Transgenes Of The Human Hypervariable Minisatellite MS32

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Maxine Johanna Allen BSc. (London)
Department of Genetics
University of Leicester

November 1994
Transgenes of the hypervariable minisatellite MS32
Maxine J. Allen, Department of Genetics, University of Leicester.

Abstract.

A high rate of mutation at some minisatellite loci can result in exceptional allelic variability. Often, in addition to length polymorphism, sequence variation exists between the minisatellite repeat units in an allele. The interspersion patterns of minisatellite variant repeat units can be determined using MVR-PCR and in this work the allelic hypervariability observed in earlier studies of the MS32 locus in Caucasians was shown to extend to Asian alleles. Previous analysis of changes in mutant MS32 MVR-PCR patterns, revealed a polarity of mutation at one end of the repeat array, implicating the possible influence of cis-acting factors on mutation at the MS32 locus. Transgenic mice were created carrying an MS32 transgene to assess for effects of flanking DNA and transgene structure on MS32 mutation processes. Eight separate insertion events were characterised, five single-copy and three multi-copy, and a 'preference' for integration into mouse gamma satellite DNA was observed. Southern blot pedigree analysis and SP-PCR of sperm DNA were performed to assay for mutations at the transgenic loci. Mutation rates at the single-copy loci were 10 to 100-fold lower than at the endogenous human locus. MVR-PCR analysis of SP-PCR recovered transgene 110D mutation events indicated regions of repeat unit duplication, deletion and switching, and a gene conversion event was recovered at the transgene 110C locus. In contrast, mutation rates at the multi-copy loci were high (5 - 9.8%) and unusual types of minisatellite mutation including early embryonic events, the frequent loss of a single restriction site and the common deletion of the entire locus were observed. Environmental and genotypic effects on minisatellite mutation were also studied using SP-PCR on DNA from transgene 110D positive mice irradiated with $^{60}$Co, and assaying for minisatellite instability in human cell line DNA with a aberrant mismatch repair phenotype.
For my parents.
## Contents

**Abbreviations**

**Acknowledgements**

**Publications**

### Chapter 1 Introduction

| Variation in genome sequence copy number | 1 |
| Mammalian dispersed repeats, | |
| SINEs | 2 |
| LINEs | 3 |
| Highly repeated satellite DNA | 3 |
| Tandemly arranged rRNA genes | 4 |
| Gene clusters and telomere repeats | 5 |
| **The continuum of tandemly repeated sequences,** | |
| STRs | 5 |
| Midisatellites | 8 |
| Minisatellites | 8 |
| Figure 1.1 | |
| Mouse minisatellites | 12 |
| Tandem repeat sequence turnover processes | 13 |
| This work | 13 |

### Chapter 2 Materials and methods

| Materials, Chemical reagents | 1 |
| Table 2.1 | |
| Methods, Mice | 2 |
| DNA preparation | 2 |
| DNA quantitation | 2 |
| DNA manipulation | 3 |
| Recovery of DNA | 3 |
| Southern blotting | 4 |
| Sequencing | 4 |
| Cloning | 4 |
| Polymerase chain reactions, general, MVR-PCR | 5 |
| Vectorette PCR | 5 |
| Figure 2.1 | |
| Small pool PCR | 7 |
| Computing | 7 |

### Chapter 3 MVR-PCR at MS32 for a series of families of Asian origin

| Summary | 1 |
| Introduction, The MS32 locus | 1 |
| Internal sequence variation | 2 |
| MVR-PCR | 2 |
| Figure 3.1 | |
| Diploid coding and null repeat units | 3 |
| Individual allele codes | 4 |
Chapter 6 Mutation analysis of the MS32 transgene loci

Summary.......................................................................................................................... 1

Introduction,
Minisatellite mutation.................................................................................................... 2
Flanking marker analysis of mutants........................................................................ 2
MVR-PCR analysis of mutants.................................................................................. 3
Mutations identified using SP-PCR.......................................................................... 4
Figure 6.1
This work..................................................................................................................... 5

Results,
Mutation analysis of the single-copy MS32 transgene loci:
Rates of mutation and transmission.................................................................. 6
Table 6.1
Figure 6.2
SP-PCR at transgene locus 110D............................................................................ 7
Figure 6.3
SP-PCR at transgene locus 110C............................................................................ 8
Figure 6.4
SP-PCR at the transgene locus 102A...................................................................... 9
Figure 6.5
Figure 6.6
Homozygosity.......................................................................................................... 10

Mutation analysis of the multi-copy loci:
Rates of mutation and transmission.................................................................. 11
Figure 6.7 - 6.9
MVR-PCR of pedigree mutations........................................................................ 12
Figure 6.10 - 6.12
SP-PCR at loci 109 and 110A.............................................................................. 13
Figure 6.13

Discussion,
Single-copy loci:
Mutation frequencies.......................................................................................... 14
Recessive lethal phenotype of 110D......................................................................... 16
MVR-PCR of SP-PCR mutations............................................................................ 16

Multi-copy loci:
Mutation frequencies.......................................................................................... 17
Early embryonic mutation...................................................................................... 17
Specific loss of a Hind III site................................................................................ 18
Full deletion of the 109 locus................................................................................ 19
MVR-PCR analysis of mutants............................................................................. 20

