
NotI XbaI SpeI BamHI SmaI
BamHI
Muc-1 Translation start site
pBS-5'Muc-1LacZ
~7.8kbps
SmaI
BamHI
ClaI
HindIII
EcoRV
EcoRI
LacZ gene
4. EcoRI-HindIII fragment from pKJ-1 (indicated in bold) containing the PGKneopA cassette was ligated into the EcoRI-HindIII sites of pBS-5'Muc-1LacZ (indicated in bold).
NotI XbaI SpeI BamHI SmaI BamHI
Muc-1 Translation start site
SmaI
BamHI
pBS-5' Muc-1LacZneo
~ 9.6kbps
ClaI
HindIII
BamHI
EcoRI
5. 1 kbp EcoRI fragment of plasmid pPGK (E/T) LacZ containing last 19 codons of LacZ, termination codons and SV40 poly-adenylation signals was ligated into EcoRI site of pBS-5'Muc-1LacZneo (indicated in bold).
NotI XbaI SpeI BamHI SmaI
BamHI
Muc-1 Translation start site
SmaI
BamHI
pBS-5'Muc-1LacZSVpAneo
~10.5kbps
ClaI
HindIII
BamHI
SV40polyA
BamHI
EcoRI
EcoRI
LacZ stop codon
OPEN READING FRAMES or EXONS

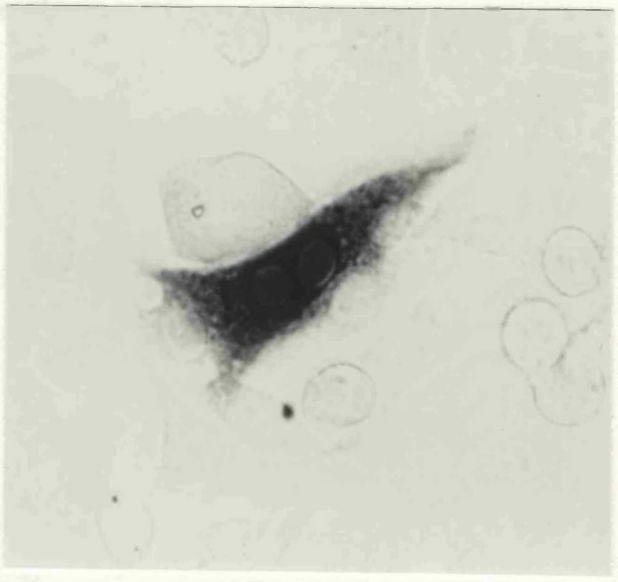

NotI
XbaI SpeI BamHI SmaI
BamHI
Muc-1 Translation start site
SmaI
BamHI
ClaI
HindIII
BamHI
pBS-5'Muc-1LacZSVpAneo
~10.5kbps
SV40polyA
BamHI
EcoRI
EcoRI
LacZ stop codon
6. The entire insert of pBS-5' Muc-1 LacZSVpAneo was cut out as a NotI-HindIII (indicated in bold) fragment, filled in with Klenow and ligated into the two SmaI sites (indicated in bold) of the vector 129 Muc-1E2 to create the final 129 Muc-1 targeting vector 129 Muc-1GT.
SmaI
NotI XbaI SpeI BamHI PstI EcoRI
SmaI
BamHI
ClaI
HindIII
EcoRV
EcoRI
129 Muc-1 E2
~14.0kbps
SmaI
HindIII
NotI XbaI SpeI BamHI XbaI SpeI BamHI SmaI
BamHI
Muc-1 Translation start site
SmaI
BamHI
ClaI
HindIII
EcoRV
EcoRI
129Muc-1GT
~17.5kbps
EcoRI
LacZ stop codon
SV40polyA
BamHI
EcoRI
HindIII
BamHI
The 129Muc-1GT targeting vector was designed so that it could be linearised with NotI (indicated in bold) prior to electroporation
OPEN READING FRAMES or EXONS

5.53 A) Sequence analysis of the Muc-1/LacZ ligation junction. Sequence was obtained from the sense strand utilising a synthetic oligonucleotide that primed at the transcription start site of the Muc-1 gene. Sequence was obtained using double-stranded plasmid DNA sequencing utilising the USB Sequenase™ version 2.0 sequencing kit. ** indicates the starting point for the sequence detailed below. The SmaI and BamHI sites utilised in the construction are underlined. **B) Expression of the Muc-1/LacZ fusion protein in an HP-1 hamster pancreatic carcinoma cell.** Approximately 5 μg plasmid were transfected into HP-1, ZR-75-1 and HT1080 cells by calcium-phosphate precipitation, as described. Approximately 24 hours after the transfection, cells were fixed and stained to detect the presence of the Muc-1/LacZ fusion protein. Only cells in the HP-1 and ZR-75-1 cultures showed positive staining. Both of these cell types have been demonstrated to express Muc-1 at high levels. Shown is a typical LacZ positive HP-1 cell.

A.

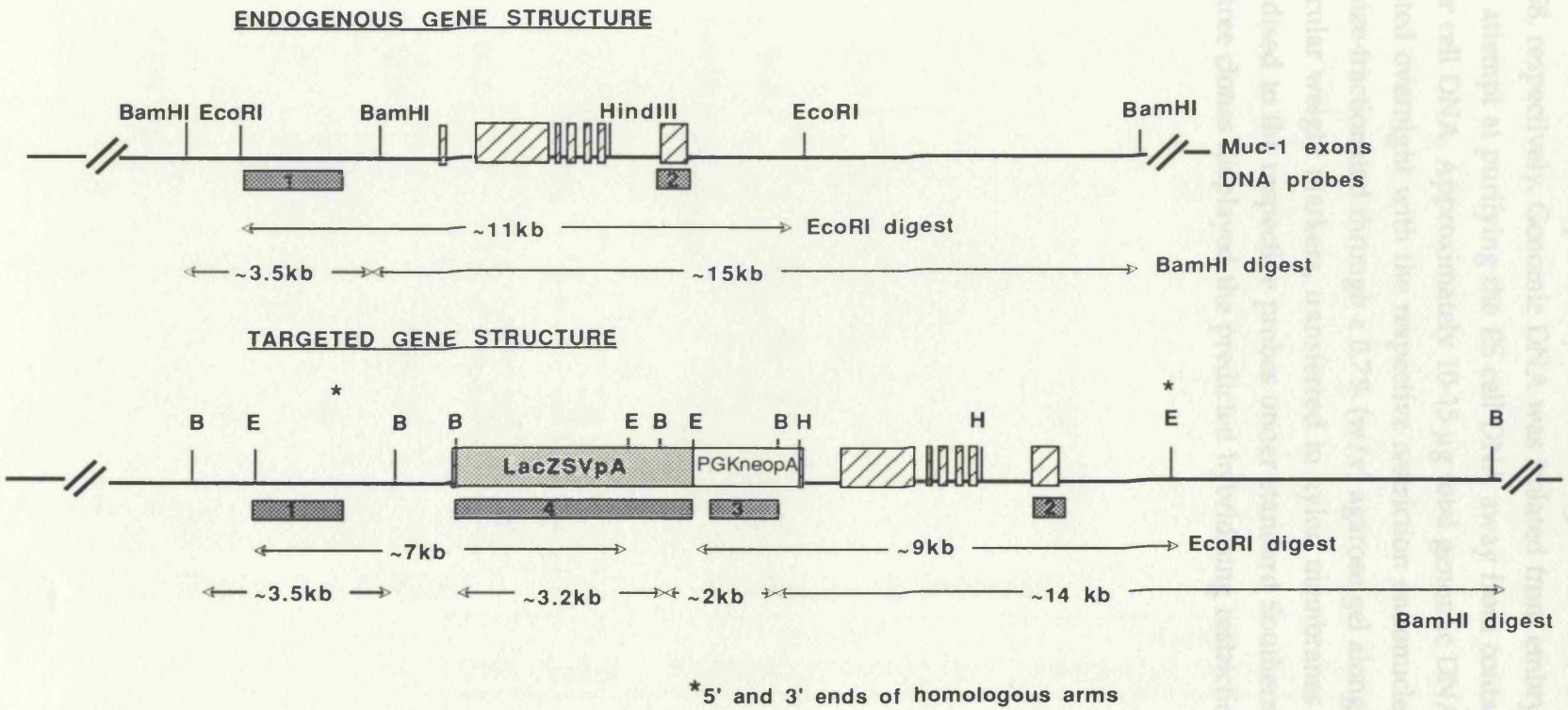


B.

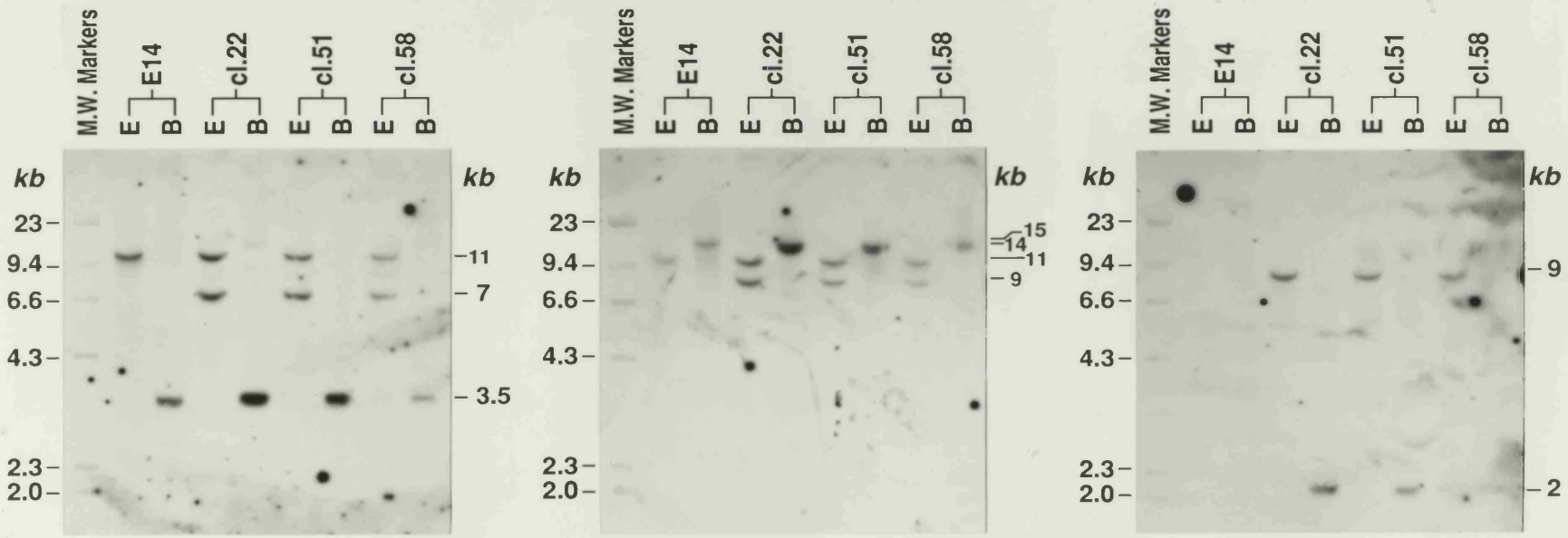


```
      Kozak    Met Thr Pro Gly Asp Pro Val Val Leu Gln
...caccaccacc ATG ACC CCG GGG GAT CCC GTC GTT TTA CAA ...
--------Muc-1 ---------->          <---------LacZ -------
```

5.61 Predicted structure of the Muc-1 gene locus after targeted replacement by the vector 129Muc-1GT. Targeted replacement of the Muc-1 gene by the 129Muc-1GT vector was predicted to result in the insertion of the LacZ gene, SV40 polyadenylation sequence and the pgkneopA gene into the Muc-1 gene. As a result of this insertion, novel sites for the restriction endonuclease EcoRI were also incorporated into the locus. In order to detect the targeted replacement event, a 5' flanking probe (EcoRI-SmaI restriction fragment) was utilised. This probe would hybridise to an 11 kilobase pair endogenous EcoRI restriction fragment, and to a targeted fragment of approximately 7 kilobase pairs. Hybridisation of the 5' flanking probe to BamHI digests of both parental and targeted DNA was predicted to hybridise to an invariant 3.5 kilobase pair fragment. In addition, the internal probes, 2 (Muc-1 cDNA probe pMuc10) and 3 (neo probe) were predicted to detect specific fragments as shown, and were utilised in order to screen for the insertion of additional copies of the vector DNA at ectopic locations. * indicates the 5' and 3' ends of the arms of homology.


ENDOGENOUS GENE STRUCTURE
BamHI
EcoRI
BamHI
HindIII
EcoRI
BamHI
Muc-1 exons
DNA probes
1
2
~11kb
EcoRI digest
~3.5kb
~15kb
BamHI digest
TARGETED GENE STRUCTURE
*
*
B
E
B
B
E
B
E
B
H
H
E
B
LacZSVpA
PGKneopA
1
4
3
2
~7kb
~9kb
EcoRI digest
~3.5kb
~3.2kb
~2kb
~14 kb
BamHI digest
*5' and 3' ends of homologous arms

5.62 Targeted inactivation of the mouse Muc-1 gene in E14TG2a cells by an isogenic Muc-1/LacZ replacement vector. Shown are the results obtained for three independently derived targeted clones, numbers, 22, 51 and 58, respectively. Genomic DNA was isolated from embryoid bodies in an attempt at purifying the ES cell DNA away from contaminating feeder cell DNA. Approximately 10-15 μg total genomic DNA were digested overnight with the respective restriction endonuclease. DNA was size-fractionated through a 0.7% (w/v) agarose gel alongside λHindIII molecular weight markers, transferred to nylon membranes and hybridised to the respective probes under standard Southern conditions. All three clones displayed the predicted hybridising restriction fragments.



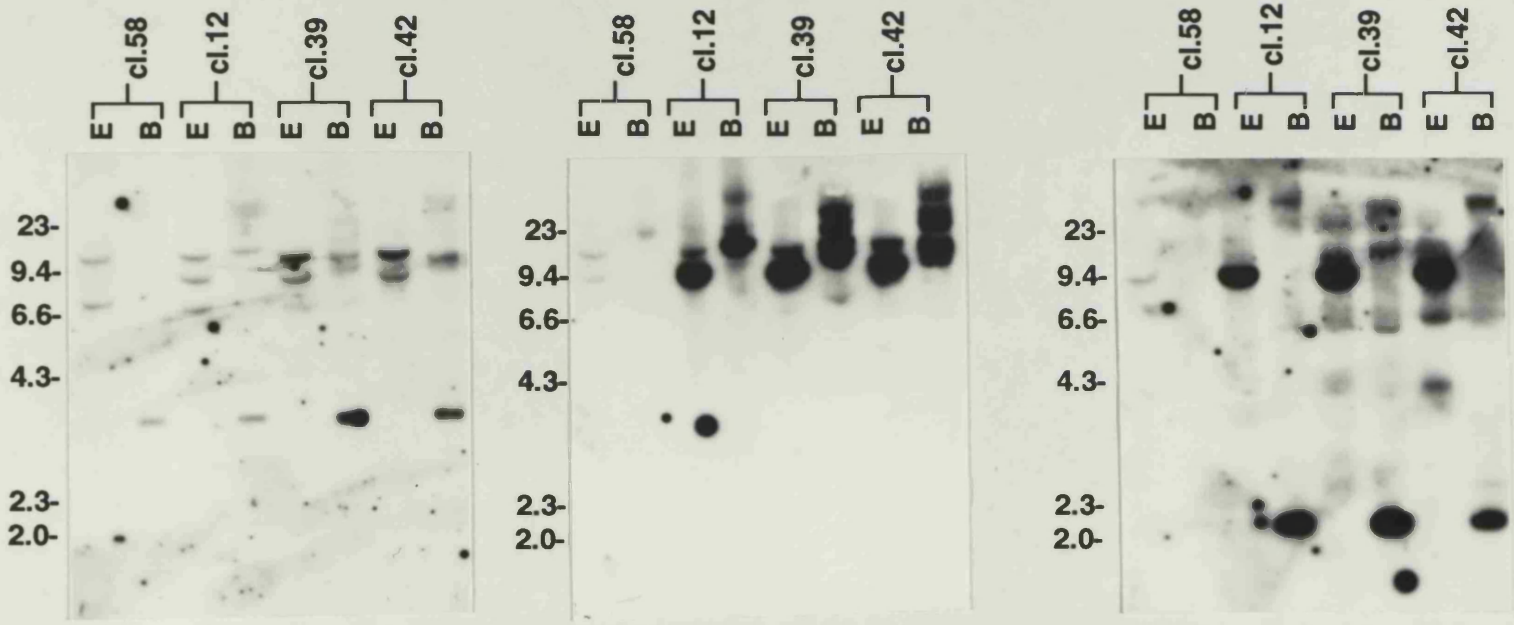
Probe 1 – 5' flank

Probe 2 – 3' internal

Probe 4 – neo

DNA Probes	EcoRI PARENTAL/TARGETED	BamHI PARENTAL/TARGETED
1	~11.0/~11.0 + ~7.0	~3.5/~3.5
2	~11.0/~11.0 + ~9.0	~15.0/~14.0
3	ND/~7.0	ND/~3.1
4	ND/~9.0	ND/~2.0

5.63 Southern analysis of aberrantly targeted ES clones. In addition to the correctly targeted clones obtained with the 129Muc-1GT replacement vector, several aberrantly targeted clones were obtained. In the majority of cases these were characterised by a shifted Muc-1 EcoRI restriction fragment of approximately 9 kilobase pairs (probe 1). One clone, #12, however, demonstrated three hybridising EcoRI restriction fragments, of 11 kilobase pairs, 9 kilobase pairs and approximately 6.5 kilobase pairs. Hybridisation of probes 2 (Muc-1 3' internal) and 3 (neo) to the aberrantly targeted clone digests indicated that these clones were the result of insertion of at least 10 copies of the targeting vector (compare intensity of band obtained for correctly targeted clone #58 = ~ 1 copy, with clones #12, #39 and #42) into the Muc-1 locus and possibly at additional locations within the genome. The concentration of G418 used in our experiments appeared to be significantly higher than that reportedly used by other laboratories and this may have contributed to selection of clones carrying more than one copy of the neo gene.

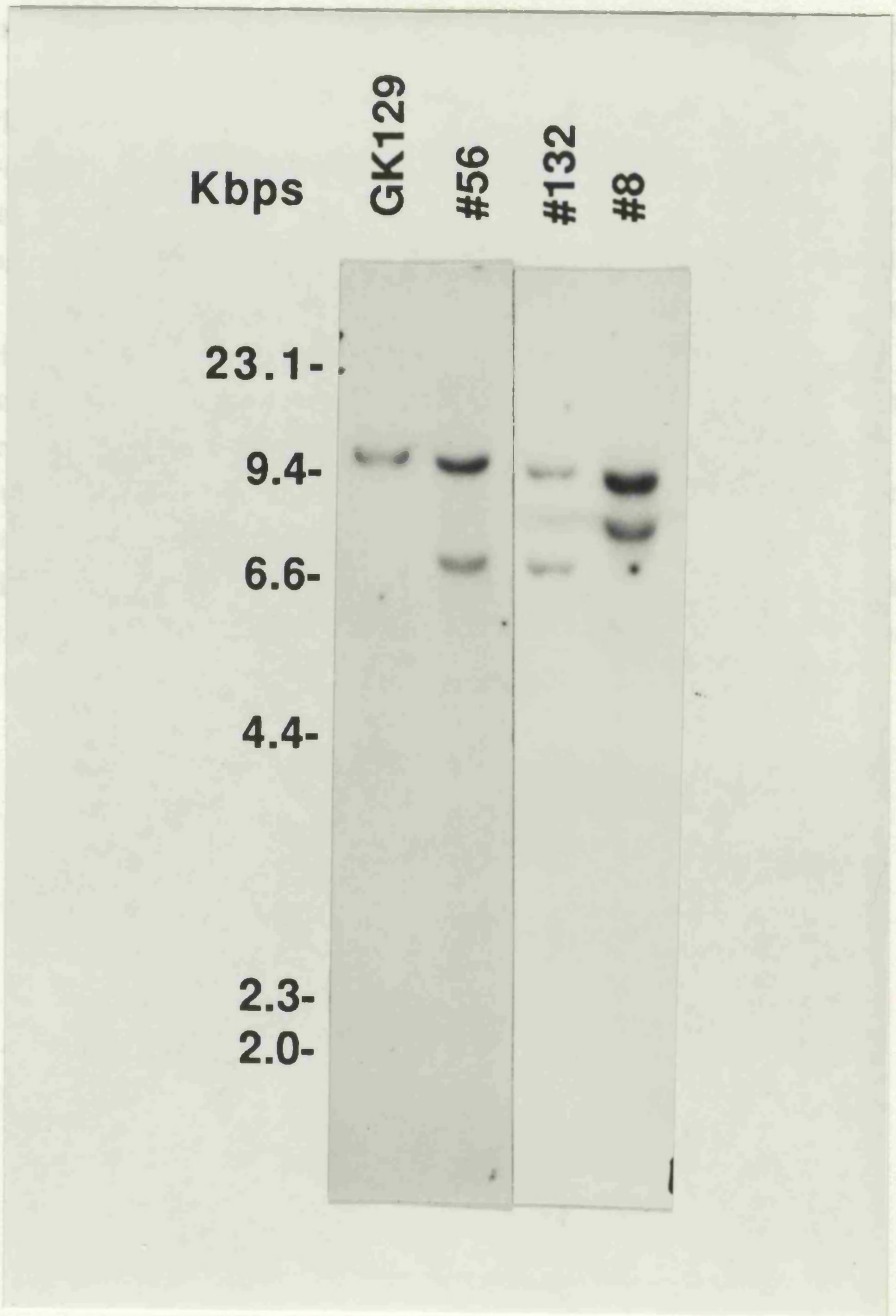


Probe 1-5' flank

Probe 2-3' internal

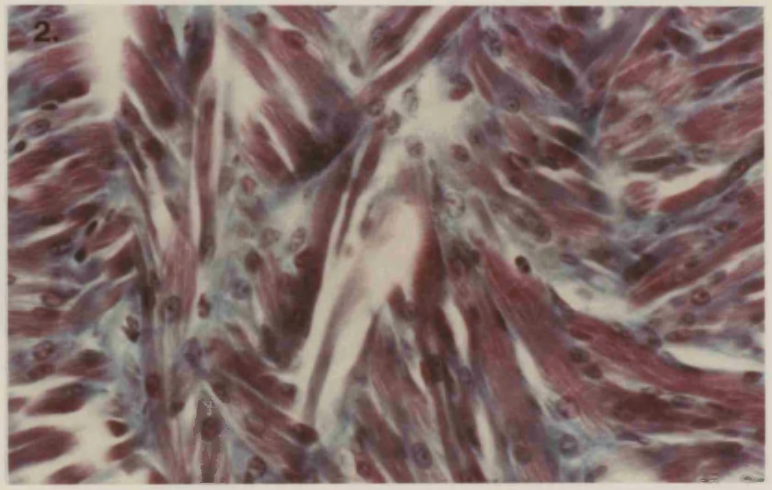
Probe 4-neo

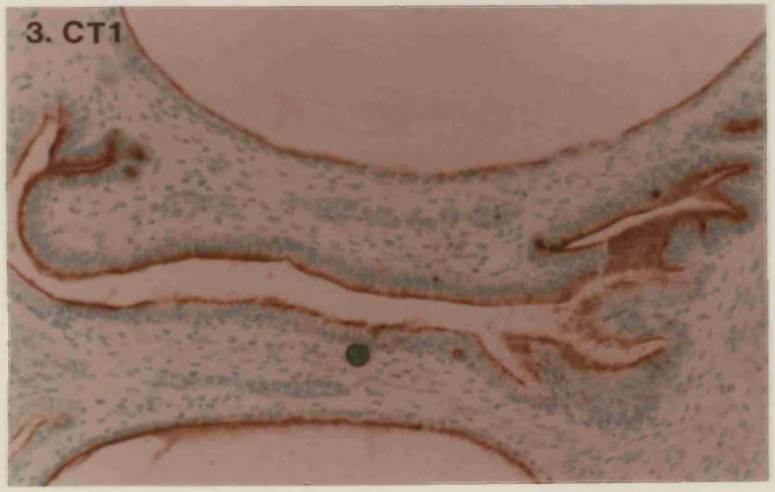
5.64 Targeted inactivation of the mouse Muc-1 gene in GK129 cells by an isogenic Muc-1/LacZ replacement vector. Total genomic DNA was prepared, as described previously, from the respective ES clones. DNA was digested overnight with EcoRI, size-fractionated through a 0.7% (w/v) agarose gel, transferred to nylon membranes and hybridised to the diagnostic 5' flanking probe. In this case, clones diagnosed as being correctly targeted (bands of 11 kbp and 7 kbp) were #56, #132, whereas clone #8 displayed the characteristic pattern for an aberrantly targeted clone (bands of 11 kbp and 9 kbp).

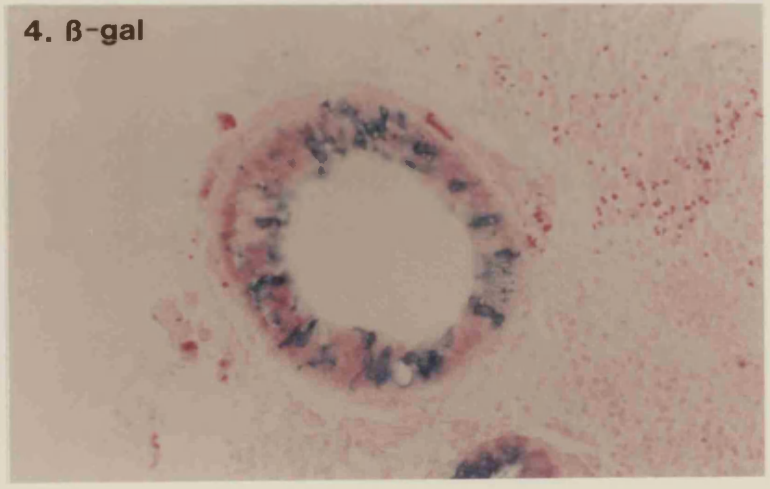

Kbps
GK129
#56
#132
#8
23.1-
9.4-
6.6-
4.4-
2.3-
2.0-

5.71 Histochemical and immunohistochemical analysis of ES-derived teratocarcinomas. Teratocarcinomas were formed through the subcutaneous injection of approximately 2 X 10^6 targeted ES cells into athymic nude mice. After 2-3 weeks, mice were sacrificed and tumours were removed for histochemical and immunohistochemical analysis as described. Panels A-D represent sections obtained from a teratocarcinoma derived from GK129 targeted clone #56. Panels A and B were trichrome stained for morphological investigation. Panel A, epithelial branching morphogenesis, 180 X mag; Panel B, differentiated smooth muscle, 360 X mag. Panel C, equivalent region of the teratocarcinoma as seen in Panel A but sectioned at a different level and stained with the polyclonal antiserum, CT1, to detect the presence of Muc-1 protein. Muc-1 protein was clearly seen on the apical surface of a large number of epithelial lumens (counterstained with methyl green), 180 Xmag. Panel D, stained to detect the presence of Muc-1 driven LacZ activity as described. LacZ positive areas were confined to cells within epithelial lumens and in all cases the staining pattern appeared to be heterogeneous (counterstained with eosin) 180 X mag.


1.

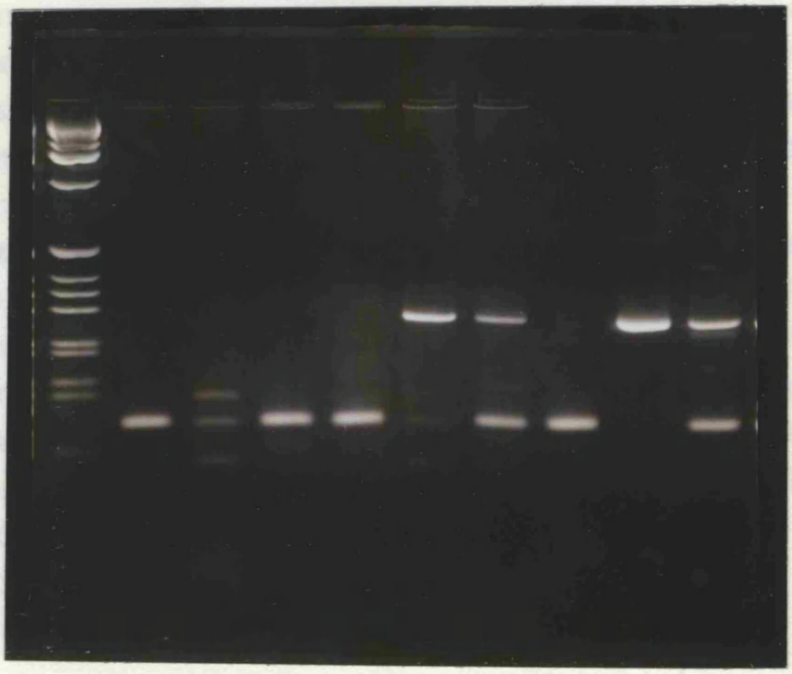

2.


3. CT1

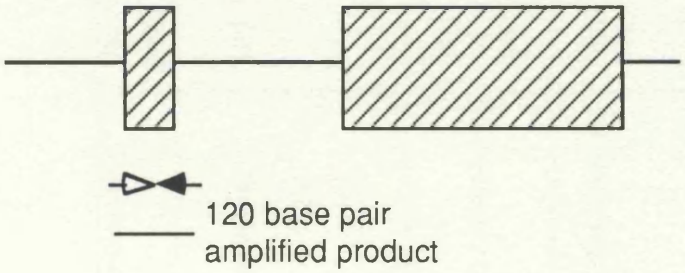

4. ß-gal

5.72 Chimaeras formed through microinjection of the targeted GK129 clone #56 into C57Bl/6 blastocysts. Between 10-20 ES cells from the Muc-1 targeted ES clone #56 were microinjected into C57Bl/6 blastocysts. Injected blastocysts were transferred into CD1 or (CBA x C57Bl/6)F_1 pseudopregnant foster mothers. Shown is one litter of mice obtained through a typical injection of GK129 clone #56. The mouse at the bottom was judged not to be chimaeric. The two mice in the centre of the photo were male and judged to be 100% ES derived, according to the amount of black blastocyst derived coat colour (none in these two cases). The mouse at the top was female and judged to be approximately 40-50% ES derived.

5.8 Germline transmission analysis of a Muc-1/LacZ mutation in agouti offspring of GK129 Muc-1 chimaeras. Genomic DNA was prepared from tail snips from agouti offspring of chimaeras. Between 500 ng and 1 μg total genomic DNA was subjected to PCR analysis with combinations of three specific oligonucleotide primers. Two oligonucleotides synthesised according to Muc-1 sequence were predicted to specifically amplify a fragment of approximately 120 base pairs from an endogenous wild-type Muc-1 allele. The 5' Muc-1 oligonucleotide, in combination with an oligonucleotide synthesised according to the sequence of the *E. coli* β-galactosidase gene was expected to amplify a fragment of approximately 250 base pairs from a targeted Muc-1 allele. Combination of all three oligonucleotides was predicted to amplify up both the 120 base pair fragment and the 250 base pair fragment from genomic DNA prepared from agouti offspring heterozygous for the designed Muc-1 mutation. PCR was carried out in a total volume of 100 μl. Twenty microlitres of the reactions were analysed by agarose gel electrophoresis through a 2.5% (w/v) agarose gel. Lane 1, BRL 1 kilobase pair ladder; Lanes 2-4, negative control C57Bl/6 mouse DNA; Lanes 5-7, agouti offspring DNA; Lanes 8-10, positive control plasmid DNA. The positive control was prepared through a 1: 1 mixture of the plasmid vectors pMucEco and 129Muc-1GT. Four nanograms of this mixture were utilised in the PCR amplification. Lanes 2, 5 and 8 represent the PCR amplified product obtained with the two Muc-1 oligonucleotides only. Lanes 3, 6 and 9 represent the PCR amplified product obtained with the 5' Muc-1 oligo and the β-gal oligo only. Lanes 4, 7 and 10 represent the PCR amplified products obtained with all three oligonucleotides. Through this analysis it could be seen that the agouti mouse was, indeed, heterozygous for the designed disruption at the Muc-1 gene locus. In the diagram below, the hatched boxes represent exons 1 and 2 of the mouse Muc-1 gene. Arrows indicate the positions of the respective PCR oligonucleotide primers.



ENDOGENOUS Muc-1 ALLELE



TARGETED Muc-1 ALLELE

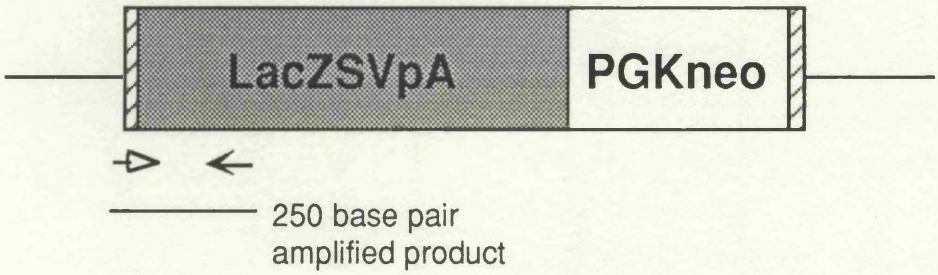


Table 5. Summary of Muc-1 gene targeting results. An unusual aspect of the results obtained through the analysis of germline transmission of the targeted GK129 clones was that a surprisingly large number of female chimaeras were observed to transmit the ES-derived agouti coat colour to their offspring. This is unusual, for reasons previously described. As ES cells have the male chromosome constitution (XY), colonisation of the germline by such cells is expected to specify the formation of a male germline. It is thought that the only way that a male ES cell line can contribute to the formation of a female germline is through previous loss of the Y chromosome, making the cells karyotypically XO. As a significant proportion of the female chimaeras that were obtained were found to transmit the ES coat colour to a small percentage (generally <5%) of their offspring this might imply that the targeted GK129 ES cell lines are made up of a mixture of XY cells along with a small percentage of XO cells.

TABLE 5

Construct	Targeting Frequency	No. of Injections (Live Births) per clone	No.Chimaeras ♂	No.Chimaeras ♀	Percentage ES Contribution	Germline
pMuc-1GT Type I	1/413 G418 resistant	#32.1 >750 (~100)	16	14	≤ 25%	NO
pMuc-1GT Type II	ND					
129 Muc-1GT	E14TG2a: 1/12 G418 resistant	#22: 18 (8)	2	2	M: 5%, 20% F: 10%, 25%	NO
		#51: 43 (8)	2		M: 5%, 10%	NO
		#58: 60 (12)	1	2	M: 10% F: 10% (2)	NO
	GK129: 1/18 G418 resistant	#22: 87 (12)	3	1	M: 95% (2), 50% F: 95%	YES
		#56: 86 (34)	13	8	M: 100% (10), 95%, 60%, 40%	YES
					F: 100% (5), 95% (2), 45%	YES
		#132: 53 (9)	5		M: 95%, 80% (2), 70%, 30%	

CHAPTER SIX:

DISCUSSION.