Conclusions............................................................................................................ 21

Chapter 7 The effects of gamma radiation on mutation processes at the MS32
transgene locus 110D

Summary..................................................................................................................... 1

Introduction............................................................................................................ 1
Table 7.1

Results and discussion......................................................................................... 5

figure 7.1

Conclusions............................................................................................................ 7
Chapter 8  A study of tandem repeat sequence instability in HNPCC DNA

Summary ........................................................................................................................... 1
Introduction,
General......................................................................................................................... 1
How might a malfunction of mismatch repair cause cancer?........................................ 4
This work......................................................................................................................... 5
Results,
PCR analysis of genomic DNA for instability............................................................ 6
SP-PCR of HNPCC samples.......................................................................................... 6
Table 8.1
Figure 8.1
STR analysis of the HNPCC samples........................................................................... 7
Figure 8.2
Discussion,
STR loci......................................................................................................................... 7
Minisatellite loci............................................................................................................. 9
Conclusions.................................................................................................................... 9

Chapter 9  Discussion

Human MS32 locus hypervariability............................................................................. 1
Forensic applications (incl. MVR-PCR).......................................................................... 1
Proposed mechanisms of minisatellite mutation.......................................................... 3
Figure 9.1
MS32 transgenic mice.................................................................................................... 7
Figure 9.2
Evidence for a cis-acting initiator................................................................................ 9
Comparison of STR and minisatellite mutation............................................................. 11
Putative STR mutation mechanisms............................................................................ 14
Minisatellites and disease............................................................................................ 15
Future work.................................................................................................................... 15

Chapter 10  References
Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP</td>
<td>adenosine 5' triphosphate</td>
</tr>
<tr>
<td>bp, kbp, Mbp</td>
<td>base pair, kilobase pair, megabase pair</td>
</tr>
<tr>
<td>BSA</td>
<td>bovine serum albumin</td>
</tr>
<tr>
<td>CEPH</td>
<td>Centre d'Etude du Polymorphisme Humain</td>
</tr>
<tr>
<td>Ci</td>
<td>Curie</td>
</tr>
<tr>
<td>cM</td>
<td>centimorgan</td>
</tr>
<tr>
<td>CRC</td>
<td>colorectal cancer</td>
</tr>
<tr>
<td>DM</td>
<td>Dystrophia Myotonica, Myotonic Dystrophy</td>
</tr>
<tr>
<td>DSB</td>
<td>double strand break</td>
</tr>
<tr>
<td>DSBR</td>
<td>double strand break repair</td>
</tr>
<tr>
<td>dATP</td>
<td>2'-deoxyadenosine 5'-triphosphate</td>
</tr>
<tr>
<td>dCTP</td>
<td>2'-deoxycytosine 5'-triphosphate</td>
</tr>
<tr>
<td>dGTP</td>
<td>2'-deoxyguanosine 5'-triphosphate</td>
</tr>
<tr>
<td>dTTP</td>
<td>2'-deoxythymidine 5'-triphosphate</td>
</tr>
<tr>
<td>dH2O</td>
<td>distilled water</td>
</tr>
<tr>
<td>DMSO</td>
<td>di-methyl-sulphoxide</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DTT</td>
<td>ditiothreitol</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediaminetetra-acetic acid</td>
</tr>
<tr>
<td>FAP</td>
<td>familial adenomatous polyposis</td>
</tr>
<tr>
<td>GEM</td>
<td>gap expansion model</td>
</tr>
<tr>
<td>g, mg, μg, ng</td>
<td>grams, milligrams, micrograms, nanograms</td>
</tr>
<tr>
<td>gw</td>
<td>genomes worth</td>
</tr>
<tr>
<td>hr</td>
<td>hour</td>
</tr>
<tr>
<td>IAP</td>
<td>intercisternal A type particle</td>
</tr>
<tr>
<td>IPTG</td>
<td>isopropyl-β-D-galactopiranoside</td>
</tr>
<tr>
<td>ml, μl</td>
<td>millilitre, microlitre</td>
</tr>
<tr>
<td>LINE</td>
<td>long interspersed repeated element</td>
</tr>
<tr>
<td>LTR</td>
<td>long terminal repeat</td>
</tr>
<tr>
<td>μ</td>
<td>mutation rate</td>
</tr>
<tr>
<td>M, mM, μM, pM</td>
<td>molar, millimolar, micromolar, picomolar</td>
</tr>
<tr>
<td>mtDNA</td>
<td>mitochondrial DNA</td>
</tr>
<tr>
<td>MVR</td>
<td>minisatellite variant repeat</td>
</tr>
<tr>
<td>MVR-PCR</td>
<td>MVR mapping by PCR</td>
</tr>
<tr>
<td>OD</td>
<td>optical densitity</td>
</tr>
<tr>
<td>PBS</td>
<td>phosphate buffered saline</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
</tbody>
</table>