6.1 The mouse homologue of the human tumour-associated mucin gene, MUC1

Biochemical investigation of mucin-like glycoproteins is difficult due to their extremely large size and their high viscosity. Recently, through the use of molecular cloning techniques, clones have been isolated for several mucin genes, and it is through this kind of investigation that more is beginning to be learned about the biology of these proteins. The first gene for a human mucin-like glycoprotein that was cloned was the human MUC1 gene (Gendler, 1990a; Lan, 1990; Ligtenberg, 1990; Wreschner, 1990). Attention was initially focused on the protein product of this gene due to the fact that it was found to be highly expressed by human mammary carcinomas (Taylor-Papadimitriou, 1981). Since the initial isolation of this gene, it has been demonstrated to be expressed by the majority of simple secretory epithelia, including the mammary gland, trachea, lung, salivary glands, stomach, pancreas, kidney and uterus, in addition to being overexpressed by the majority of carcinomas (Zotter, 1988; Peat, 1992; Braga, 1992).

Mucin-like glycoproteins are traditionally thought of as being the major protein constituent of the mucous secretion that covers the surface of epithelial cells that are in contact with environmental elements. As such, they have been assumed to play primarily a protective role, protecting the epithelia for instance from bacterial attack or dessication. Mucin glycoproteins are generally thought of as being large disulphide-linked multimers that form a dense mucous gel. It was not surprising, therefore, to find that the majority of the mucin genes that have been cloned are characterised by the presence of cysteine-rich domains that presumably function in the linkage of mucin subunits (Probst, 1990; Bhargava, 1990; Eckhardt, 1991; Gum, 1991; Gum, 1992; Xu, 1992). The

sequence obtained for the human MUC1 gene, however, was found to encode an integral membrane protein. The majority of the protein was comprised of an external portion made up of multiple copies of a 20 amino acid repeat motif that was rich in potential O-glycosylation sites (serines and threonines). In addition, the sequence encoded a 31 amino acid membrane-spanning segment and a 69 amino acid cytoplasmic tail (Gendler, 1990a; Lan, 1990; Ligtenberg, 1990; Wreschner, 1990).

Since the initial characterisation of the human MUC1 gene, numerous functions have been attributed to its protein product. Functions ranging from a potential role in the metastatic spread of carcinomas to a potential role in epithelial organogenesis have been attributed to this molecule (Ligtenberg, 1992b; Braga, 1992), yet its true biological function in the normal tissues and tumours that express it remains unclear. In an attempt to progress towards an investigation of the function of MUC1, the mouse homologue was sought. It was reasoned that the advent of mouse embryonic stem cell technology in combination with gene targeting approaches, made the mouse an attractive model organism for an *in vivo* investigation of the function of the Muc-1 protein.

The human MUC1 protein has been demonstrated to be associated with elements of the actin cytoskeletal network (Parry, 1990). It was assumed that this interaction was mediated through the cytoplasmic portion of the Muc-1 molecule and would, therefore, be reflected by conservation of this sequence across species. Utilising a human cDNA probe containing sequence of the membrane-spanning and cytoplasmic portions of the protein, cDNA clones for the mouse homologue, Muc-1, were obtained. Through a variety of cloning procedures, the full-length cDNA and genomic sequence of the mouse Muc-1 gene was elucidated. The genomic structure of the mouse gene was found to be similar to its human homologue, there being seven exons and six introns. The cDNA sequence encoded an integral membrane protein with a high proportion of its coding capacity made up of serine, threonine and proline, an amino acid composition typical for a highly O-glycosylated mucin glycoprotein.

The regions of the Muc-1 protein that were observed to share the highest similarity between the human and mouse sequences were the membrane-spanning and cytoplasmic domains, suggesting an important

role in the function of the protein. It is feasible that this portion of the protein interacts with the actin microfilaments through a link protein. Significantly, the sequence conservation of the repetitive domain was below 40% overall, but this disguised the fact that the potential sites for O-linked attachment of carbohydrate side chains were highly conserved. This result fitted well with the proposal of Gendler, 1990a, that the primary function of the Muc-1 repeat domain may be to act as a scaffold for the attachment of carbohydrate side chains.

The poor conservation of sequence of the repeat as a whole explained the fact that the human MUC1 protein was so imunogenic in mice. Antibodies that recognise epitopes within the human MUC1 repeat sequence were observed to exhibit no cross-reaction with the mouse Muc-1 protein (Peat, 1992 and this study). However, a second piece of evidence that reinforced the hypothesis that the primary role of the Muc-1 tandem repeat region is to act as a carbohydrate scaffold, was obtained through the use of a monoclonal antibody (LB2) that was directed against the carbohydrate component of the human mammary MUC1 glycoprotein (Moss, 1988). In contrast to the lack of species cross-reactivity of the monoclonal antibody HMFG2, that recognises a human MUC1 core protein epitope, the monoclonal antibody LB2 was found to cross-react strongly with the mouse mammary Muc-1 mucin. This suggested that where the protein sequence of the Muc-1 molecule has not been highly conserved, the carbohydrate side-chains that are attached to the repeat region have been maintained.

The isolation of cDNA and genomic clones for the mouse Muc-1 gene enabled an investigation of its expression pattern. Utilising oligonucleotides derived from the mouse Muc-1 sequence, Braga, 1992, investigated the expression profile of the mouse Muc-1 gene during embryonic development. This investigation utilised reverse transcriptase-PCR in combination with immunohistochemistry and northern analysis of mRNA, and revealed that expression of the mouse Muc-1 gene was regulated in both a spatial and temporal fashion during embryonic development. Muc-1 protein could be detected early in epithelial organogenesis and correlated well with the onset of epithelial differentiation in each organ. In each instance, the protein was tightly localised to the apical cell surface. These results suggested that the Muc-1

protein could play an important role in the embryonic development and differentiation of epithelial organs.

Although a large amount of data is available regarding the expression of human MUC1 in carcinomas, there was no evidence that the mouse Muc-1 gene was similarly overexpressed. Investigation of Muc-1 expression in several spontaneous mouse mammary tumours indicated that the mouse gene was expressed at levels comparable to that observed in the lactating breast. In addition, the intensity of expression observed correlated well with the morphology of the tumour. Areas of hyperproliferation were observed to express Muc-1 protein at much higher levels than adjacent more normal areas. This type of heterogeneity of expression of Muc-1 is also observed in human carcinomas.

The human MUC1 protein has been demonstrated to be capable of blocking cell adhesion molecules, such as E-cadherin (Ligtenberg, 1992b, Wesseling, 1992). This is thought to be mediated through the large size of the MUC1 protein in comparison to adhesion molecules. The human MUC1 protein has been estimated to be greater than 250 nanometers in length (Bramwell, 1986). This is extremely long in comparison to the extracellular domain of cellular adhesion molecules, such as L-CAM and N-CAM, of which the extracellular domains have a bent rod-like structure with a length of only 28.4 nm and 27.8 nm, respectively (Becker, 1989). Thus, it has been hypothesised that overexpression of the MUC1 protein in carcinoma cells may have a similar effect to the down-modulation of cell-adhesion molecules (Ligtenberg, 1992b). An inverse correlation has been found between the amount of the human homologue of the cellular homotypic adhesion molecule, E-cadherin, on human carcinoma cells and their ability to invade collagen gels (Frixen, 1991). A high concentration of MUC1 on the cell surface of tumour cells may, therefore, influence their invasiveness.

The fact that the expression pattern of both the human and mouse Muc-1 proteins was found to be similar, suggested that the regulatory elements controlling their expression would be conserved. A comparison of the sequences of approximately 600 base pairs of the two promoters indicated that there was, indeed, high conservation of sequence. In particular, sequences corresponding to elements that have recently been

identified within the human MUC1 promoter as being responsible for controlling aspects of the tissue-specific expression (Abe, 1993 and Kovarik, 1993) were observed to be highly conserved.

One of the primary reasons for obtaining clones for the mouse homologue of the human tumour-associated mucin gene, MUC1, was to progress towards experiments aimed at elucidating the biological function of this molecule *in vivo*. For the mouse to be a good model organism for an investigation of function, the expression patterns should presumably be maintained and, similarly, the amino acid sequence should be conserved. The fact that the expression patterns of the human and mouse Muc-1 genes appeared to be maintained in both normal and tumour cells, that the sequence of the protein was conserved and that, in addition, the carbohydrate side-chains present on the repetitive portion of the protein appeared to be conserved, suggested that the mouse would represent an ideal organism for an investigation of the *in vivo* function of Muc-1.

6.2 Approaching the function of the Muc-1 glycoprotein

The ability to specifically mutate a single gene and subsequently to introduce this mutation into the germline of mice, with the goal being to develop mice deficient in a particular gene product, has become a powerful method for investigating the possible biological function of a protein. Utilising mouse embryonic stem cell technology in combination with the gene targeting strategy, it is now possible to make precise alterations to the mouse genome. In order to create a mouse line that was specifically mutated at the Muc-1 gene locus, experiments were designed to disrupt the mouse Muc-1 gene in embryonic stem cells.

Mouse Muc-1 Balb/c genomic DNA clones were obtained through cosmid library screening and a restriction map of the Muc-1 gene was derived. Genomic clones were utilised in the construction of standard replacement targeting vectors. A selectable neomycin resistance gene, flanked by the promoter and polyadenylation sequences of the mouse phosphoglycerate kinase-1 (Pgk-1) gene, was inserted into the first exon of the Muc-1 gene in order to disrupt the coding domain. The Muc-1 targeting vector was introduced into the mouse embryonic stem cell line E14TG2a (Hooper, 1987), and resistant colonies were selected. Through

screening by Southern analysis, one correctly targeted clone out of 172 was obtained. This represented an overall frequency of 1 correctly targeted clone in 413 G418-resistant colonies. This targeted clone was analysed extensively prior to microinjection into 3.5 day blastocysts. Although a large number of chimaeras were generated from this clone, none of them were observed to transmit the ES coat colour to their offspring.

Recently, it has been shown that the use of isogenic-derived targeting vectors greatly increases the frequency of homologous recombination in mouse ES cells (te Riele, 1992). Prior to this time, the majority of mouse genomic libraries that have been commercially available were Balb/c in origin. However, all ES cell lines in regular use today have been isolated from blastocysts of the mouse strain 129. It is presumed that the presence of small strain-specific point mutations within non-isogenic-derived targeting constructs results in a general lowering of the targeting efficiency (te Riele, 1992). This correlates well with previous research that has been carried out in bacteria and mammalian cells. In bacteria, a reduction in homology from 100% to 92% was observed to result in a decrease in recombination frequency up to 45-fold (Shen, 1986). Similarly, it has been shown that the efficiency of intrachromosomal recombination in mammalian cells also decreases as a function of percentage homology between two sequences (Bollag, 1989), and can be sensitive to as little as a single mismatch within a 1 kilobase pair interval (Letsou, 1987).

For these reasons, the mouse Muc-1 gene was re-isolated from a 129Sv cosmid library. An investigation of RFLPs present between Balb/c Muc-1 and 129Sv Muc-1 indicated that 4 out of 17 restriction endonucleases yielded informative RFLPs. Utilising the 129Sv derived Muc-1 clones, a second Muc-1 targeting vector was constructed. It has been shown that the targeting frequency in ES cells increases exponentially with an increase in the absolute length of homology (Deng, 1992). In an effort to increase the efficiency of homologous recombination further, a greater absolute amount of homology was incorporated into the new Muc-1 targeting construct (9 kilobase pairs as opposed to 5 kilobase pairs in the original targeting vector). In addition, a β-galactosidase gene was included in the targeting vector, as a Muc-1/LacZ fusion protein. A Muc-1 allele that was subsequently correctly targeted by

this construct would thus be null in terms of Muc-1 expression, and in addition express the β-galactosidase gene under the control of the Muc-1 promoter and regulatory elements.

The targeting frequency obtained with the isogenic Muc-1/LacZ replacement vector was 1 in 12 neomycin resistant colonies in the ES cell line E14TG2a. This represented a 34-fold increase in targeting frequency over the Balb/c-derived vector and, presumably, was the result of both an increase in the absolute homology incorporated into the vector, and the use of isogenic DNA in the targeting vector. All clones were analysed extensively prior to their microinjection into 3.5 day C57Bl/6 blastocysts. Numerous chimaeras were obtained from all the respective targeted clones but, again, none of these chimaeras were observed to transmit the ES coat colour to their offspring. In an attempt to overcome these difficulties, a second ES cell line, GK129, obtained at earlier passage number, was utilised. The 129 Muc-1GT targeting vector was electroporated into these cells and correctly targeted clones were obtained at a frequency of 1 in 18 neomycin resistant colonies. In this instance, chimaeras were obtained at high frequency and in many cases appeared to be 100% ES-derived for coat colour. The first chimaeras, including males and females, that were bred through these cells were found to transmit the ES agouti coat colour to their offspring and, on average, 50% of the agouti offspring were observed to be heterozygous for the specific Muc-1 mutation. To obtain mice deficient in Muc-1 protein, heterozygous animals will be crossed, and offspring will be typed. Assuming the mutation is not an early embryonic lethal, homozygous mice should be obtained at a frequency of 25%. The creation of mice deficient in Muc-1 glycoprotein is expected to be instrumental in discerning the *in vivo* biological function of this molecule.

Through these experiments, several important points were noted regarding gene targeting experiments in mouse embryonic stem cells. By far the most important of these points was the use of a stable ES cell line that has been demonstrated to routinely give rise to germline chimaeras. This should be determined prior to starting gene-disruption experiments, by immediately injecting a mouse ES cell line, obtaining chimaeras and screening for germline transmission. The use of an ES cell line that has already differentiated to the point where chimaeras derived from it will not transmit through the germline is both time-wasting and frustrating.

The second point that was noted was the use of isogenic versus non-isogenic DNA in targeting vectors, and the incorporation of the maximum amount of homology that is possible. From the results of te Riele, 1992, and others, it is apparent that wherever possible, if a clone can be obtained from an isogenic DNA library it should be. te Riele, 1992, utilising large stretches of isogenic-derived homologous DNA in their targeting vectors, demonstrated a maximum level of 78% homologous recombination at the Rb gene locus. This could partly account for the elevation in targeting frequency that we observed with the isogenic-derived vector 129Muc-1GT. Through the use of this vector it was deduced that 17% of all the targeting vector recombination events that occurred were at the correct locus. Although the use of large stretches of homology within targeting vectors precludes the use of PCR-based screening methods, the frequency of homologous recombination that is now being obtained for many genes is high enough that screening by Southern analysis has become more feasible.

6.3 Evidence for the evolution of the Muc-1 gene in mammals

The majority of mucin genes that have been characterised have been found to have a large portion of their coding capacity comprised of multiple copies of a tandem repeat which codes for a sequence rich in serine and/or threonine and proline (Gendler, 1987a, 1988 and 1990; Hoffmann, 1988; Sorimachi, 1988; Timpte, 1988; Gum, 1989; Gum, 1990; Probst, 1990; Van Cong, 1990; Porchet, 1991; Aubert, 1991; Hauser, 1992 and Toribara, 1993). Mucin glycoproteins have between 50 to 90% of their molecular mass made up of O-linked carbohydrate, and it is to the repetitive domain of the proteins that this carbohydrate is attached. In many cases, the presence of this repetitive domain gives the gene the characteristics of a hypervariable minisatellite locus. Hypervariable minisatellite sequences are generally thought of as being non-coding so-called 'junk' DNA. Therefore, the fact that the repetitive portion of mucin genes actually codes for a expressed protein product, makes the mucin gene sequences members of a small class of DNA sequence, the expressed minisatellite sequences.

Multiple copies of a 60 base pair tandem repeat encoding a 20 amino acid repeat motif were found make up the majority of the coding capacity of the human MUC1 gene (Gendler, 1990; Lan, 1990; Ligtenberg, 1990 and Wreschner, 1990). The number of tandem repeats that were

observed per allele was found to range from a low of around 20 up to a high of over 100 (Gendler, 1990). Such a large variation in the number of repeats per allele translates into protein products with a similar large variation in size. Gendler, 1990, suggested that size variations implied that the length of the protein molecule is not critically important to the function, but that the primary function of the repetitive domain was probably as a scaffold for the attachment of carbohydrate. Alignment of the human and mouse Muc-1 consensus repeat sequences suggested that this hypothesis was correct. Although the overall homology between the two respective repeat sequences was low, sequences corresponding to the potential O-linked carbohydrate attachment sites were found to be conserved.

As the repetitive portion of the Muc-1 gene is translated into a protein product, the variable number tandem repeat (VNTR) polymorphism can be observed at the RNA and protein levels, in addition to the DNA levl. The protein polymorphism is easily detected through an invetigation of milk-fat-globule (MFG) proteins. The Muc-1 protein is a major glycoprotein constituent of the MFG. SDS-PAGE analysis of milk-fat-globule proteins revealed that the mammary associated Muc-1 protein isolated from the milk of a variety of mammalian species exhibited the characteristic polymorphism; two major protein products originating from two alleles with differing numbers of tandem repeats. Of the mammals studied, the rodents were the only mammalian group that did not exhibit Muc-1 polymorphism at the protein level. Determination of sequence for the repetitive domain of the mouse Muc-1 gene indicated that the repeat sequences were very poorly conserved, an average 75% similarity between repeats in contrast to the 98-100% conservation between repeats observed in the human MUC1 gene. Due to the lack of similarity between mouse Muc-1 repeats, the level of naturally occurring mouse Muc-1 polymorphisms was assessed by a screen of greater than 50 wild rodent genomic DNA samples. Surprisingly, no repeat-mediated polymorphisms could be detected, and all rodents investigated appeared to possess Muc-1 alleles of the same length, 16 repeats.

Close inspection of the sequence of the mouse Muc-1 gene indicated that the 'modern' mouse Muc-1 gene had, however, evolved through a series of repeat duplications. Five of the sixteen Muc-1 repeats

had an additional codon, in each instance in an identical position within the repeat. Five repeats with this extra codon could only have been generated through a series of duplications. Pairwise alignments of the sequence of each repeat revealed strong evidence in support of a duplication of 5 repeats occurring at some time in the evolutionary past.