Abbreviations
PEG: polyethylene glycol
PFGE: pulsed field gel electrophoresis
PIC: polymorphism information content
RER+: replication error repair positive DNA
RFLP: restriction fragment length polymorphism
RI: recombinant inbred
RNA: ribonucleic acid
rpm: revolutions per minute
SBMA: Spinocerebellar Ataxia type 1
SDS: sodium dodecyl (lauryl) sulphate
SINE: short interspersed repeated element
SLP: single locus probe
SMD: single molecule dilution
SP-PCR: small pool PCR
SSC: saline sodium citrate
SSCP: single stranded conformational polymorphism
STR: simple tandem repeat
TAE: Tris-acetate EDTA
TBE: Tris-borate EDTA
TEMED: N, N', N'-tetramethyl ethylenediamine
Tris: Tris-(hydroxymethyl) methylamine [2-amino-(2-hydroxymethyl)-propan-1,3-diol]
UV: ultra violet
Vec1: vectorette linker number 1
Vec3: vectorette linker number 3
VNTR: variable number tandem repeat
YAC: yeast artificial chromosome

Abbreviations
Above all I would like to thank Alec for his continual excellent supervision and encouragement. It has been a great privilege to work alongside Alec in the laboratory (loss of my millipores notwithstanding!) during such an exciting period in the laboratory’s work, and to witness his great enthusiasm for science.

I would also like to thank all the members of G19, past and present for their support and friendship throughout my time in Leicester, and for so graciously allowing me to thrash them at Laserquest! Thanks to Annette, Andy, Celia, Darren, David, Duncan, Ester, Ila, Jane, Jo, John, Keiji, Mark, Moira, Nicola, Rita, Tara, and Yuri. In particular I would like to thank Ila, for sparing me the trauma of having to be organised, John for providing the “alternative Maniatis” and to Andy, Nicola, Ester, David, John, Duncan and Yuri for lively discussions on the transgenic world. Special thanks must go to Andy Collick, for shedding light on some of the more unusual pathways in science and for his masterful management of an ever expanding mouse colony in Leicester. In addition, many thanks are extended to the whole Genetics department for providing such a friendly environment in which to work. Much of the credit for the skilfully engineered running of the department must go to Terry Lynn, Jenny Whiteman and Margaret Peake.

I should also like to thank all those who made the East Midlands the epicentre of civilisation during my stay. This mighty achievement owes much to the plottings of one Nicola Rose (the Stick) and to Will, Lee, David (Q.), Tara, Bianca, Milkey, Claire and other residents past and present of No.17., to David (N.) and Helen for many a relaxing evening spent gossiping and to Lesley for being a real dancing queen.

Finally, I would like to thank my family, and friends from afar, Mal, Claire and David (K.) for their unflagging support and encouragement throughout my student years.
Some of this work has been published:


Chapter 1

Introduction

Variation in genome sequence copy number

The C-value paradox. Even before the direct analysis of DNA sequences between organisms became possible, early investigators such as Mirsky and Ris (1951) were able to observe extraordinary variation in the chromosomal DNA content of related organisms. Initial experiments compared the amount of DNA in the haploid genomes of various organisms. This became known as the C-value, expressed in bp of DNA. The C-value was shown to increase with the complexity of organisms from prokaryotes ($10^6$bp) through to the higher eukaryotes ($10^5 - 10^{11}$bp), assumed to reflect the greater number of genes required for the development of higher organisms. However the C-value between related higher eukaryotic organisms was found to vary greatly, for example the amphibians have C-values up to 25 times larger than those of mammals and C-values can vary considerably between different amphibian species. Subsequent estimates of the number of genes and the average gene size related to genome size led to the conclusion that there was an additional, variable, component of the genome beyond that which is necessary to encode the functions required for the development and maintenance of an organism.

Classical DNA reassociation experiments performed to measure the speed at which sheared DNA fragments reannealed following denaturation indicated that not all DNA sequences are represented equally in the genome. The proportion of sequences that reannealed most rapidly were predicted to be present with a high copy number in the genome (Britten and Kohne 1968). These high copy number sequences were found to be ubiquitous in the genomes of higher eukaryotes and to comprise a large proportion of these genomes. For example it has been estimated that 20-30% of the human and mouse genomes are composed of highly and moderately repeated sequences (Schmid and Jelinek 1982; Singer 1982). Additional experiments showed that the highly repetitive sequences were often composed of a type of DNA referred to as satellite DNA. These sequences were mainly centromeric. The moderately repetitive sequences were generally interspersed with single-copy sequences (reviewed in Singer 1982 and Hastie 1989). The remainder of the excess DNA was later identified to be unique sequence. Estimates of the number of non-repetitive sequences expressed and represented in RNA were determined directly in terms of the proportion of the DNA able to hybridise with RNA and suggested that actual coding sequences comprised only 10% of the genome (Ohno 1986).

Repetitive DNA sequence families in mammals

Many of the repeated DNA families in mammals have examples common to both the mouse and human genomes. For the purpose of this review I shall discuss dispersed repeat sequences and satellite repeats in general, and the common mouse and human examples of these sequences. In addition this review will cover telomere repeats and gene clusters in higher eukaryotes in general and describe the continuum of tandem repeat sequences from mononucleotides to satellite repeat units.
Mammalian dispersed repeats

The first mammalian dispersed repetitive sequences to be isolated and sequenced were members of the human 'Alu' family (Houck et al., 1979). Since the discovery of the Alu sequence repeats, many other highly repetitive sequences have been identified in primates and rodents and these will be discussed in more detail. These sequences all appeared to fit into two categories in terms of size, those of the order of several hundred base pairs in length called SINEs (short interspersed elements) and those several thousand base pairs in length, called LINEs (long interspersed elements).

Short interspersed repeated elements (SINEs). SINE families often contain more than $10^9$ member sequences although smaller families exist. Although a variety of functions have been suggested for these sequences, their physiological significance remains unknown. All the major SINEs appear to be homologous to well-characterised small cellular RNAs (such as the 7SL RNA) from which they are thought to derive by the reverse transcription of the homologous RNA species (Baltimore 1985). The general term retroposon was introduced to describe these sequences. An additional characteristic of SINEs is an (A)-rich region of variable length in the part of the SINE thought to correspond the 3' terminus of the RNA from which the SINE derived.

The archetypal primate SINE is the ~300bp Alu retroposon repeat (Houck et al., 1979; Schmid and Jelinek 1982). The Alu element is essentially a head to tail dimer of a 135bp sequence interrupted by ~30bp of an A-rich sequence. The human genome contains over 500,000 members of this family, 5-6% of the total genome, interspersed approximately every 4kb through the genome. The structure of Alu elements suggests that they are dispersed in the genome by replication through an RNA intermediate. Sequencing of Alu repeat units found that they were flanked on either side by a direct repeat of a 10-20bp stretch. This feature was reminiscent of the situation found for a number of transposable elements in bacteria, yeast and drosophila. Insertion of such elements is invariably associated with a duplication of a short stretch of the target site, resulting from staggered nicks at the site of insertion. A run of deoxyadenosine residues typically precedes the 3' direct repeat in an equivalent position to the poly (A) tail of an RNA transcript. The retroposition intermediate is thought to be an RNA polymerase III (RPiII) transcript, since Alu elements contain an internal RPiII promoter. The Alu repeat shows sequence homology with the 7SL RNA component of the signal recognition particle (Ullu and Tschudi 1984). The 7SL RNA is approximately 300nt long. When subjected to limited nuclease digestion the 7SL RNA is cleaved into three fragments, 5' 100nt, 3' 45nt and a central 155nt portion. Only the 5' and 3' regions are represented in SINE sequences. It has been proposed (Hastie 1989) that the 5' and 3' fragments could ligate and be reverse transcribed to produce the ancestral SINE progenitor.

In the mouse genome there are two principle SINE families, the B1 SINE (Krayev et al., 1980) and the B2 SINE (Krayev et al., 1981). The most highly repeated dispersed element in the mouse genome is the B1 sequence, of the order of 130-150bp in length and present at a copy number of 130,000-180,000. This mouse SINE bears strong homology to the human Alu sequence, but the two SINEs are structurally dissimilar, the Alu being a dimeric unit and the B1 sequence a monomer. Like Alu the B1 sequence is homologous to regions of the 7SL RNA from which it is thought to have derived. Mouse B2 elements are ~190bp long and present in the mouse genome at a copy number of ~80,000-100,000. B2 elements share an internal RPiII homology with Alu elements, but are structurally distinct. Individual B2 sequences show an 8% divergence from the consensus repeat, compared to a mean 13%
divergence in humans from the Aku consensus sequence. Rogers (1985) first noted that parts of the B2 SINE bear
homology to tRNAs from which they may have evolved. It is likely that all SINE families have small RNA gene
progenitors from which they have evolved into high copy number dispersed repeats.

**Long interspersed repeated elements (LINEs).** Dispersed repeated elements several kb long, LINEs, have been
described in both rodents and primates (reviewed in Singer 1982). The predominant LINEs of mammalian
genomes are related to the L1 family (Singer 1982). When mouse DNA is digested with Eco R I and
electrophoresed through an agarose gel and visualised by ethidium bromide staining, prominent bands are revealed
over a background smear. These bands represent the internal fragments of the major mouse LINE. Initially this
was referred to as the 1.3RI family to reflect the size of the internal fragment produced by the enzyme Eco R I, but
is now referred to as the mouse L1md element (LINE 1 of Mus domesticus, Veliva et al., 1984). Full length
members of this family are 6-7kb, but family members are very heterogeneous and many are 5' truncated (reviewed
Hastie 1989). Their copy number varies inversely with their length from 5000 -100,000. The most common form
of the L1md element simply contains the 3' end of the element. Mouse L1md elements variably comprise 1-3% of
the mouse genome.

All mammals investigated show sequence identity with L1md (Hastie 1989). The most prominent human LINE is
the L1 element which is 6.3kb in size. It is also often variably truncated at the 5' end. It is thought that the L1 copy
number is around 12,000 (Singer 1982). L1 elements are retroposons and are therefore flanked by direct repeats
and have a poly(A) tail at the 3' end. L1 elements may be dispersed in the genome via a RPIII transcript. Loeb et
al (1986) isolated a full length, 6.85kb, L1md element in the mouse genome and found that the central region of
sequence contained two open reading frames, overlapping by 14bp, the longer of which showed homology to viral
reverse transcription domains. Therefore it appears that the success of the L1 retroposons lies in their ability to
encode a self copying enzyme. Continuing integration of LINEs by retroposition can result in variation in
contemporary genomes. For example, independent de novo L1 insertions have been reported to be responsible for
two cases of Haemophilia A in man (Antonarkis et al., 1988).