In an effort to investigate the evolution of the Muc-1 gene, clones were obtained for a variety of mammals. Clones containing repetitive sequence were obtained for gibbon, bovine and rabbit. Clones for the conserved 3' portion of the gene were obtained from gibbon, bovine, rabbit, guinea-pig and hamster. Comparison of the sequences for the respective membrane-spanning and cytoplasmic domains indicated very high conservation of sequence. In particular, a region of 27 amino acids within the cytoplasmic domain was found to be highly conserved. This sequence contained six conserved tyrosine residues, in addition to a conserved consensus site for potential phosphorylation by protein kinase C. It is possible that this region of the cytoplasmic tail binds to a link protein that in turn attaches the molecule to the actin microfilaments.

Comparison of the sequence obtained for the human, gibbon, bovine, rabbit and mouse Muc-1 consensus tandem repeats indicated that all the sequences had evolved from a common ancestor. The human, gibbon, bovine and mouse tandem repeats were 60 base pairs (20 amino acids) whereas the rabbit Muc-1 tandem repeat was a 19 amino acid variant. The human, gibbon, bovine and rabbit Muc-1 repeat sequences were found to be precise repeats, i.e. sharing 100% homology with the next repeat in most cases, whereas, as previously stated, the sequence of the mouse Muc-1 repeats was very poorly maintained. As predicted, the most highly conserved residues within the tandem repeat sequence were discovered to be those corresponding to potential O-glycosylation sites, in addition to proline residues. This would suggest that the important residues within the Muc-1 tandem repeat sequence are the serines, threonines and prolines. In combination with the observation that epitopes within the carbohydrate side-chains present on the human and mouse Muc-1 mammary mucins have been conserved, this would suggest that the carbohydrate component of the external portion of the Muc-1 molecule may be the most important in terms of function.

Of the species for which Muc-1 sequence was obtained, the mouse has been demonstrated to have diverged first during the mammalian radiation (Li, 1990). This would imply that the sequence and the length of the ancestral Muc-1 tandem repeat was specified prior to, or at the time of, the rodent divergence between 80 to 100 million years ago (Li, 1990). As the mouse Muc-1 gene is also the only Muc-1 gene that has been found not to exhibit VNTR polymorphism, two alternative hypotheses can be put forth concerning the evolution of the repeat-mediated polymorphism. The first of these two hypotheses considers that the ancestral Muc-1 gene was not hypervariable prior to the divergence of the rodents, but subsequent to the rodent divergence the gene became hypervariable. The alternative hypothesis considers that the ancestral Muc-1 repeat sequence was hypervariable or at least exhibited a degree of variability that was mediated through a 60 base pair repeat sequence. After the divergence of the rodents, events proceeded that resulted in the fixation of the mouse Muc-1 gene as a monomorphic locus. In the rest of the mammalian orders the Muc-1 gene remained polymorphic.

One piece of evidence that suggests that the alternative hypothesis may be more likely, came from a determination of the chromosomal localisation of the mouse Muc-1 gene. The human MUC1 gene has been previously localised to chromosome 1q21 (Swallow, 1987b and Middleton-Price, 1988), and thus lies within an area of the genome that has been demonstrated to be highly conserved between human and mouse (Kingsmore, 1989 and Moseley, 1989a). Through haplotype analysis of interspecific backcross mice, the mouse Muc-1 gene was localised to mouse chromosome 3 and was demonstrated to cosegregate with markers for the genes, thrombospondin-3 (Thbs-3), cluster designation 1 (Cd1) and liver red cell pyruvate kinase (Pklr). This localisation positioned the mouse Muc-1 gene within 5 megabases of a breakpoint in homology between human chromosome 1 and mouse chromosomes 1 and 3 that has been previously identified (Moseley, 1989b). A systematic comparison of genes present on human chromosome 1 and mouse chromosomes 1 and 3 has indicated strong evidence for a major rearrangement event that resulted in the translocation of a large portion of the mammalian genome to a new genomic environment (Kingsmore, 1989; Moseley, 1989a and 1989b; Oakey, 1992a and 1992b). It is not yet known when this event occurred during mammalian evolutionary

history. However, this rearrangement would have resulted in the translocation of the Muc-1 gene into a new genomic environment.

A number of studies suggest that the described rearrangement event could have been instrumental in a 'loss' of polymorphism of the mouse Muc-1 gene. For instance, it has been suggested that the mutation rate can vary in different regions of the genome (Wolfe. 1989). If the Muc-1 gene was translocated to an area of the genome in which the mutation rate was elevated this may have resulted in the repeats rapidly accumulating point mutations and the gene becoming fixed at 16 repeats. In addition, the fact that hypervariable minisatellite sequences are not distributed randomly through the genome but tend to localise to specific sites (Royle, 1988) suggests that the chromosomal localisation of the Muc-1 gene may influence its hypervariability. If, for instance, the rodent Muc-1 gene was translocated to a region of the genome in which the recombination frequency was reduced, or suppressed, this may have contributed to its eventual fixation.

The evidence presented herein suggests that the repeat domain of the Muc-1 gene is ancient in origin and that the mouse Muc-1 gene represents a rare non-polymorphic variant of it. The 'loss' of polymorphism observed at the mouse Muc-1 locus may have been partly due to a major rearrangement of the genome that occurred within 5 megabases of the gene. Due to the observation that the majority of mucin genes that have been isolated appear to exhibit repeat-mediated polymorphisms (Sorimachi, 1988; Timpte, 1988; Griffiths, 1990; Hauser, 1990; Eckhardt, 1991; Porchet, 1991; Toribara, 1991; Toribara, 1993) it appears that once hypervariability has been established at these loci, it is generally maintained. This is in contrast to the predictions that have been made with respect to non-expressed minisatellite sequences (Gray, 1991). Hypervariability of these loci has been predicted to be transient in nature. In order to answer further questions regarding the evolutionary origin of the Muc-1 gene, more ancestral Muc-1 homologues need to be investigated. As described previously, this may be aided by the fact that the protein polymorphism is accessible through SDS-PAGE analysis of milk proteins, and that the gene forms part of a tightly linked cluster of genes that may have been conserved through evolution.

6.4 Future directions and concluding remarks

In this thesis, data has been presented regarding the mouse homologue of the human tumour-associated MUC1 gene and its protein product. The mouse Muc-1 gene was isolated and fully characterised through cDNA and genomic sequence determination, through an investigation of its expression profile, through chromosomal localisation studies and through a consideration of its evolution. Through an investigation of the evolution of the mouse Muc-1 gene, the evolution of the mammalian Muc-1 gene and its tandem repeat domain was analysed. This investigation indicated that the Muc-1 repetitive domain is relatively ancient and is hypervariable in the majority of mammalian orders. The isolation of Muc-1 clones for species from a number of mammalian orders represents the first evolutionary investigation that has been carried out on expressed minisatellite mucin sequences. Through this study, it appears that the evolution of expressed hypervariable minisatellite sequences may proceed in a different manner than that of the more typical non-expressed minisatellites. A continued investigation into the evolution of this gene locus may allow models to be proposed for the evolution of mucin genes.

In addition to an investigation of the evolution of the Muc-1 gene locus, mice carrying a specific disruption in the Muc-1 gene were created through homologous recombination in mouse embryonic stem cells. In the following section, potential uses for mice deficient in the Muc-1 membrane-associated mucin-like glycoprotein will be considered.

Firstly, the role of Muc-1 during epithelial organogenesis in the developing embryo can be considered. If mice deficient in Muc-1 survive, the role of Muc-1 as a protective molecule can be assessed. For instance, its possible role in protecting epithelia from bacterial and/or viral infection. In addition, the potential role of Muc-1 as a nutritive factor present in the milk may be assessed. Mammary mucin, present in the milk, has been shown to inhibit the attachment of pathogenic *E. coli* strains to epithelial cells (Schroten, 1992). Elsewhere in the digestive system, the presence of large amounts of a highly glycosylated molecule in the gastrointestinal tract may help in, for instance, the lubrication of substances in the gut. A lack of Muc-1 may mean that secretory processes in epithelial organs are affected. In addition to the potential role of Muc-1

in normal tissues, the role that Muc-1 may play during tumour generation and metastasis can also be assessed. This could be achieved through the generation of either radiation-induced, chemically-induced or, alternatively, transgene-induced tumours. These investigations may allow us to deduce the role that Muc-1 may be playing in processes such as tumour spread and invasion. The presence of a Muc-1/LacZ fusion at the target site may help in the histochemical investigation of such tumours.

The Muc-1 glycoprotein is expressed by cells within the female reproductive tract (Zotter, 1988; Braga, in press). It has recently been demonstrated that the expression of mouse Muc-1 is reduced at around the time of embryo implantation (Braga, in press). This reduction in the level of Muc-1 glycoprotein present on the cells lining the uterus might facilitate the exposure of relevant molecule/s that are involved in mediating blastocyst attachment. In addition, Muc-1 glycoprotein was found to be expressed by granular metrial gland (GMG) cells within the placenta (Braga, in press). These cells are lymphoid in origin and their function within the placenta is unknown (Stewart, 1991). This cell type represents the first non-epithelial cell type in which Muc-1 expression has been characterised, and the possible role Muc-1 may be playing in these cells is unclear. Thus, through the development of mice deficient in Muc-1 glycoprotein, several aspects of its potential function within the female reproductive tract and during pregnancy can be considered.

If mice lacking Muc-1 glycoprotein do not survive, the presence of the created Muc-1/LacZ fusion protein will aid in an investigation of the cause of lethality. A lethal mutation might imply that the Muc-1 molecule is required during epithelial organogenesis, during processes such as lumen formation. If mice homozygous for a lack of Muc-1 are found to be nonviable, a variety of 'rescue' experiments could be carried out. The effect of re-introducing DNA fragments containing sequence coding for the various domains of the protein, and combinations of them, may allow an elucidation of the role each plays in the function of the molecule. In addition, the effect of inserting a transgene coding for a mutant Muc-1 protein can be determined. For instance, the role of the cysteine residues within the membrane-spanning domain, and the role of the tyrosine residues and the potential phosphorylation site for protein kinase C within the cytoplasmic domain can be investigated.

The creation of mice carrying a β-galactosidase gene under the control of the Muc-1 promoter at its endogenous site within the genome, allows the design of a large number of experiments aimed at elucidating factors involved in the regulation of expression of the Muc-1 gene. In particular, this type of investigation may enable more to be discerned regarding the up-regulation of expression of Muc-1 in metastatic carcinoma of the colon and the ovary, tissues that normally do not express Muc-1. The creation of mice carrying a Muc-1 regulated LacZ gene may also allow an investigation of the hormonal regulation of expression of Muc-1 in, for instance, the mammary gland.

In conclusion, the creation of mice deficient in the Muc-1 membrane-associated mucin-like glycoprotein represents the first chance to investigate the *in vivo* function of a mucin-like glycoprotein. Detailed analysis of Muc-1 deficient mice is expected to answer numerous questions regarding the biology of this molecule that have been unanswerable until now. Muc-1 deficient mice will allow an investigation of the potential function that Muc-1 may be playing in both normal tissues and during carcinogenesis. In turn, these studies may provide important novel information regarding how epithelial organs develop in the embryo and are maintained in the adult, and may also provide new insights into the multi-stage process of cancer.

APPENDIX

Appendix 1: Calculation of an approximate time for the duplication of a portion of the mouse Muc-1 gene.

In an attempt to derive an approximate figure for the length of time that has elapsed since the duplication that was identified within the mouse Muc-1 repeat domain occurred, the method of Nei, 1986, was adopted. Using this method, the number of synonymous (s) and non-synonymous (n) substitutions that had occurred between the two duplicated sequences was calculated. To provide an evolutionary reference, the sequence of the human MUC1 tandem repeat domain was utilised.

The two mouse Muc-1 sequences (from nucleotides +1010 to +1214 and from nucleotides +1319 to +1522, Fig. 3.61) were compared codon by codon and the number of synonymous and non-synonymous nucleotide differences were calculated. In addition, the mouse Muc-1 sequence from +1010 to +1214 was compared codon by codon with an equivalent human MUC1 sequence of 201 base pairs. The number of nucleotide differences were calculated as follows: where there was only a single nucleotide difference between two codons the substitution could be rapidly classified as either synonymous or non-synonymous. For example, if the codon pairs compared were CCC (Pro) and CCA (Pro), there was a single synonymous difference. The number of synonymous and non-synonymous differences per codon were denoted by s_d and n_d, respectively. In the example, $s_d = 1$ and $n_d = 0$. When two nucleotide differences were found to exist between the two codons compared, there were two possible ways to arrive at those differences. For example, in the comparison of GCC (Ala) and CCA (Pro), the two pathways are as follows: pathway I: GCC (Ala) → CCC (Pro) → CCA (Pro); pathway II : GCC (Ala) → GCA (Ala) → CCA (Pro). Pathway I involves one synonymous and one non-synonymous substitution, respectively. Similarly, pathway II involves one synonymous and one non-synonymous substitution. Assuming that pathways I and II occur with equal probability the s_d and n_d become 1.0 and 1.0, respectively. When there were three nucleotide differences between the codons being compared, there were six different possible pathways between the codons, and within each pathway there were three mutational steps. Considering all the pathways and

mutational steps, the s_d and n_d were evaluated in the same way as in the case of two nucleotide differences.

The total number of synonymous differences and non-synonymous differences was obtained by summing up these values over all codons; that is $S_d = \Sigma^r_{j=1} s_{dj}$. and $N_d = \Sigma^r_{j=1} n_{dj}$, where s_{dj} and n_{dj} were s_d and n_d for the *j*th codon, resectively, and r was the number of codons compared. $S_d + N_d$ was equal to the total number of nucleotide differences between the two DNA sequences being compared.

In order to estimate the proportion of synonymous (p_s) and non-synonymous (p_n) differences, the following equations were used:

$$p_s = S_d/S \qquad 1,$$

and

$$p_n = N_d/N \qquad 2,$$

where S and N were the average number of synonymous and non-synonymous sites for the two sequences being compared. The number of synonymous (s) and non-synonymous (n) sites present in a sequence was calculated by considering each base of each codon in turn. The fraction of synonymous changes at the *i*th position of a given codon (where i= 1, 2, 3) was denoted by f_i. The s and n for a codon were then given by $s = \Sigma^3_{i=1} f_i$ and n= (3 - s), respectively (Kafatos, 1977). For example, in the case of the codon TTA (Leu), $f_1 = 1/3$ (T→ C), $f_2 = 0$, and $f_3 = 1/3$ (A→ G). Thus, s = 2/3 and n = 7/3. For a DNA sequence of r codons, the total number of synonymous and non-synonymous sites was, therefore, given by $S = \Sigma^r_{j=1} S_i$ and N = (3r - S), respectively, where s_i was the value of s for the *i*th codon. When two sequences were compared, the averages of S and N for the two sequences were used.

Putting these formulae into practice, the following was derived:

1.Total number of synonymous substitutions (S_d)
that had occurred between the duplicated mouse
Muc-1 sequences = **1**

2. Total number of non-synonymous substitutions (N_d)

that had occurred between the duplicated mouse Muc-1 sequences = **7**

3. Total number of synonymous substitutions (S_d) that had occurred between mouse Muc-1 +1010 to +1214 and an equivalent human MUC1 sequence = **28**

4. Total number of non-synonymous substitutions (N_d) that had occurred between mouse Muc-1 +1010 to +1214 and an equivalent human MUC1 sequence = **57**

5. Total number of synonymous sites (S) and non-synonymous sites (N) present in mouse Muc-1 sequence +1010 to +1214 = **58 $^2/_3$** and **145 $^1/_3$**, respectively.

6. Total number of synonymous sites (S) and non-synonymous sites (N) present in mouse Muc-1 sequence +1319 to +1522 = **57 $^1/_3$** and **146 $^2/_3$**, respectively.

7. From 5 and 6, the average number of synonymous and non-synonymous sites present in the duplicated Muc-1 sequences = **58** and **146**, respectively.

8. Total number of synonymous (S) and non-synonymous sites present in the human MUC1 sequence = **61** and **140**, respectively.

9. From 5 and 8, the average number of synonymous and non-synonymous sites present in mouse Muc-1 +1010 to +1214 and human MUC1 = **59 $^5/_6$** and **142 $^2/_3$**, respectively.

Utilising equations 1 and 2 above, estimates for the proportion of synonymous (p_s) and non-synonymous (p_n) substitutions that have occurred between the two mouse Muc-1 sequences and between one mouse Muc-1 sequence and its human equivalent were obtained:

p_s mouse = 1÷ 58 or **0.017**

p_n mouse = 7÷ 146 or **0.048**

p_s mouse/human = 28÷ 59 $^5/_6$ or **0.468**

p_n mouse/human = 57 ÷ 142 $^2/_3$ or **0.400**

To estimate the number of synonymous substitutions (d_s) and non-synonymous substitutions (d_n) per site, the formula developed by Jukes and Cantor, 1969 was utilised as follows:

$$d = -\,^3/_4 \ln (1 - \,^4/_3 p)$$

mouse Muc-1 d_s = **0.017**
mouse Muc-1 d_n= **0.050**

mouse/human MUC1 d_s = **0.733**
mouse/human MUC1 d_n = **0.572**

The mouse and human lineages have been estimated to have diverged between 80 and 100 million years ago (Li, 1990). Taking these estimates into consideration, along with the approximate figures for the number of synonymous and non-synonymous substitutions per site that have occurred between the duplicated mouse Muc-1 sequence and the mouse and human Muc-1 sequences, this would place an approximate timescale of between 2 and 9 million years for the mouse Muc-1 duplication. A comparison of the rates of synonymous substitution gave an approximate time of (80-100 Myr) ÷ (0.733 ÷ 0.017) = 1.86 to 2.23 million years. A comparison of the rates of non-synonymous substitutions gave an approximate time of (80-100 Myr) ÷ (0.572 ÷ 0.050) = 6.99 to 8.74 million years. However, extrapolation from the estimated values for the number of non-synonymous and synonymous substitutions per site in the rodents (1.1×10^{-9} and 6.5×10^{-9}, respectively (Li, 1987)), yielded approximate timescales for the mouse Muc-1 duplication of between 45 million years and 2.6 million years. The reason for the large range in this calculation was due to the fact that the rate of non-synonymous substitution appears to have been higher than the rate of synonymous substitution within the mouse Muc-1 duplication. This might suggest that different areas within the repeat sequences are subjected to different levels of selection pressure. This, in turn, would fit with the hypothesis that it is primarily the potential O-glycosylation sites that are important

within the repeat sequence. The selection pressure that operates upon other residues within the repeat unit may be markedly different.