**Highly repeated satellite DNA.** The ability of a CsCl density gradient to separate distinct components of the
genome was first observed using mouse DNA (Kis 1961). In these experiments the majority of mouse DNA
(>90%) was shown to have a G/C content of 42% but there was an additional band observed, representing DNA
with a lower G/C base content of around 30%. This fraction, which was separate from the major band in the CsCl
gradient, was termed the satellite band. Other (cryptic) satellite sequences were masked in the main genomic DNA
band. Since then satellite DNA has been isolated from a number of other higher eukaryotic species, including
humans and comprises a varying amount of these genomes from 5% (human), to 50% (Kangaroo rat, Dipodomys
ordii, Fry and Salser 1977). *In situ* hybridisation shows that these sequences are generally located in the non-
expressed constitutive heterochromatin regions of chromosomes, although there are examples of non-centromeric,
ectopic blocks of satellite (e.g. Pardae and Gall 1970, Jones and Corneo 1971). Therefore Singer (1982) redefined
satellites as those highly repeated sequences which are centromerically located to include the cryptic satellites
which cannot be separated from main-band DNA by virtue of their density.
Satellite DNA is organised into long tandemly repeated arrays of a unit sequence. The first human satellite sequences were identified as a separate series of bands by density gradient centrifugation (Miklos and John 1979) and at least 5 classes of satellite DNA have been identified since (Tyler-Smith and Brown 1987). Sequence data indicated that the 'classical' human satellites I-IV were composed of large tandem arrays of relatively short (5-25nt) repeat units (Singer 1982; Prosser et al., 1986). In contrast the most abundant human satellite, alpha-satellite, consists of arrays of a 171bp repeat unit. Alpha-satellite has been shown to be the predominant satellite sequence in centromeric heterochromatin, spanning between 1-3Mb across each human centromere (Willard 1991). The length of these centromeric arrays is highly variable (Waye et al., 1987; Tyler-Smith and Brown 1987) and in addition, further polymorphism within arrays was observed (Waye and Willard 1986). Variant restriction enzyme sites in human alpha-satellite repeat monomers have been observed at human centromeres and have been utilised to identify a number of chromosome specific higher order repeats within alpha-satellite. This variation suggests that satellite DNA is a rapidly evolving component of the human genome, consistent with the divergent organisation of the satellite sequences of Mus musculus (Hastie 1989).

The human beta-satellite is less well characterised, but appears to comprise a considerable proportion of the short arms of acrocentric chromosomes (Greig and Willard 1992). The basic beta-satellite repeat monomer is 68bp in length and has been shown to form the long tandemly repeated arrays typical of other satellite sequences (Greig and Willard 1992).

Two major classes of satellite DNA have been identified in the mouse genome. To reflect differences in their abundance, they are referred to as the major and minor mouse satellites. The organisation of the major mouse satellite was elucidated by Southern (1975) using restriction endonucleases which cleave most repeat units (such as Eco R II) and digested the mouse satellite into a ladder of small bands which were multimers of a basic repeating unit estimated to be 220-260bp in length. The satellite monomer was sequenced and shown to be 234bp long (Horz and Altenburger, 1981).and is estimated to comprise 5-10% of the mouse genome (Waring and Britten 1966; Vissel and Choo 1989; Hastie 1989). The mouse major satellite has been shown to exist in blocks ranging in size from 240kb to 2000kb (Vissel and Choo 1989). In situ hybridisation revealed that most blocks are centromerically located, except on the Y chromosome (Pardee and Gall 1970). To distinguish the mouse major satellite from the other mammalian satellites, it was proposed to call this 234bp repeat unit gamma-satellite (Horz and Altenburger 1981; Vissel and Choo 1989) and in this thesis it is referred to as gamma-satellite.

Pietris et al (1983) first identified the minor mouse satellite and showed that it was composed of 120bp monomers. A ladder based on this monomer is visible by Msp I digestion of mouse genomic DNA. This satellite appears to be centromerically located with the exception of the Y-chromosome (Pietris et al., 1983). The same study showed that the minor satellite was sequence related to the major satellite, and that the 120bp repeat unit is composed of two smaller internal repeats very similar to those found in the major satellite, suggesting that both these satellites evolved from a common ancestor.

Tandemly arranged rRNA genes. The short arms of the five human acrocentric chromosomes contain tandemly repeated arrays of the genes for the 18S and 28S ribosomal RNA (rRNA). Each chromosome has ~40 repeats of a 44kb monomer composed of a 13kb transcribed portion and a larger non-transcribed spacer (NTS) region (Arnheim
The transcribed region encodes the 18S and 28S rRNAs and shows high interspecies conservation, whilst the NTS shows high interspecies variability in sequence and length even between closely related species such as the great apes (Arnheim et al., 1980). There are however some species specific variants that have become fixed at dispersed positions throughout the genome suggesting that rRNA evolution involves unequal exchange between repeats on sister chromatids, homologous and non-homologous chromosomes at a high rate over a relatively short evolutionary period (Arnheim et al., 1980; Qu et al., 1991).

Gene clusters. Many other genes are arranged in linear clusters which may contain both identical genes and genes with related functions. These clusters are assumed to have arisen from a single gene source by duplication and repeated rounds of unequal exchange (Maeda and Smithies 1986). Examples of such gene clusters in man include the immunoglobulin genes (Baltimore 1981), the histone genes (Holitz et al., 1981), the human globin genes, α- and β-clusters (Weatherall and Clegg 1982) and the haptoglobin genes (Maeda and Smithies 1986).

Telomere repeats. The molecular ends of mammalian chromosomes terminate in a tandem array of a hexamer sequence (TTAGGG) averaging ~10kb in length (Moyzis et al., 1988). The exact length of the telomere hexamer array is extremely heterogeneous within individuals such that Southern blot analysis using a telomeric probe produces a smear (Hastie and Allshire 1989). The G-rich strand of all telomeres is oriented 5' to 3' towards the chromosome terminus and protrudes 12-16bp beyond the complementary C-rich strand. A specialised ribonucleoprotein reverse transcriptase called 'telomerase' is able to prime from this overhang and allows complete synthesis of the linear chromosomal end, which otherwise would be shortened by progressive rounds of replication. There is some evidence that somatic cells lack telomerase activity and that telomere shortening may lead to gross chromosome instability and contribute to cell senescence (Hastie et al., 1990).

The continuum of tandemly repeated sequences; micro-, mini- and midisatellites. Within eukaryotic genomes there exists a continuum of tandem repeat unit lengths ranging from a single nucleotide to hundreds of bp. In addition these repeat units show a variation in copy number, repeat unit length and total array length. This continuum has been somewhat arbitrarily broken down into classes loosely based upon array length from the satellite arrays of 200-5000kb, through midisatellites with an array size 10-500kb, minisatellite loci which are typically defined as having an array size of >0.5kb to about 30kb to microsatellites (commonly referred to as simple tandem repeats or STRs) which are defined as having a short <0.5kb array (Epplen et al., 1991). Although some of the individual repeat unit lengths from different groups are the same, the different loci can be grouped by the physical methods required to resolve single allele. The larger satellite and midisatellite arrays require pulsed field gel electrophoresis (Mahtani and Willard 1990; Oakey and Tyler-Smith 1990), minisatellites can be resolved by conventional agarose gel electrophoresis (Wong et al., 1990) and STRs are usually resolved by polyacrylamide gel electrophoresis (Weber and May 1989; Litt and Luty 1989).

Simple tandem repeats (STRs)/microsatellites:

1. Mononucleotides. The best characterised mononucleotide repeats are the poly(dA) tracts associated with retroposon tails (Singer 1981, 1982) found in mice and humans. These can show extensive allele length variability which can be assessed by PCR amplification followed by polyacrylamide gel electrophoresis. In addition there
have been a number of mononucleotide repeats identified within human genes such as the APP gene (Munt et al., 1991) and generally within the human and rodent genomes (Beckman and Weber 1992).

2. Dinucleotides. The \((CA)_n\) dinucleotide repeats were the first microsatellite loci to be described (Miesfeld et al., 1981; Hamada et al., 1982; Tautz and Renz 1984) and it is estimated that there are 50-100,000 of these loci in the human genome (Hamada and Kalunaga 1982). They are also widely found in the mouse genome (Dietrich et al., 1994). The small size of these microsatellite alleles makes them particularly amenable to amplification by the polymerase chain reaction (Saiki et al., 1988) and PCR analysis of these loci indicated that they are often polymorphic in repeat unit copy number with heterozygosities of up to 90% (Litt and Luty 1989; Weber and May 1989; Tautz 1989; Weber 1990). The low variability and heterozygosity of microsatellites (when compared to minisatellites) makes them less informative genetic markers, nevertheless their high incidence and dispersion throughout the mouse and human genomes has made them valuable tools in modern genetic research. These properties of dinucleotide loci have been exploited extensively in the creation of linkage maps for both mouse and human genomes (Dietrich et al., 1992; Copeland et al., 1993), in the localisation of numerous disease related genes (reviewed by Collins 1992), in predictive diagnosis and in forensic identification (Edwards et al., 1992).

The use of microsatellite markers for individual identification is limited by their low heterozygosites, although the statistical power of identifications can be cumulatively improved by increasing the number of loci analysed. Multiplex PCRs as developed by Chamberlain et al (1988) increase the number of loci that can be analysed in a single reaction. The small size of microsatellite alleles means that they are ideal for use in the study of ancient or degraded DNA samples and polymorphic microsatellites have been used in the identification from skeletal remains (Hagelberg et al., 1991; Jeffreys et al., 1992).