In general, therefore, the most important piece of evidence that was obtained regarding the evolution of the mouse Muc-1 duplication was the fact that Southern analysis of rat genomic DNA restriction endonuclease digests revealed that the mouse and rat Muc-1 alleles were the same size (Spicer, 1991). The mouse and rat have been estimated to have diverged approximately 15 million years ago, and this implies that the Muc-1 duplication occurred at least 15 million years ago.

Appendix 2: Oligonucleotides utilised in the construction of the Balb/c targeting vectors pMuc-1GT Type I and pMuc-1GT Type II, and in the PCR-based screening for homologous recombinants.

The oligonucleotides utilised were as follows:

A) pMuc-1GT Type I 3' arm of homology:

5' oligo: 5'- CTCACGGACGCTACGTGCCC-3'
corresponding to nucleotides +3028 to +3047 (Fig 3.61)

3' oligo: 5'- CCCGCGGCCGCCCCAGTGTCCCCCAGGGC-3'
corresponding to nucleotides +3523 to +3504 (antisense) (Fig 3.61)

B) pMuc-1GT Type II 3' arm of homology:

5' oligo: 5'- CCCAAGCTTGTCACCTCATCTCAGGACACC-3'
corresponding to nucleotides +938 to +958 (Fig 3.61)

3' oligo: 5'- CCCGCGGCCGCGTCTAGACTGGTAGCTGAGCC-3'
corresponding to nucleotides +1905 to +1885 (antisense) (Fig 3.61)

PCR amplified products were restriction digested with HindIII and NotI (sites underlined) and ligated into the vector pBS-HSV-tk to form the vectors 3' Type I and 3' Type II, respectively.

Vectors constructed for use in determination of the conditions for PCR detection of Muc-1 homologous recombinants were made as follows:

The Type I control vector 3' arm was an approximately 3.8 kilobase pair HindIII-EcoRI restriction endonuclease fragment, stretching from the HindIII site present within the mouse Muc-1 last intron to the EcoRI site approximately 2.5 kilobase pairs downstream of the Muc-1 polyadenylation signal.

The Type II control vector 3' arm was amplified by PCR utilising the same 5' oligonucleotide employed for the production of the Type II targeting vector 3' arm of homology, but with an alternative 3' oligo:

3' oligo: 5'- CCCGCGGCCGCTGCAGAAACTGGGAGAAGAGG-3'
corresponding to nucleotides +2286 to +2266 (antisense) (Fig 3.61)

After PCR amplification, the respective 3' arms of homology were digested with the restriction endonucleases HindIII and NotI and ligated into the vector pBS-HSV-tk.

For use in the PCR-based detection of homologous recombinants the following oligonucleotides were designed and constructed:

Two nested oligonucleotides were designed for each construct and were used in conjunction with two nested oligonucelotides synthesised according to the sequence for the mouse phosphoglycerate kinase-1 polyadenylation sequence:

PGK oligo 1: 5'- CCTGAAGAACGAGTCAGCAG -3'
PGK oligo 2 (nested): 5'- GCCTCTGTTCCACATACACT -3'

Muc-1 Type I oligo 1: 5'- GGGTTCCCATCCCTGTCTCC -3'
corresponding to nucleotides +3582 to +3563 (antisense) (Fig 3.61)

Muc-1 Type I oligo 2 (nested):
5'- GGCCAGTCCTTCTGAGAGCC -3'
corresponding to nucleotides +3553 to +3524 (antisense) (Fig 3.61)

Muc-1 Type II oligo 1: 5'- AGGGGACACATGGCAGAGGC -3'
corresponding to nucleotides + 2215 to +2196 (antisense) (Fig 3.61)

Muc-1 Type II oligo 2 (nested):
5'- GGAGCTGGGGTCTTCCAGAG -3'
corresponding to nucleotides + 2166 to +2147 (antisense) (Fig 3.61)

REFERENCES

Abe, M., and Kufe, D. (1993). Characterization of cis-acting elements regulating transcription of the human DF3 breast carcinoma-associated antigen (MUC1) gene. *Proc. Natl. Acad. Sci. U.S.A.* **90:** 282-286.

Andreason, G. L., and Evans, G.A. (1989). Optimization of electroporation for transfection of mammalian cell lines. *Anal. Biochem.* **180:** 269-275.

Appleby, M. W., Gross, J.A., Cooke, M.P., Levin, S.D., Qian, X., and Perlmutter, R.M. (1992). Defective T cell receptor signalling in mice lacking the thymic isoform of $p59^{fyn}$. *Cell* **70:** 751-763.

Arklie, J., Taylor-Papadimitriou, J., Bodmer, W., Egan, M., and Millis, R. (1981). Differentiation antigens expressed by epithelial cells in the lactating breast are also detectable in breast cancers. *Int. J. Cancer* **28:** 23-29.

Armour, J. A. L., Patel., I., Thein, S.L., Fey, M.F., and Jeffreys, A.J. (1989a). Analysis of somatic mutations at minisatellite loci in tumours and cell lines. *Genomics* **4:** 328-334.

Armour, J.A.L., Wong, Z., Wilson, V., Royle, .J., and Jeffreys, A.J. (1989b). Sequences flanking the repeat arrays of human minisatellites: association with tandem and dispersed repeat elements. *Nucl. Acids Res.* **17:** 4925-4935.

Ashall, F., Bramwell, M.E., and Harris, H. (1982). A new marker for human cancer cells. *Lancet* **ii:** 1-6.

Aubert, J. P., Porchet, N., Crepin, M., Duterque-Coquillaud, M., Vergnes, G., Mazzuca, M., Debuire, B., Petitprez, D., Degand, P. (1991). Evidence for different human tracheobronchial mucin peptides deduced from nucleotide cDNA sequences. *Am. J. Respir. Cell Mol. Biol.* **5:** 178-185.

Balk, S.P., Bleicher, P.A., and Terhorst, C. (1989). Isolation and characterization of a cDNA and gene coding for a fourth CD1 molecule. *Proc. Natl. Acad. Sci. U.S.A.* **86:** 252-256.

Ball, R. K., Friis, R.R., Schonenberger, C.-A., Doppler, W., and Groner, B. (1988). Prolactin regulation of beta-casein gene expression and of a cytosolic 120-kd protein in a cloned mouse mammary epithelial cell line. *EMBO J.* **7:** 2089-2095.

Barcellos-Hoff, M. H., Aggeler, J., Ram, T.G., and Bissell, M.J. (1989). Functional differentiation and alveolar morphogenesis of primary mammary cultures on reconstituted basement membranes. *Development* **105:** 223-235.

Baribault, H., and Oshima, R.G. (1991). Polarized and functional epithelia can form after the targeted inactivation of both mouse keratin 8 alleles. *Development* **115:** 1675-1684.

Barnd, D. L., Lan, M.S., Metzgar, R.S., and Finn, O.J. (1989). Specific, major histocompatibility complex-unrestricted recognition of tumor-associated mucins by human cytotoxic T cells. *Proc. Natl. Acad. Sci. U.S.A.* **86:** 7159-7163.

Becker, J.W., Erickson, H.P., Hoffman, S., Cunningham, B.A., and Edelman, G.M. (1989). Topology of cell adhesion molecules. *Proc. Natl. Acad. Sci. U.S.A.* **86:** 1088-1092.

Behrens, J., Mareel, M.M., VanRoy, F.M., and Birchmeier, W. (1989). Dissecting tumor cell invasion: epithelial cells acquire invasive properties following the loss of uvomorulin-mediated cell-cell adhesion. *J. Cell Biol.* **108:** 2435-2447.

Bernelot Moens, C., Auerbach, A.B., Conlon, R.A., Joyner, A.L., and Rossant, J. (1992). A targeted mutation reveals a role for N-myc in branching morphogenesis in the embryonic mouse lung. *Genes Dev.* **6:** 691-704.

Bhargava, A. K., Woitach, J.T., Davidson, E.A., and Bhavanandan, V.P. (1990). Cloning and cDNA sequence of a bovine submaxillary gland mucin-like protein containing 2 distinct domains. *Proc. Natl. Acad. Sci. U.S.A.* **87:** 6798-6802.

Bishop, D.T. (1988). The information content of phase-known matings for ordering genetic loci. *Genetic Epidemiology* **2:** 349-361.

Boland, C. R., Montgomery, C.K., and Kim, Y.S. (1982). Alterations in human colonic mucin occurring with cellular differentiation and malignant transformation. *Proc. Natl. Acad. Sci. U.S.A.* **79:** 2051-2055.

Bollag, R. J., Waldman, A.S., and Liskay, R.M. (1989). Homologous recombination in mammalian cells. *Annu. Rev. Genet.* **23:** 199-225.

Boshell, M., Lalani, E-N., Pemberton, L., Burchell, J., Gendler, S., and Taylor-Papadimitriou, J. (1992). The product of the human MUC1 gene when secreted by mouse cells transfected with the full-length cDNA lacks the cytoplasmic tail. *Biochem. Biophys. Res. Commun.* **185:** 1-8.

Bradley, A., Evans, M.J., Kaufman, M.H., and Robertson, E.J. (1984). Formation of germ line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature* **309:** 255-256.

Braga, V. M. M., Pemberton, L.F., Duhig, T., and Gendler, S.J. (1992). Spatial and temporal expression of an epithelial mucin, Muc-1, during mouse development. *Development* **115:** 427-437.

Braga, V.M.M., and Gendler, S.J. Modulation of Muc-1 mucin expression in the mouse uterus during the estrus cycle, early pregnancy and placentation. *J. Cell Sci.* in press.

Bramwell, M. E., Wiseman, G., and Shotton, D.M. (1986). Electron-microscopic studies of the Ca antigen, epitectin. *J. Cell Sci.* **86:** 249-261.

Braun, T., Rudnicki, M.A., Arnold, H-H., and Jaenisch, R. (1992). Targeted inactivation of the muscle regulatory gene Myf-5 results in abnormal rib development and perinatal death. *Cell* **71:** 369-382.

Burchell, J., Durbin, H., and Taylor-Papadimitriou, J. (1983). Complexity of expression of antigenic determinants recognised by monoclonal antibodies HMFG1 and HMFG2, in normal and malignant human mammary epithelial cells. *J. Imunol.* **131:** 508-513.

Burchell, J., Wang, D., and Taylor-Papadimitriou, J. (1984). Detection of the tumour associated antigens recognised by the monoclonal antibodies HMFG-1 and 2 in serum from patients with breast cancer. *Int. J. Cancer* **34:** 763-768.

Burchell, J., Taylor-Papadimitriou, J., Boshell, M., Gendler, S., and Duhig., T. (1987). Development and characterization of breast cancer reactive monoclonal antibodies directed to the core protein of the human milk mucin. *Int. J. Cancer* **44:** 691-696.

Burchell, J., Taylor-Papadimitriou, J., Boshell, M., Gendler, S., and Duhig, T. (1989). A short sequence, within the amino acid tandem repeat of a cancer-associated mucin, contains immunodominant epitopes. *Int. J. Cancer* **44:** 691-696.

Campana, W. M., Josephson, R.V., and Patton, S. (1992). Presence and genetic polymorphism of an epithelial mucin in milk of the goat (*Capra hircus*). *Comp. Biochem. Physiol.* **103B:** 261-266.

Carraway, K. L., Fregien, N., Carraway III, K.L.,and Carraway, C.A.C. (1992). Tumor sialomucin complexes as tumor antigens and modulators of cellular interactions and proliferation. *J. Cell Sci.* **103:** 299-307.

Ceriani, R. L., Thompson, K., Peterson, J.A., and Abraham, S. (1977). Surface differentiation antigens on human mammary epithelial cells carried on the human milk fat globule. *Proc. Natl. Acad. Sci. U.S.A.* **74:** 582-586.

Ceriani, R. L., Peterson, J.A., Lee, J.Y., Moncada, R., and Blank, E.W. (1983). Characterisation of cell surface antigens of human mammary epithelial cells with monoclonal antibodies prepared against human milk fat globule. *Som. Cell Genet.* **9:** 415-.

Chandley, A. C., and Mitchell, A.R. (1988). Hypervariable minisatellite regions are sites for crossing-over at meiosis in man. *Cytogenet. Cell Genet.* **48:** 152-155.

Charron, J., Malynn, B.A., Robertson, E.J., Goff, S.P., and Alt, F.W. (1990). High-frequency disruption of the N-myc gene in embryonic stem and

pre-B cell lines by homologous recombination. *Mol. Cell Biol.* **10:** 1799-1804.

Chase, J. W., and Williams, K.R. (1986). Single-stranded DNA binding proteins required for DNA replication. *Ann. Rev. Biochem.* **55:** 103-136.

Chirgwin, J. M., Przybyla, A.E., MacDonald, R.J., and Rutter, W.J. (1979). Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* **18:** 5294-5299.

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

Chisaka, O., Musci, T.S., and Capecchi, M.R. (1992). Developmental defects of the ear, cranial nerves and hindbrain resulting from targeted disruption of the mouse homeobox gene Hox-1.6. *Nature* **355:** 516-520.

Church, G. M., and Gilbert, W. (1984). Genomic sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **81:** 1991-1995.

Clark, J. M. (1988). Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases. *Nucl. Acids Res.* **16:** 9677-9686.

Codington, J.F., Sanford, B.H., and Jeanloz, R.W. (1972). Glycoprotein coat of the TA3 cell. Isolation and partial characterization of a sialic acid containing glycoprotein fraction. *Biochemistry* **11:** 2559-2564.

Codington, J.F., and Frim, D.M. (1983). Cell-surface macromolecular and morphological changes related to allotransplantability in the TA3 tumor. *Biomembranes* **11:** 207-258.

Collick, A., and Jeffreys, A.J. (1990). Detection of a novel minisatellite-specific DNA-binding protein. *Nucl. Acids Res.* **18:** 625-629.

Collick, A., Dunn, M.G., and Jeffreys, A.J. (1991). Minisatellite binding protein Msbp-1 is a sequence-specific single-stranded DNA-binding protein. *Nucl. Acids Res.* **19:** 6399-6404.

Cooney, A. J., Tsai, S.Y., O'Malley, B.W., and Tsai, M.-J. (1992). Chicken ovalbumin upstream promoter transcription factor (COUP-TF) dimers bind to different GGTCA response elements, allowing COUP-TF to repress hormonal induction of the vitamin D3, thyroid hormone, and retinoic acid receptors. *Mol. Cell. Biol.* **12:** 4153-4163.

Del Sal, G., Manifioletti, G., and Schneider, C. (1988). A one-tube plasmid DNA mini-preparation suitable for DNA sequencing. *Nucl. Acids Res.* **16:** 9878.

Deng, C., and Capecchi, M.R. (1992). Reexamination of gene targeting frequency as a function of the extent of homology between the targeting vector and the target locus. *Mol. Cell. Biol.* **12:** 3365-3371.

Devine, P. L., Warren, J.A., Ward, B.G., McKenzie, I.F.C., and Layton, G.T. (1990). Glycosylation and the exposure of tumor-associated epitopes on mucins. *J. Tumor Marker Oncol.* **5:** 11-26.

Diamond, J.M. (1990). Old dead rats are valuable. *Nature* **347:** 334.

Dijan, P., and Green, H. (1989a). Vectorial expansion of the involucrin gene and the relatedness of the hominoids. *Proc. Natl. Acad. Sci. U.S.A.* **86:** 8447-8451.

Dijan, P., and Green, H. (1989b). The involucrin gene of the orangutan: generation of the late region as an evolutionary trend in the hominoids. *Mol. Biol. Evol.* **6:** 469-477.

Doetschman, T. C., Eistetter, H., Katz, M., Schmidt, W., and Kemler, R. (1985). The *in vitro* development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium. *J. Embryol. Exp. Morph.* **87:** 27-45.

Donehower, L. A., Harvey, M., Slagle, B.L., McArthur, M.J., Montgomery Jr., C.A., Butel, J.S., and Bradley, A. (1992). Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature* **356:** 215-221.

Dorin, J.R., Emslie, E., and Van Heyningen, V. (1990). Related calcium-binding proteins map to the same subregion of chromosome 1q and to an extended region of synteny on mouse chromosome 3. *Genomics* **8:** 420-426.

Dressler, D., and Potter, H. (1982). Molecular mechanisms in genetic recombination. *Ann. Rev. Biochem.* **51:** 727-761.

Duwe, A.K., and Ceriani, R.L. (1989). Human milk-fat-globule membrane derived mucin is a disulfide-linked heteromer. *Biochem. Biophys. Res. Commun.* **165:** 1305-1311.

Eckert, R. L., and Green, H. (1986). Structure and evolution of the human involucrin gene. *Cell* **46:** 583-589.

Eckhardt, A. E., Timpte, C.S., Abernethy, J.L., Zhao, Y., and Hill, R.L. (1991). Porcine submaxillary mucin contains a cystine-rich carboxyl-terminal domain in addition to a highly repetitive, glycosylated domain. *J. Biol. Chem.* **266:** 9678-9686.

Evans, M. J. and Kaufman, M.H. (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292:** 154-156.

Feinberg, A. P., and Vogelstein, B. (1984). A technique for radio-labelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **137:** 266-267.