3. Trinucleotides. Recent efforts have identified the existence of a number of trimeric repeat units and these have been shown to be widely distributed throughout the human genome, where they are estimated to occur every 300-500kb (Edwards et al., 1992), and in addition have been isolated from the mouse genome (Abbott and Chambers 1993). This type of repeat unit has become the subject of intense study, following the recent observation of expansions in triplet repeat arrays associated with a rapidly growing list of human inherited genetic diseases (reviewed in Caskey et al., 1992; Kuhl and Caskey 1993; Mandel 1993, 1994). Expanded triplet repeat units have been identified in the non-translated regions 5' of the FMR1 gene at the fragile X locus (FRAXA) (Kremer et al., 1991) and downstream of the 3' terminus of the myotonic dystrophy (DM) gene (Brook et al., 1992; Fu et al., 1991; Madhadevan et al., 1992). Triplet repeat expansions have also been observed within the coding region of some genes, for example in the polyglutamine tracts of the spino-bulbar muscular atrophy (SBMA) gene (La Spada et al., 1992), in the spinocerebellar ataxia type 1 (SCA1) gene (Orr et al., 1993), in the Huntington's disease gene (Huntington’s Disease Collaborative Research Group., 1993) and within the gene for DRPLA (Dentatorubral pallidoluysian atrophy). All the expanded loci to date have a CAG repeat with the exception of the fragile X associated repeat (FRAXA) and the FRAXE fragile site associated with mild mental retardation which have CGG repeats.

Searches for other trimeric repeats have since been performed using sequence database searches (Riggins et al., 1992), cDNA library screening with synthetic repeat unit probes (Li et al., 1993) and by hybridisation selection...
followed by PCR (Armour et al., 1994). Most of these approaches result in a bias of CG-rich trimers since they exclude the use of synthetic probes based upon the most abundant trinucleotides AAT, AAC, and AAG which are usually associated with retrovirus tails. In addition recovery of trimeric loci in these studies revealed loci of comparatively low informativeness when compared to the highly polymorphic triplet repeat loci associated with human disease (Armour et al., 1994). This suggests that the disease associated triplet repeats may be atypically unstable and not representative of the majority of trimeric repeat sequences in the human genome. Seven mouse genes containing CAG repeats in their 5' translated regions were recently identified (Abbott and Chambers 1994). In addition the mouse homologue of the Human Huntington disease gene has also been isolated and sequenced (Lin et al., 1994). There is strong sequence identity between the mouse and human cDNAs. The human CAG repeat is conserved, but present at half the 'normal' copy number on the human chromosome, a median of seven repeats rather than nineteen. The CGG repeat is present with a similar copy number (seven to twelve triplets) as on the human chromosome and is the polymorphic repeat unit in the mouse (Lin et al., 1994). No doubt more murine STR loci will be discovered with the elevated interest in these sequences.

4. Tetranucleotides. Tetrameric loci have been estimated to occur at roughly the same frequency as trimeric loci, every 300-500kb in the human genome (Edwards et al., 1991). A few tetramer repeat loci were among the first STRs to be recovered by Epplen et al., (1982) who isolated GATAAGACA repeat units from snailc satellite DNA. Subsequently many other tetrameric repeats have been identified by using oligonucleotide probes (Medis et al., 1993) and by using hybridisation selection (Armour et al., 1994). Other tetramer repeats have been fortuitously identified due to their association with genes of interest. For example the (CTTT)₉ sequence closely linked to the retinoblastoma gene (Yandell and Dryja 1989) and the (TGGAA)₉ sequence which is 5' upstream of the myelin basic protein gene (Boylan et al., 1990). Some of the gene associated tetramer repeats have been observed to be highly polymorphic; the (AAAG)₉ of a β-actin related processed pseudogene (Polymeropoulos et al., 1992) and the DXS981 (TATC)₉ locus (Mahtani and Willard 1993). Many of the tetrameric loci recovered have been shown to be more polymorphic than trimeric loci (Armour et al., 1994). This may correlate with the length of the tetrameric repeat arrays recovered in this study, which were longer than the corresponding trimeric repeat units.

5. Pentanucleotides. There are few examples of pentanuclear loci reported in the literature, though this is likely to reflect only a paucity of investigation of these loci rather than a true under-representation of these loci in the genome. Some pentanuclear loci have been fortuitously identified in sequences associated with genes of interest. For example there is a (AAAAAG)₉ repeat at the CD4 locus (Edwards, M.C. et al., 1991) and an (ATAAA)₉ repeat in the 5' flanking DNA of the glutathione-S-transferase PI-gene (Harada et al., 1994). In addition sequence database screens have been used to identify pentanumeric repeat unit sequences within mammalian genomes (Bornske et al., 1994).

Minisatellites. Repeat unit lengths at the human minisatellite loci characterised to date range from a minimum of 9bp at the MSI 1 locus (Wong et al., 1987) to maximum of 90bp at the MS607A locus (Armour, 1990b), and total minisatellite tandem repeat array sizes are typically in the order of 0.5-30kb. While some minisatellite loci appear to be monomorphic with respect to allele length in the populations studied (Armour et al., 1990), others frequently exhibit very high variability. The work in this thesis is based upon a study of variability at minisatellite loci and...
the controls that modulate the creation of hypervariable loci and therefore the history, isolation and characterisation of these loci will be described in greater detail at the end of this section.