Ferrari, S., Calabretta, B., DeRiel, J.K., Battini, R., Ghezzo, F., Lauret, E., Griffin, C., Emanuel, B.S., Gurrieri, F., and Baserga, R. (1987). Structural and functional analysis of a growth-related gene, the human calcyclin. *J. Biol. Chem.* **262:** 8325-8332.

Fiedler, F. (1987). Effects of secondary interactions on the kinetics of peptide and peptide ester hydrolysis by tissue kallikrein and trypsin. *Eur. J. Biochem.* **163:** 303-312.

Fontenot, J. D., Tjandra, N., Bu, D., Ho, C., Montelaro, R.C., and Finn, O.J. Biophysical characterization of one-, two-, and three-tandem repeats of the human mucin (MUC1) protein core. *Manuscript submitted.*

Frixen, U.H., Behrens, J., Sachs, M., Eberle, G., Voss, B., Warda, A., Löchner, D., and Birchmeier, W. (1991). E-cadherin-mediated cell-cell adhesion prevents invasiveness of human carcinoma cells. *J. Cell Biol.* **113:** 173-185.

Fung-Leung, W.-P., Schilman, M.W., Rahemtulla, A., Kündig, T.M., Vollenweider, M., Potter, J., van Ewijl, W., and Mak, T.W. (1991). CD8 is needed for development of cytotoxic T cells but not helper T cells. *Cell* **65:** 443-449.

Gendler, S. J., Burchell, J.M., Duhig, T., Lamport, D., White, R., Parker, M., and Taylor-Papadimitriou, J. (1987a). Cloning of partial cDNA encoding differentiation and tumor-associated mucin glycoproteins expressed by human mammary epithelium. *Proc. Natl. Acad. Sci. U.S.A.* **84:** 6060-6064.

Gendler, S., Burchell, J., Girling, A., Millis, R., Duhig, T., and Taylor-Papadimitriou, J. (1987b). Cloning the polymorphic gene for the mammary mucin abnormally glycosylated in carcinomas. In *Breast cancer: Scientific and Clinical Progress.* Kluwer Academic Publishers. pp. 112-126. Norwell, MA, USA.

Gendler, S. J., Taylor-Papadimitriou, J., Duhig, T., Rothbard, J., and Burchell, J. (1988). A highly immunogenic region of a human polymorphic epithelial mucin expressed by carcinomas is made up of tandem repeats. *J. Biol. Chem.* **263:** 12820-12823.

Gendler, S. J., Lancaster, C.A., Taylor-Papadimitriou, J., Duhig, T., Peat, N., Burchell, J., Pemberton, L., Lalani, E-N., and Wilson, D. (1990a). Molecular cloning and expression of human tumor-associated polymorphic epithelial mucin. *J. Biol. Chem.* **265:** 15286-15293.

Gendler, S.J., Cohen, E.P., Craston, A., Duhig, T., Johnstone, G., and Barnes, D. (1990b). The locus of the polymorphic epithelial mucin (PEM) tumour antigen on chromosome 1q21 shows a high frequency of alteration in primary human breast tumours. *Int. J. Cancer* **45:** 431-435.

Gendler, S.J., Spicer, A.P., Pemberton, L., Lancaster, C.A., Duhig, T., Peat, N., Taylor-Papadimitriou, J., and Burchell, J. (1991). Characterization and

evolution of an expressed hypervariable gene for a tumor-associated mucin, MUC-1. In *Breast Epithelial Antigens.* (R.L. Ceriani ed). Plenum, New York, USA.

Ginn s, E. I., Choudary, P.V., Tsuji, S., Martin, R., Stubblefield, B., Sawyer, J., Hozier, J., and Barranger, J.A. (1985). Gene mapping and leader polypeptide sequence of human glucocerebrosidase: Implications for Gaucher disease. *Proc. Natl. Acad. Sci. U.S.A.* **82:** 7101-7105.

Girling, A., Bartkova, J., Burchell, J., Gendler, S., Gillett, C., and Taylor-Papadimitriou, J. (1989). A core protein epitope of the polymorphic epithelial mucin detected by the monoclonal antibody SM-3 is selectively exposed in a range of primary carcinomas. *Int. J. Cancer* **43:** 1072-1076.

Gossler, A., Doetschman, T., Korn,R., Serfling, E., and Kemler, R. (1986). Transgenesis by means of blastocyst-derived embryonic stem cell lines. *Proc. Natl. Acad. Sci. U.S.A.* **83:** 9065-9069.

Gottschling, D. E., and Cech, T.R. (1984). Chromatin structure of the molecular ends of Oxytricha macronuclear DNA: phased nucleosomes and a telomeric complex. *Cell* **38:** 501-510.

Graham, F. L., and van der Eb, A.J. (1973). Transformation of rat cells by DNA of human adenovirus 5. *Virology* **52:** 456-467.

Graham, R. A., Seif, M.W., Aplin, J.D., Li, T.C., Cooke, I.D., Rogers, A.W., and Dockery, P. (1990). An endometrial factor in unexplained fertility. *Brit. Med. J.* **300:** 1428-1431.

Graur, D., Hide, W.A., and Li, W-H. (1991). Is the guinea-pig a rodent? *Nature* **351:** 649-652.

Gray, I. C., and Jeffreys, A.J. (1991). Evolutionary transience of hypervariable minisatellites in man and the primates. *Proc. R. Soc. Lond. B.* **243:** 241-253.

Green, E.L. (1981). Linkage, recombination and mapping. In *Genetics and Probability in Animal Breeding Experiments.* (E. Green, ed). Macmillan, New York, USA.

Griffiths, B., Matthews, D.J., West, L., Attwood, J., Povey, S., Swallow, D.M., Gum, J.R., and Kim, Y.S. (1990). Assignment of the polymorphic intestinal mucin gene (MUC2) to chromosome 11p15. *Ann. Hum. Genet.* **54:** 277-285.

Grusby, M. J., Johnson, R.S., Papaioannou, V.E., and Glimcher, L.H. (1991). Depletion of CD4+ T cells in major histocompatability complex class II-deficient mice. *Science* **253:** 1417-1420.

Gum, J. R., Byrd, J.C., Hicks, J.W.,Toribana, N.W., Lamport, D.T.A., and Kim, Y.S. (1989). Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evidence for genetic polymorphism. *J. Biol. Chem.* **264:** 6480-6487.

Gum, J. R., Hicks, J.W., Swallow, D.M., Lagace, R.L., Byrd, J.C., Lamport, D.T.A., Siddiki, B., and Kim, Y.S. (1990). Molecular cloning of cDNAs derived from a novel human intestinal mucin gene. *Biochem. Biophys. Res. Commun.* **171:** 407-415.

Gum, J. R. H., J.W., Lagace, R.E., Byrd, J.C., Toribara, N.W., Siddiki, B, Fearney, F.J., Lamport, D.T.A., and Kim, Y.S. (1991). Molecular cloning of rat intestinal mucin. *J. Biol. Chem.* **266:** 22733-22738.

Gum Jr, J.R., Hicks, J.W., Toribara, N.W., Rothe, E-M., Lagace, R.E., and Kim, Y.S. (1992). The human MUC2 intestinal mucin has cysteine-rich subdomains located both upstream and downstream of its central repetitive region. *J. Biol. Chem.* **267:** 21375-21383.

Hanisch, F.-G., Uhlenbruck, G., Peter-Katalinic, J., Egge, H., Dabrowski, J., Dabrowski, U. (1989). Structures of neutral O-linked polylactosaminoglycans on human skim milk mucins. *J. Biol. Chem.* **264:** 872-873.

Hareuveni, M., Gautier, C., Kieny, M-P., Wreschner, D.H., Chambon, P., and Lathe, R. (1990). Vaccination against tumor cells expressing breast cancer epithelial tumor antigen. *Proc. Natl. Acad. Sci. U.S.A.* **87:** 9498-9502.

Hasty, P., Rivera-Pérez, J., Chang, C., and Bradley, A. (1991a). Target frequency and integration pattern for insertion and replacement vectors in embryonic stem cells. *Mol. Cell. Biol.* **11:** 4509-4517.

Hasty, P., Ramírez-Solis, R., Krumlauf, R., and Bradley, A. (1991b). Introduction of a subtle mutation into the Hox-2.6 locus in embryonic stem cells. *Nature* **350:** 243-246.

Hauser, F., Gertzen, E.M., and Hoffmann, W. (1990). Expression of spasmolysin (FIM-A.1): an integumentary mucin from *Xenopus laevis. Exp. Cell Res.* **189:** 157-162.

Hauser, F., and Hoffmann, W. (1992). P-domains as shuffled cysteine-rich modules in integumentary mucin C.1 (FIM-C.1) from *Xenopus laevis. J. Biol. Chem.* **267:** 24620-24624.

Hayes, D. F., Sekine, H., Ohno, T., and Kufe, D. (1985). Detection of circulating plasma DF3 antigen levels in breast cancer patients. *J. Clin. Invest.* **75:** 1671-1678.

Hebert, J. M., Basilico, C., Goldfarb, M., Haub, O., and Martin, G.R. (1990). Isolation of cDNAs encoding four mouse FGF family members and characterization of their expression patterns during embryogenesis. *Dev. Biol.* **138:** 454-463.

Hennighausen, L. G., Steudle, A., and Sippel, A.E. (1982). Nucleotide sequence of cloned cDNA for mouse ε casein. *Eur. J. Biochem.* **126:** 569-572.

Heyderman, E., Steele, K., and Ormerod, M.G. (1979). A new antigen on the epithelial membrane: Its immunoperoxidase localization in normal and neoplastic tissues. *J. Clin. Pathol.* **32:** 35-.

Hilkens, J., Buijs, F., Hilgers, J., Hageman, P., Calafat, J., Sonnenberg, A., and van der Valk, M. (1984). Monoclonal antibodies against human milk-fat globule membranes detecting differentiation antigens of the mammary gland and its tumors. *Int. J. Cancer* **34:** 197-206.

Hilkens, J., Kroezen, V., Bonfrer, J.M., De Jong-Bakker, M., and Bruning P.F. (1986). MAM-6 antigen, a new serum marker for breast cancer monitoring. *Cancer Res.* **46:** 2586-2587.

Hilkens, J., and Buijs, F. (1988). Biosynthesis of MAM-6, an epithelial sialomucin. Evidence for involvement of a rare proteolytic cleavage step in the endoplasmic reticulum. *J. Biol. Chem.* **263:** 4215-4222.

Hoadley, M. F., Seif, M.W., and Aplin, J.D. (1990). Menstrual-cycle-dependent expression of keratan sulphate in human endometrium. *Biochem J.* **266:** 757-763.

Hoff, S. D., Matsushita, Y., Ota, D.M., Cleary, K.R., Yamori, T., Hakomori, S., and Irimura, T. (1989). Increased expression of sialyl-dimeric Le^X antigen in liver metastasis of human colorectal carcinoma. *Cancer Res.* **49:** 6883-6888.

Hoffmann, W. (1988). A new repetitive protein from *Xenopus laevis* skin highly homologous to pancreatic spasmolytic polypeptide. *J. Biol. Chem.* **263:** 7686-7690.

Hogan, B., Constantini, F., and Lacy, E. (1986). *Manipulating the Mouse Embryo: A Laboratory Manual.* Cold Spring Harbor, NY,

Hooper, M., Hardy, K., Handyside, A., Hunter, S., and Monk, M. (1987). HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature* **326:** 292-294.

Huebner, K., Palumbo, A.P., Isobe, M., Kozak, C.A., Monaco, S., Rovera, G., Croce, C.M., and Curtis, P.J. (1985). The α-spectrin gene is on chromosome 1 in mouse and man. *Proc. Natl. Acad. Sci. U.S.A.* **82:** 3790-3793.

Hull, S., Bright, A., Carraway, K., Abe, M., Hayes, D., and Kufe, D. (1989). Oligosaccharide differences in the DF3 sialomucin antigen from normal human milk and the BT20 human breast carcinoma cell line. *Cancer Comm.* **1:** 261-267.

Irimura, T., Carlson, D.A., Price, J., Yamori, T., Giavazzi, R., Ota, D.M., and Cleary, K.R. (1989). Differential expression of a sialoglycoprotein with an approximate molecular weight of 900,000 on metastatic human colon carcinoma cells growing in culture and in tumor tissues. *Cancer Res.* **48:** 2353-2360.

Irimura, T., McIsaac, A.M., Carlson, D.A., Yagita, M., Grimm, E.A., Menter, D.G., Ota, D.M., and Cleary, K.R. (1990). Soluble factor in normal tissues that stimulates high molecular weight sialoglycoprotein production by colon carcinoma cells. *Cancer Res.* **50:** 3331-3338.

Irimura, T., Matsushita, Y., Hoff, S.D., Yamori, T., Nakamori, S., Frazier, M.L., Giacco, G.G., Cleary, K.R., and Ota, D.M. (1991). Ectopic expression of mucins in colorectal cancer metastasis. *Seminars in Cancer Biol.* **2:** 129-139.

Jarman, A.P., and Wells, R.A. (1989). Hypervariable minisatellites: recombinators or innocent bystanders? *Trends Genet.* **5:** 367-371.

Jeanotte, L., Ruiz, J.C., and Robertson, E.J. (1991). Low level of Hox1.3 gene expression does not preclude the use of promoterless vectors to generate a targeted gene disruption. *Mol. Cell. Biol.* **11:** 5578-5585.

Jeffreys, A. J., Wilson, V., and Thein, S.L. (1985a). Hypervariable 'minisatellite' regions in human DNA. *Nature* **314:** 67-73.

Jeffreys, A. J., Wilson, V., and Thein, S.L. (1985b). Individual-specific 'fingerprints' of human DNA. *Nature* **316:** 76-79.

Jeffreys, A.J., Royle, N.J., Wilson, V., and Wong, Z. (1988a). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332:** 278-281

Jeffreys, A. J., Wilson, V., Neumann, R., and Keyte, J. (1988b). Amplification of human minisatellites by the polymerase chain reaction: towards DNA fingerprinting of single cells. *Nucl. Acids. Res.* **16:** 10953-10971.

Jeffreys, A.J., Neumann, R., and Wilson, V. (1990). Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* **60:** 473-485.

Jeffreys, A. J., MacLeod, A., Tamaki, K., Neil, D.L., and Monckton, D.G. (1991). Minisatellite repeat coding as a digital approach to DNA typing. *Nature* **354:** 204-209.

Jentoft, N. (1990). Why are proteins O-glycosylated? *Trends Biochem.* **15:** 291-294.

Johnson, R. S., Sheng, M., Greenberg, M.E., Kolodner, R.D., Papaioannou, V.E., and Spiegelman, B.M. (1989). Targeting of nonexpressed genes in embryonic stem cells via homologous recombination. *Science* **245:** 1234-1236.

Jukes, T.H., and Cantor, C.R. (1969). Evolution of protein molecules. pp 21-32. In *Mammalian protein metabolism III.* (H.N. Munro, ed). Academic press, New York, USA.

Kafatos, F.C.A., Efstratiadis, B.G., Forget, B.G., and Weissman, S.M. (1977). Molecular evolution of human and rabbit β-globin mRNAs. *Proc. Natl. Acad. Sci. U.S.A.* **74:** 5618-5622.

Kalnins, A., Otto, K., Rüther, U., and Müller-Hill, B. (1983). Sequence of the LacZ gene of *Escherichia coli. EMBO J.* **2:** 593-598.

Karlsson, S., Swallow, D.M., Griffiths, B., Corney, G., Kopkinson, D.A., Dawnay, A., and Cantron, J.P. (1983). A genetic polymorphism of a human urinary mucin. *Ann. Hum. Genet.* **47:** 263-269.

Kaufman, D. L., and Evans, G.A. (1990). Restriction endonuclease cleavage at the termini of PCR products. *BioTechniques* **9:** 304-306.

Kelly, R., Gibbs, M., Collick, A., and Jeffreys, A.J. (1991). Spontaneous mutation at hypervariable mouse minisatellite locus Ms6-hm: flanking DNA sequence and analysis of germline and early somatic mutation events. *Proc. R. Soc. Lond. B.* **245:** 235-245.

Keydar, I., Chen, L., Karby, S., Weiss, F.R., Delerea, J., Radu, M., Chaitchik, S., and Brenner, H.J. (1979). Establishment and characterization of a cell line of human breast carcinoma origin. *Eur. J. Cancer* **15:** 659-670.

Keydar, I., Chou, C.S., Hareuveni, M., Tsarfaty, I., Sahar, E., Selzer, G., Chaitchik, S., and Hizi, A. (1989). Production and characterization of monoclonal antibodies identifying breast tumor-associated antigens. *Proc. Natl. Acad. Sci. U.S.A.* **86:** 1362-1366.

Kim, H.-S., and Smithies, O. (1988). Recombinant fragment assay for gene targeting based on the polymerase chain reaction. *Nucl. Acids Res.* **16:** 8887-8903.

Kimber, S. J., and Lindenberg, S. (1990). Hormonal control of a carbohydrate epitope involved in implantation in mice. *J. Reprod. Fert.* **89:** 13-21.

Kingsmore, S.F., Watson, M.L., Howard, T.A., and Seldin, M.F. (1989) A 6000 kb segment of chromosome 1 is conserved in human and mouse. *EMBO J.* **8:** 4073-4080.

Kingsmore, S.F., Watson, M.L., and Seldin, M.F. Genetic mapping of the high affinity nerve growth factor gene, *Ntrk1*, to mouse chromosome 3. *Genomics*, in press.

Kioussis, D., Wilson, F., Daniels, C., Leveton, C., Taverne, J., and Playfair, J.H.L. (1987). Expression and rescuing of a cloned human necrosis factor gene using an EBV-based shuttle cosmid vector. *EMBO J.* **6:** 355-361.

Koop, B. F., Goodman, M., Xu, P., Chan, K., and Slightom, J.L. (1986). Primate η-globin DNA sequences and man's place among the great apes. *Nature* **319:** 234-238.

Korhonen, T. K., Valtonen, M.V., and Parkkinen, J. (1985). Serotypes, hemolysin production and receptor recognition of *Escherichia coli* strains associated with neonatal sepsis and meningitis. *Infect. Immun.* **48:** 486-491.

Kovarik, A., Peat, N., Wilson, D., Gendler, S.J., and Taylor-Papadimitriou, J. (1993). Analysis of the tissue-specific promoter of the MUC1 gene. *J. Biol. Chem.* **268:** 9917-9926.

Kuehn, M. R., Bradley, A., Robertson, E.J., and Evans, M.J. (1987). A potential animal model for Lesch-Nyan syndrome through introduction of HPRT mutations into mice. *Nature* **326:** 295-298.

Kufe, D., Inghirami, G., Abe, M., Hayes, D., Justi-Wheeler, H., and Schlom, J. (1984). Differential activity of a novel monoclonal antibody (DF3) with human malignant versus benign breast tumours. *Hybridoma* **3:** 223-232.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227:** 680-685.

Lalani, E.-N., Berdichevsky, F., Boshell, M., Shearer, M., Wilson, D., Stauss, H., Gendler, S.J., and Taylor-Papdimitriou, J. (1991). Expression of the gene coding for a human mucin in mouse mammary tumor cells can affect their tumorigenicity. *J. Biol. Chem.* **266:** 15420-15426.

Lan, M. S., Batra, S.K., Qi, W-N., Metzgar, R.S., Hollingsworth, M.A. (1990). Cloning and sequencing of a human pancreatic tumor mucin cDNA. *J. Biol. Chem* **265:** 15294-15299.

Lancaster, C. A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitriou, J., and Gendler, S.J. (1990). Structure and expression of the human polymorphic epithelial mucin gene: an expressed VNTR unit. *Biochem. Biophys. Res. Commun.* **173:** 1019-1029.

Le Mouellic, H., Lallemand, Y., and Brûlet, P. (1992). Homeosis in the mouse induced by a null mutation in the Hox-3.1 gene. *Cell* **69:** 251-264.

Lee, K.-F., Li, E., Huber, J., Landis, S.C., Sharpe, A.H., Chao, M.V., and Jaenisch, R. (1992). Targeted mutation of the gene encoding the low affinity NGF receptor p75 leads to deficits in the peripheral sensory nervous system. *Cell* **69:** 737-749.

Letsou, A., and Liskay, R.M. (1987). Effect of the molecular nature of mutation on the efficiency of intrachromosomal gene conversion in mouse cells. *Genetics* **117:** 759-769.

Levinson, G., and Gutman, G.A. (1987). Slipped strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4:** 203-221.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methlytransferase gene results in embryonic lethality. *Cell* **69:** 915-926.

Li, M.-L., Aggeler, J., Farson, D.A., Hatier, C., Hassell, J., and Bissell, M.J. (1987). Influence of a reconstituted basement membrane and its components on casein gene expression and secretion in mouse mammary epithelial cells. *Proc. Natl. Acad. Sci. U.S.A.* **84:** 136-140.

Li, W-H., Wu, C-I., and Luo, C-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2:** 150-174.

Li, W-H., Tanimura, M., and Sharp, P.M. (1987). An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* **25:** 330-342.

Li, W-H., Gouy, M., Sharp, P.M., O'hUigin, C., and Yang, Y-W. (1990). Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. *Proc. Natl. Acad. Sci. U.S.A.* **87:** 6703-6707.

Ligtenberg, M. J. L., Vos, H.L., Gennissen, A.M.C., and Hilkens, J. (1990). Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene encoding splice variants with alternative amino termini. *J. Biol. Chem.* **265:** 5573-5578.

Ligtenberg, M. J. L., Gennissen, A.M.C., Vos, H.L., and Hilkens, J. (1991). A single nucleotide polymorphism in an exon dictates allele dependent differential splicing of episialin mRNA. *Nucl. Acids Res.* **19:** 297-301.

Ligtenberg, M. J. L., Kruijshaar, L., Buijs, F., van Meijer, M., Litvinov, S.V., and Hilkens, J. (1992a). Cell-associated episialin is a complex containing two proteins derived from a common precursor. *J. Biol. Chem.* **267:** 6171-6177.

Ligtenberg, M.J.L., Buijs, F., Vos, H.L., and Hilkens, J. (1992b). Suppression of cellular aggregation by high levels of episialin. *Cancer Res.* **52:** 2318-2324.

Little, P. F. (1985). Choice and use of cosmid vectors. In *DNA cloning: a practical approach.* (D.M. Glover, ed). Oxford, U.K., IRL Press Ltd.

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

Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual,* Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA.

Mansour, S. L., Thomas, K.R., and Capecchi, M.R. (1988). Disruption of the proto-oncogene int-2 in mouse embryo-derived stem cells: a general strategy for targeting mutations to non-selectable genes. *Nature* **336:** 348-352.

Mansour, S. L., Thomas, K.R., Deng, C., and Capecchi, M.R. (1990). Introduction of a lacZ reporter gene into the mouse int-2 locus by homologous recombination. *Proc. Natl. Acad. Sci. U.S.A.* **87:** 7688-7692.

Marchuk, D., Drumm, M., Saulino, A., and Collins, F.S. (1991). Construction of T-vectors, a rapid and general system for direct cloning of unmodified PCR products. *Nucl. Acids Res.* **19:** 1154.

Martin, G. R., and Evans, M.J. (1975). Differentiation of clonal lines of teratocarcinoma cells: Formation of embryoid bodies *in vitro. Proc. Natl. Acad. Sci. U.S.A.* **72:** 1441-1445.

Martin, G. R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc. Natl. Acad. Sci. U.S.A.* **78:** 7634-7638.

Matzuk, M. M., Finegold, M.J., Su, J-G. J., Hsueh, A.J.W., and Bradley, A. (1992). α-Inhibin is a tumour-suppressor gene with gonadal specificity in mice. *Nature* **360:** 313-319.

McBurney, M. W., Sutherland, L.C., Adra, C.N., Leclair, B., Rudnicki, M.A., and Jardine, K. (1991). The mouse PGK-1 gene promoter contains an upstream activator sequence. *Nucl. Acids Res.* **19:** 5755-5761.

McGuire, E. J., and Roseman, S. (1967). Enzymatic synthesis of the protein-hexosamine linkage in sheep submaxillary mucin. *J. Biol. Chem.* **242:** 3745-3747.

McMahon, A. P., and Bradley, A. (1990). The Wnt-1 (Int-1) proto-oncogene is required for the development of a large region of the mouse brain. *Cell* **62:** 1073-1085.

Metzgar, R. S., Gaillard, M.T., Levine, S.J., Tuck, F.L., Bossen, E.H., and Borowitz, M.J. (1982). Antigens of human pancreatic adenocarcinoma cells defined by murine monoclonal antibodies. *Cancer Res.* **42:** 601-608.

Metzgar, R. S., Rodriguez, N., Finn, O.J., lan, M.S., Daasch, V.N., Fernsten, P.D., Meyers, W.C., Sindelar, W.F., and Sandler, R.S. (1984). Detection of a pancreatic cancer-associated antigen (DU-PAN-2 antigen) in serum and ascites of patients with adenocarcinoma. *Proc. Natl. Acad. Sci. U.S.A.* **81:** 5242-5246.

Middleton-Price, H., Gendler, S., and Malcolm, S. (1988). Close linkage of PUM and SPTA within chromosome band 1q21. *Ann. Hum. Genet.* **52:** 273-278.

Miller, C. C. J., McPheat, J.C., and Potts, W.J. (1992). Targeted integration of the Ren-1D locus in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* **89:** 5020-5024.

Mole, S. E., Iggo, R.D., and Lane, D.P. (1989). Using the polymerase chain reaction to modify expression plasmids for epitope mapping. *Nucl. Acids Res.* **17:** 3319.

Mombaerts, P., Clarke, A. R., Hooper, M.L., and Tonegawa, S. (1991). Creation of a large deletion at the T-cell antigen receptor β-subunit locus in mouse embryonic stem cells by gene targeting. *Proc. Natl. Acad. Sci. U.S.A.* **88:** 3084-3087.

Morrissey, J. H. (1981). Silver stain for proteins in polyacrylamide gels: a modified procedure with enhanced uniform activity. *Anal. Biochem.* **117:** 307-310.

Mortensen, R. M., Zubiaur, M., Neer, E.J., and Seidman, J.G. (1991). Embryonic stem cells lacking a functional inhibitory G-protein subunit (αi2) produced by gene targeting of both alleles. *Proc. Natl. Acad. Sci. U.S.A.* **88:** 7036-7040.

Moseley, W.S., and Seldin, M.F. (1989a). Definition of mouse chromosome 1 and 3 gene linkage groups that are conserved on human chromosome 1: evidence that a conserved linkage group spans the centromere of human chromosome 1. *Genomics* **5:** 899-905.

Moseley, W.S., Watson, M.L., Kingsmore, S.F., and Seldin, M.F. (1989b) CD1 defines conserved linkage group border between human chromosomes 1 and mouse chromosomes 1 and 3. *Immunogenetics* **30:** 378-382.

Moss, L., Greenwalt, D., Cullen, B., Dinh, N., Ranken, R., and Parry, G. (1988). Cell-to-cell heterogeneity in the expression of carbohydrate based epitopes of a mucin-type glycoprotein on the surface of human mammary carcinoma cells. *J. Cell Physiol.* **137:** 310-320.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. *Mol. Biol. Evol.* **3:** 418-426.

Oakey, R.J., Watson, M.L., and Seldin, M.F. (1992a). Construction of a physical map on mouse and human chromosome 1: comparison of 13 Mb of mouse and 11 Mb of human DNA. *Hum. Molec. Genet.* **1:** 613-620.

Oakey, R.J., Howard, T.A., Hogarth, P.M., Tani, K., and Seldin, M.F. (1992b). Chromosomal mapping of the high affinity Fcγ receptor gene. *Immunogenetics* **35:** 279-282.

Ojakian, G.K., and Schwimmer, R. (1988). The polarized distribution of an apical cell surface glycoprotein is maintained by interactions with the cytoskeleton of Madin-Darby Canine Kidney cells. *J. Cell Biol.* **107:** 2377-2387.

Ozawa, M., Ringwald, M., and Kemler, R. (1990). Uvomorulin-catenin complex formation is regulated by a specific domain in the cytoplasmic region of the cell adhesion molecule. *Proc. Natl. Acad. Sci. U.S.A.* **87:** 4246-4250.

Parry, G., Cullen, B., Kaetzel, C.S., Kramer, R., and Moss, L. (1987). Regulation of differentiation and polarized secretion in mammary epithelial cells in culture. *J. Cell Biol.* **105:** 2043-2051.

Parry, G., Beck, J.C., Moss, L., Bartley, J., and Ojakian, G.K. (1990). Determination of apical membrane polarity in mammary epithelial cell cultures: the role of cell-cell, cell-substratum, and membrane-cytoskeleton interactions. *Exp. Cell Res.* **188:** 302-311.

Parry, G., Li, J., Stubbs, J., Bissell, M.J., Schmidhauser, C., Spicer, A.P., and Gendler, S.J. (1992). Studies of Muc-1 mucin expression and polarity in the mouse mammary gland demonstrate developmental regulation of Muc-1 glycosylation and establish the hormonal basis for mRNA expression. *J. Cell Sci.* **101:** 191-199.

Patek, C. E., Kerr, J.B., Gosden, R.G., Jones, K.W., Hardy, K., Muggleton-Harris, A.L., Handyside, A.H., Whittingham, D.G., and Hooper, M.L. (1991). Sex chimaerism, fertility and sex determination in the mouse. *Development* **113:** 311-325.

Patton, S., and Huston, G.E. (1986). A method for isolation of milk fat globules. *Lipids* **21:** 170-174.

Patton, S., Huston, G.E., Jennes, R., and Vaucher, Y. (1989). Differences between individuals in high-molecular weight glycoproteins from mammary epithelia of several species. *Biochim. Biophys. Acta.* **980:** 333-338.

Patton, S., and Patton, R.S. (1990). Genetic polymorphism of PAS-1, the mucin-like glycoprotein of bovine milk fat globule membrane. *J. Dairy Sci.* **73:** 3567-3574.

Patton, S., and Muller, L.D. (1992). Genetic polymorphism of the epithelial mucin, PAS-1, in milk samples from the major dairy breeds. *J. Dairy Sci.* **75:** 863-867.

Pease, S., and Williams, R.L. (1990). Formation of germ line chimaeras from embryonic stem cells maintained with recombinant leukaemia inhibitory factor. *Exp. Cell. Res.* **190:** 209-211.

Peat, N., Gendler, S.J., Lalani, E-N., Duhig, T., and Taylor-Papadimitriou, J. (1992). Tissue-specific expression of a human polymorphic epithelial mucin (MUC1) in transgenic mice. *Cancer Res.* **52:** 1954-1960.

Pemberton, L., Taylor-Papadimitriou, J., and Gendler, S.J. (1992). Antibodies to the cytoplasmic domain of the MUC1 mucin show conservation throughout mammals. *Biochem. Biophys. Res. Commun.* **185:** 167-175.

Philpott, K.L., Viney, J.L., Kay, G., Rastan, S., Gardiner, E.M., Chae, S., Hayday, A.C., and Owen, M.J. (1992) Lymphoid development in mice congenitally lacking T cell receptor αβ-expressing cells. *Science* **256:** 1448-1452.

Pigman, W., Moschera, J., Weis, M., and Tettamanti, G. (1973). The occurrence of repetitive glycopeptide sequences in bovine submaxillary glycoprotein. *Eur. J. Biochem.* **32:** 148-.

Porchet, N., Van Cong., N., Dufosse, J., Audie, J.P., Guyonnet-Duperat, V., Gross, M.S., Denis, C., Degand, P., Bernheim, A., and Aubert, J.P. (1991). Molecular cloning and chromosomal localization of a novel human tracheo-bronchial mucin cDNA containing tandemly repeated sequences of 48 base pairs. *Biochem. Biophys. Res. Commun.* **175:** 414-422.

Poustka, A., and Lehrach, H. (1985). Genetic approaches to the cloning, modification and characterization of cosmid clones and clone libraries. In *DNA cloning: a practical approach.* (D.M. Glover, ed). Oxford, U.K., IRL Press Ltd.

Price, M., Edwards, S., Owainati, A., and Robins, A. (1985). Multiple epitopes on a breast carcinoma associated antigen. *Int. J. Cancer* **36:** 567-572.

Probst, J. C., Gertzen, E.M., and Hoffmann, W. (1990). An integumentary mucin (FIM-B.1) from *Xenopus laevis* homologous with von Willebrand factor. *Biochemistry* **29:** 6240-6244.

Probst, J. C., Hauser, F., Joba, W., and Hoffmann, W. (1992). The polymorphic integumentary mucin B.1 from *Xenopus laevis* contains the short consensus repeat. *J. Biol. Chem.* **267:** 6310-6316.

Rayssiguier, C., Thaler, D.S., and Radman, M. (1989). The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature* **342:** 396-401.

Robertson, E., Bradley, A., Kuehn, M., and Evans, M. (1986). Germ-line transmission of genes introduced into cultured pluripotential cells by retroviral vector. *Nature* **323:** 445-448.

Royle, N. J., Clarkson, R.E., Wong, Z., and Jeffreys, A.J. (1988). Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. *Genomics* **3:** 352-360.

Rudnicki, M. A., Braun, T., Hinuma, S., and Jaenisch, R. (1992). Inactivation of MyoD in mice leads to up-regulation of the myogenic HLH gene Myf-5 and results in apparently normal muscle development. *Cell* **71:** 383-390.

Saga, Y., Yagi, T., Sakakura, T., and Aizawa, S. (1992). Mice develop normally without tenascin. *Genes Dev.* **6:** 1821-1831.

Sanes, J. R., Rubenstein, J.L.R., and Nicolas, J.-F, (1986). Use of a recombinant retrovirus to study post-implantation cell lineage in mouse embryos. *EMBO J.* **5:** 3133-3142.

Sanger, F., Nicklen, S., Coulson, AR (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74:** 5463-5467.

Schimizu, M., and Yamauchi, K. (1982). Isolation and characterisation of mucin-like glycoprotein in human milk fat globule membrane. *J. Biochem. (Tokyo)* **91:** 515-519.

Schmidt, A. M. A., Herterich, S.U., and Krauss, G. (1991). A single-stranded DNA binding protein from *S. cerevisiae* specifically recognizes the T-rich strand of the core sequence of ARS elements and discriminates against mutant sequences. *EMBO J.* **10:** 981-985.

Schroten, H., Lethen, A., Hanisch., F.G., Plogmann, R., Hacker, J., Nobis-Bosch, R., and Wahn, V. (1992). Inhibition of adhesion of S-fimbriated *Escherichia coli* to epithelial cells by meconium and feces of breast-fed and formula-fed newborns: mucins are the major inhibitory component. *J. Pediat. Gastroent. Nutrit.* **15:** 150-158.

Schwartzberg, P. L., Robertson, E.J., and Goff, S.P. (1990). Targeted gene disruption of the endogenous c-abl locus by homologous recombination with DNA encoding a selectable fusion protein. *Proc. Natl. Acad. Sci. U.S.A.* **87:** 3210-3214.

Schwartzberg, P. L., Stall, A.M., Hardin, J.D., Bowdish, K.S., Humaran, T., Boast, S., Harbison, M.H., Robertson, E.J., and Goff, S.P. (1991). Mice homozygous for the abl^{m1} mutation show poor viability and depletion of selected B and T cell populations. *Cell* **65:** 1165-1172.

Sears, D.W., Osman, N., Tate, B., McKenzie, I.F.C., and Hogarth, P.M. (1990). Molecular cloning and expression of the mouse high affinity Fc receptor for IgG. *J. Immunol.* **144:** 371-378.

Seldin, M.F., Morse III, H.C., Reeves, J.P., Scribner, C.L., LeBoeuf, R.C., and Steinberg, A.D. (1988). Genetic analysis of autoimmune *gld* mice: 1. Identification of a restriction fragment length polymorphism closely linked to the *gld* mutation within a conserved linkage group. *J. Exp. Med.* **167:** 688-693.

Shen, P., and Huang, H.V. (1986). Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics* **112:** 441-457.

Sheng, Z., Hull, S.R., and Carraway, K.L. (1990). Biosynthesis of the cell surface sialomucin complex of ascites 13762 rat mammary adenocarcinoma cells from a high M_r precursor. *J. Biol. Chem.* **265:** 8505-8510.

Sheng, Z., Wu, K., Carraway, K.L., and Freigien, N. (1992). Molecular cloning of the transmembrane component of the 13762 mammary adenocarcinoma sialomucin complex: a new member of the epidermal growth factor superfamily. *J. Biol. Chem.* **267:** 16341-16346.

Siddiqui, J., Abe, M., Hayes, D., Shani, E., Yunis, E., and Kufe, D. (1988). Isolation and sequencing of a cDNA coding for the human DF3 breast carcinoma-associated antigen. *Proc. Natl. Acad. Sci. U.S.A.* **85:** 2320-2323.

Simon, M., Philips, M., Green, H., Stroh, H., Glatt, K., Bruns, G., and Latt, S.A. (1989). Absence of a single repeat from the coding region of the human involucrin gene leading to RFLP. *Am. J. Hum. Genet.* **45:** 910-916.

Simon, M., Phillips, M., and Green, H. (1991). Polymorphism due to variable number of repeats in the human involucrin gene. *Genomics* **9:** 576-580.

Smith, A. G., and Hooper, M.L. (1987). Buffalo rat liver cells produce a diffusible activity which inhibits the differentiation of embryonal carcinoma and embryonic stem cells. *Dev. Biol.* **121:** 1-9.

Smith, R. A., Seif, M.W., Rogers, A.W., Li, T.C., Dockery, P., Cooke, I.D., and Aplin, J.D. (1989). The endometrial cycle: the expression of a scretory

component correlated with the luteinizing hormone peak. *Human Reproduction* **4:** 236-242.

Smithies, O., Gregg, R.G., Boggs, S.S., Koralewski, M.A., and Kucherlapati, R.S. (1985). Insertion of DNA sequences into the human chromosome β-globin locus by homologous recombination. *Nature* **317:** 230-234.

Snouwaert, J. N., Brigman, K.K., Latour, A.M., Malouf, N.N., Boucher, R.C., Smithies, O., and Koller, B.H. (1992). An animal model for cystic fibrosis made by gene targeting. *Science* **257:** 1083-1088.

Soriano, P., Montgomery, C., Geske, R., and Bradley, A. (1991). Targeted disruption of the c-src proto-oncogene leads to osteopetrosis in mice. *Cell* **64:** 693-702.

Sorimachi, H., Emori, Y., Kawasaki, H., Kitajima, K., Inoue, S., Suzuki, K., and Inoue, Y. (1988). Molecular cloning and characterization of cDNAs coding for apo-polysialoglycoprotein of rainbow trout eggs. Multiple mRNA species transcribed from multiple genes contain diverged numbers of exact 39-base (13 amino acid) repeats. *J. Biol. Chem.* **263:** 17678-17688.

Spicer, A. P., Parry, G., Patton, S., and Gendler, S.J. (1991). Molecular cloning and analysis of the mouse homologue of the tumor-associated mucin, MUC1, reveals conservation of potential O-glycosylation sites, transmembrane, and cytoplasmic domains and a loss of minisatellite-like polymorphism. *J. Biol. Chem.* **266:** 15099-15109.

Stanton, B. R., Reid, S.W., and Parada, L.F. (1990). Germ line transmission of an inactive N-myc allele generated by homologous recombination in mouse embryonic stem cells. *Mol. Cell. Biol.* **10:** 6755-6758.

Stanton, B. R., Perkins, A.S., Tessarollo, L., Sassoon, D.A., and Parada, L.F. (1992). Loss of N-myc function results in embryonic lethality and failure of the epithelial component of the embryo to develop. *Genes Dev.* **6:** 2235-2247.

Stein, P. L., Lee, H-M., Rich, S., and Soriano, P. (1992). pp59^{fyn} mutant mice display differential signalling in thymocytes and peripheral T cells. *Cell* **70:** 741-750.

Steinmetz, M., Uematsu, Y., and Lindahl, K.F. (1987). Hotspots of homologous recombination in mammalian genomes. *Trends Genet.* **3:** 7-10.

Stephan, W. (1989). Tandem-repetitive noncoding DNA: forms and forces. *Mol. Biol. Evol.* **6:** 198-212.

Stewart, I. (1991). Granulated metrial gland cells: pregnancy specific leukocytes? *J. Leukocyte Biol.* **50:** 198-207.

Stubbs, J. D., Lekutis, C., Singer, K.L., Bui, A., Yuzuki, D., Srinivasan, U., and Parry, G. (1990). cDNA cloning of a mouse mammary epithelial cell surface protein reveals the existence of epidermal growth factor-like domains linked to factor VIII-like sequences. *Proc. Natl. Acad. Sci. U.S.A.* **87:** 8417-8421.

Swallow, D. M., Griffiths, B., Bramwell, M., Wiseman, G., and Burchell, J. (1986). Detection of the urinary 'PUM' polymorphism by the tumour-binding monoclonal antibodies Ca1, Ca2, Ca3, HMFG1 and HMFG2. *Disease Markers* **4:** 247-254.

Swallow, D. M., Gendler, S., Griffiths, B., Corney, G., Taylor-Papadimitriou, J., and Bramwell, M.E. (1987a). The human tumour-associated epithelial mucins are coded by an expressed hypervariable gene locus PUM. *Nature* **327:** 82-84.

Swallow, D., Gendler, S., Griffiths, B., Kearney, A., Povey, S., Sheer, D., Palmer, R., and Taylor-Papadimitriou, J. (1987b). The hypervariable gene locus PUM, which codes for the tumor associated epithelial mucins, is located on chromosome 1, within the region 1q21-24. *Ann. Hum. Gent.* **51:** 289-294.

Tani, K., Fujii, H., Nagata, S., and Miwa, S. (1988). Human liver type pyruvate kinase: Complete amino acid sequence and the expression in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* **85:** 1792-1795.

Taylor-Papadimitriou, J., Peterson, J.A., Arklie, J., Burchell, J., Ceriani, R.L., and Bodmer, W.F. (1981). Monoclonal antibodies to epithelium-specific components of the human milk fat globule membrane: production and reaction with cells in culture. *Int. J. Cancer* **28:** 17-21.

Taylor-Papadimitriou, J. (1991). Report on the first international workshop on carcinoma-assocaiated mucins. *Int. J. Cancer* **49:** 1-5.

Taylor-Papadimitriou, J., Allen, D., Granowska, M., Peat, N., Duhig, T., Spicer, A., Burchell, J., and Gendler, S.J. (1992). Molecular structure and clinical applications of a cancer-associated mucin. In *Ovarian Cancer 2.* (F. Sharp, P. Mason and W. Creasman, eds). Chapman and Hall Medical, pp 39-50.

te Riele, H., Maandag, E.R., Clarke, A., Hooper, M., and Berns, A. (1990). Consecutive inactivation of both alleles of the pim-1 proto-oncogene by homologous recombination in embryonic stem cells. *Nature* **348:** 649-651.

te Riele, H., Robanus Maandag, E., and Berns, A. (1992). Highly efficient gene targeting in embryonic stem cells through homologous recombination with isogenic DNA constructs. *Proc. Natl. Acad. Sci. U.S.A.* **89:** 5128-5132.

Thomas, K. R., and Capecchi, M.R. (1987). Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell* **51:** 503-512.

Thomas, K. R., and Capecchi, M.R. (1990). Targeted disruption of the murine int-1 proto-oncogene resulting in severe abnormalities in midbrain and cerebellar development. *Nature* **346:** 847-850.

Thomas, K.R., Deng, C., and Capecchi, M.R. (1992). High fidelity gene targeting in embryonic stem cells by using sequence replacement vectors. *Mol. Cell. Biol.* **12:** 2919-2923.

Timpte, C. S., Eckhardt, A.E., Abernethy, J.L., and Hill, R.L. (1988). Porcine submaxillary gland apomucin contains tandemly repeated, identical sequences of 81 residues. *J. Biol. Chem.* **263:** 7686-7690.

Toribara, N. W., Gum, J.R., Culhane, P.J., lagace, R.E., Hicks, J.W., Peterson, G.W., and Kim, Y.S. (1991). MUC-2 human small intestinal mucin gene structure: repeated arrays and polymorphism. *J. Clin. Invest.* **88:** 1005-1013.

Toribara, N. W., Roberton, A.M., Ho, S.B., Kuo, W-L., Gum, E., Gum, J.R., Byrd, J.C., Siddiki, B., and Kim, Y.S. (1993). Human gastric mucin: Identification of a unique species by expression cloning. *J. Biol. Chem* **268:** 5879-5885.

Tsarfaty, I., Chatchik, S., Hareuveni, M., Horev, J., Hizi, A., Wreschner, D.H., and Keydar, I. (1988). H23 monoclonal antibodies recognize a breast cancer tumor associated antigen: Clinical and molecular studies. In *Breast Cancer Immunodiagnosis and Immunotherapy.* pp 161-169. Plenum, New York, USA.

Tsarfaty, I., Hareuveni, M., Horev, J., Zaretsky, J., Weiss, M., Jeltsch, J.M., Garnier, J.M., lathe, R., Keydar, I., and Wreschner, D.H. (1990). Isolation and characterization of an expressed hypervariable gene coding for a breast-cancer-associated antigen. *Gene* **93:** 313-318.

Tseng, H., and Green, H. (1988). Remodelling of the involucrin gene during primate evolution. *Cell* **54:** 491-496.

Tseng, H., and Green, H. (1989). The involucrin gene of the owl monkey: origin of the early region. *Mol. Biol. Evol.* **6:** 460-468.

Tybulewicz, V. L. J., Crawford, C.E., Jackson, P.K., Bronson, R.T., and Mulligan, R.C. (1991). Neonatal lethality and lymphopenia in mice with a homozygous disruption of the c-abl proto-oncogene. *Cell* **65:** 1153-1163.

Tybulewicz, V. L. J., Tremblay, M.L., LaMarca, M.E., Willemsen, R., Stubblefield, B.K., Winfield, S., Zablocka, B., Sidransky, E., Martin, B.M., Huang, S.P., Mintzer, K.A., Westphal, H., Mulligan, R.C., and Ginns, E.I. (1992). Animal model of Gaucher's disease from targeted disruption of the mouse glucocerebrosidase gene. *Nature* **357:** 407-410.

Umesono, K., Murakami, K.K., Thompson, C.C., and Evans, R.M. (1991). Direct repeats as selective response elements for the thyroid hormone, retinoic acid, and vitamin D_3 receptors. *Cell* **65:** 1255-1266.

Vaidya, A. B., Lasfargues, E.Y., Sheffield J.B., and Coutinho, W.G. (1978). Murine mammary tumor virus (MuMTV) infection of an epithelial cell line established from C57Bl/6 mouse mammary glands. *Virology* **90:** 12-22.

Valancius, V., and Smithies, O. (1991). Testing an "In-Out" targeting procedure for making subtle genomic modifications in mouse embryonic stem cells. *Mol. Cell. Biol.* **11:** 1402-1408.

Van Cong, N., Aubert, J.P., Gross, M.S., Porchet, N., Degand, P., and Frezal, J. (1990). Assignment of human tracheobronchial mucin gene(s) to 11p15 and a tracheobronchial mucin-related sequence to chromosome 13. *Hum. Genet.* **86:** 167-172.

Vincek, V., Kawaguchi, H., Mizuno, K., Zaleska-Rutezynska, Z., Kasahara, M., Forejt, F., and Klein, J. (1989). Linkage map of mouse chromosome 17: Localization of 27 new DNA markers. *Genomics* **5:** 773-786.

Vleminckx, K., Vakaet, L., Jr., Mareel, M., Fiers, W., and van Roy, F. (1991). Genetic manipulation of E-cadherin expression by epithelial tumor cells reveals an invasion suppressor role. *Cell* **66:** 107-119.

von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucl. Acids Res.* **14:** 4683-4691.

Vos, H. L., Devries, Y., and Hilkens, J. (1991). The mouse episialin (Muc-1) gene and its promoter-rapid evolution of the repetitive domain in the protein. *Biochem. Biophys. Res. Commun.* **181:** 121-130.

Vos, H. L., Devarayalu, S., de Vries, Y., and Bornstein, P. (1992). Thrombospondin 3 (Thbs3), a new member of the thrombospondin gene family. *J. Biol. Chem.* **267:** 12192-12196.

Wahls, W.P., Wallace, L.J., and Moore, P.D. (1990). Hypervariable minisatellite DNA is a hotspot for homologous recombination in human cells. *Cell* **60:** 95-103.

Wahls, W. P., Swenson, G., and Moore, P.D. (1991). Two hypervariable minisatellite DNA binding proteins. *Nucl. Acids Res.* **19:** 3269-3274.

Wang, Y., Abernethy, J.L., Eckhardt, A.E., and Hill, R.L. (1992). Purification and characterization of a UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase specific for glycosylation of threonine residues. *J. Biol. Chem.* **267:** 12709-12716.

Watson, C. J., Gordon, K.E., Robertson, M., and Clark, A.J. (1991). Interaction of DNA-binding proteins with a milk protein gene promoter in vitro: identification of a mammary gland-specific factor. *Nuc. Ac. Res.* **19:** 6603-6610.

Watson, M.L., Kingsmore, S.F., Johnston, G.I., Siegelman, M.H., Le Beau, M. M., Lemons, R.S., Bora, N.S., Howard, T.A., Weissman, I.L., McEver, R.P., and Seldin, M.F. (1990). Genomic organization of the selectin family of leukocyte adhesion molecules on human and mouse chromosome 1. *J. Exp. Med.* **172:** 263-272.

Welsch, U., Schumacher, U., Buchheim, W., Schinko, I., Jennes, R., and Patton, S. (1990). Histochemical and biochemical observations on milk-fat-globule membranes from several mammalian species. *Acta Histochem.* **40:** S59-64.

Wesseling, J., Ligtenberg, M., Vos, H., van der Valk, S., Buijs, F., and Hilkens, J. (1992). The mucin-like glycoprotein episialin modulates cell-cell and cell-matrix adhesion. *Proceedings of the 2nd International Workshop on Carcinoma-Associated Mucins.* Cambridge, UK.

Wilkinson, M., Doskow, J., and Lindsey, S. (1991). RNA blots: staining procedures and optimization of conditions. *Nucl. Acids Res.* **19:** 679.

Williams, R. L., Hilton, D.J., Pease, S., Wilson, T.A., Stewart, C.L., Gearing, D.P., Wagner, E.F., Metcalf, D., Nicola, N.A., and Gough, N.M.

(1988). Myeloid leukaemia inhibitory factor (LIF) maintains the developmental potential of embryonic stem cells. *Nature* **336:** 684-687.

Wolfe, K.H., Sharp, P.M., and Li, W-H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283-285.

Wolff, R. K., Plaetke, R., Jeffreys, A.J., and White, R. (1989). Unequal crossingover between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci. *Genomics* **5:** 382-384.

Wong, Z., Wilson, V., Patel, I., Povey, S., and Jeffreys, A.J. (1987). Characterization of a panel of highly variable minisatellites cloned from human DNA. *Ann. Hum. Genet.* **51:** 269-288.

Wreschner, D. H., Hareuveni, M., Tsarfaty, I., Smorodinsky, N., Horev, J., Zaretsky, J., Kotkes, P., Weiss, M., Lathe, R., and Keydar, I. (1990). Human epithelial tumor antigen cDNA sequences: Differential splicing may generate multiple protein forms. *Eur. J. Biochem.* **189:** 463-473.

Wu, C-I., and Li, W-H. (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. U.S.A.* **82:** 1741-1745.

Xu, G., Huan, L-J., Khatri, I.A., Wang, D., Bennick, A., Fahim, R.E.F., Forstner, G.G., and Forstner, J.F. (1992). cDNA for the carboxyl-terminal region of a rat intestinal mucin-like peptide. *J. Biol. Chem.* **267:** 5401-5407.

Yagi, T., Ikawa, Y., Yoshida, K., Shigetani, Y., Takeda, N., Mabuchi, I., Yamamoto, T., and Aizawa, S. (1990). Homologous recombination at c-fyn locus of mouse embryonic stem cells with the use of Diptheria toxin A-fragment gene in negative selection. *Proc. Natl. Acad. Sci. U.S.A.* **87:** 9918-9922.

Yee, H. A., Wong, A.K.C., van de Sande, J.H., and Rattner, J.B. (1991). Identification of novel single-stranded d(TC)n binding proteins in several mammalian species. *Nucl. Acids Res.* **19:** 949-953.

Zaretsky, J. Z., Weiss, M., Tsarfaty, I., Hareuveni, M., Wreschner, D.H., and Keydar, I. (1990). Expression of genes coding for pS2, c-erbB2, estrogen receptor and the H23 breast tumor-associated antigen. A comparative analysis in breast cancer. *FEBS Lett.* **265:** 46-50.

Zhou, C., Yang, Y., and Jong, A.Y. (1990). Mini-prep in ten minutes. *BioTechniques* **8:** 172-173.

Zimmer, A., and Gruss, P. (1989). Production of chimaeric mice containing embryonic stem (ES) cells carrying a homoeobox Hox 1.1 allele mutated by homologous recombination. *Nature* **338:** 150-152.

Zotter, S., Hageman, P.C., Lossnitzer, A., Mooi, W.J., and Hilgers, J. (1988). Tissue and tumor distribution of human polymorphic epithelial mucin. *Cancer Rev.* **11-12:** 55-100.