**Midisatellites.** Within the hierarchy of tandemly repeated sequences the midisatellites represent those tandem repeat arrays intermediate in size between the minisatellites and the large megabase satellite arrays. They have monomer repeat unit lengths between 9-40bp, and are distinguished loosely from minisatellites by the total length of their arrays which are typically 10-500kb. One midisatellite, the D1Z2 locus, was identified by Nakamura et al (1987b) near the telomere of chromosome 1. It has a 40bp repeat unit and is arranged in arrays between 250-500kb in length. In addition Page et al (1987) identified a midisatellite sequence with a repeat unit length of 61bp in the pseudautosomal region, with arrays of 10-50kb. Both of these midisatellites showed polymorphism in length and internal sequence. Two further midisatellites were identified by Gray (1991), with array lengths of 50-200kb and a relatively short 9bp repeat unit. The relative paucity of these loci identified may reflect the greater difficulties involved in cloning and characterising large tandemly repeated loci.

**Minisatellite loci**

**History.** The first highly polymorphic region to be described in the human genome was a highly polymorphic restriction fragment length polymorphism (RFLP) which had more alleles than the standard dimorphic loci previously characterised (Wyman and White 1980). This polymorphism was subsequently shown to result from length changes at a tandemly repeated sequence (Wyman et al, 1986). Over the next few years a series of loci were fortuitously discovered from the sequence of essentially random clones usually containing a gene of interest. For example multiallelic loci were identified in the 3' flanking DNA of the human α-globin genes (Higgs et al, 1981), the Ha-ras gene (Capon et al, 1983) and in the human insulin gene (Bell et al, 1982) and the polymorphism was shown to arise from a variation in tandem repeat copy number for the polymorphic regions of the Ha-ras and insulin genes.

**DNA fingerprints.** The major breakthrough in the study of these sequences came with the discovery that a number of novel multiple tandem repetitive loci could be detected by Southern blot analysis by using a repeated DNA sequence probe hybridised at low stringency (Jeffreys et al., 1985a). The repeated sequence probes used in these studies was originally derived from a tandem repeated sequence in the first intron of the human myoglobin gene. Two of these probes, called 33.6 and 33.15, each detected between 10-20 apparently unlinked polymorphic loci of a similar sequence which were termed minisatellite loci (Jeffreys et al., 1986). The number of loci that cross hybridised to both these probes was estimated to be around 1000, though most were small and probably monomorphic. However the largest (4-20kb) fragments in composite profiles generated by simultaneous detection of these loci were shown to be so polymorphic as to produce an individual specific DNA profile, which was called a 'DNA fingerprint'.

Following this initial work other tandem repeat probes were identified which also detected multiple polymorphic loci when hybridised at a low stringency. For example the α-globin 3' hypervariable region (Fowler et al., 1988), a tandem repeat from the M13 pluge genome (Vasart et al., 1987) and also various synthetic oligonucleotides comprising simple sequence motifs (Ali et al., 1986; Vergnaud et al., 1991; Epplen et al 1991). DNA sequences
that detected multiple variable loci under low stringency hybridisation conditions came to be collectively known as multi-locus probes or MLPs.

**Applications of multi-locus DNA 'fingerprints'.** The DNA fingerprints detected by the sets of probes 33.6 and 33.15 were individual specific with the exception of identical twin DNA and individual bands showed Mendelian inheritance, with the exception of novel mutant bands. The potential application of this information to individual identity testing in paternity, immigration and forensic cases was quickly realised (Jeffreys et al., 1985b, 1985c; Gill et al., 1985). In addition the technique was used in monitoring bone marrow transplants (Thein et al., 1986), to follow tumour progression (Thein et al., 1987) and to enable the identification of cell lines (Thacker et al., 1988).

**Other species.** The human multi-locus core probes 33.6 and 33.15 were shown to hybridise to a variety of variable loci in mammals including mice (Jeffreys et al., 1987a), primates (Ditson et al., 1992), dogs and cats (Jeffreys and Morton 1987), cattle (Georges et al., 1991), and pigs (Signer et al., 1994) and in some bird species (Burke and Bruford 1987; Signer and Jeffreys 1993). In these cases they are typically used for the establishment of family relationships in wild populations and to maximise outbreeding in captive animals.

**Isolation.** The cross hybridising properties of the initial minisatellite probes enabled the directed cloning of more minisatellite loci (Jeffreys et al., 1985a; Wong et al., 1986). Initial approaches for the deliberate isolation of minisatellite loci cloned selected loci from an MLP profile (Wong et al., 1986, 1987). However, although this technique identified some superbly polymorphic loci, it was not efficient for detecting large numbers of variable loci and therefore new techniques for their isolation were developed. The most successful of these have used multi-locus probes to screen genomic libraries including: human \(\lambda\) libraries screened with 33.15 and 33.6 (Wong et al., 1987), human cosmid libraries (Nakamura et al., 1987a, 1988) and ordered array chromid libraries (Armour et al., 1990; Signer et al., 1994).

Most of the cloned minisatellite probes detect only their cognate single locus in a high stringency hybridisation reaction. Used in this way they have proven to be very informative genetic markers (Nakamura et al., 1987a) and to have high heterozygosities of up to 99% (Wong et al., 1987). These single-locus probes (SLPs) are also widely used in individual identification in paternity and forensic cases. They have also been used to detect allele loss in tumours (Vogelstein et al., 1989; Armour et al., 1989b) and in the detection of chromosomal abnormalities such as uniparental disomy in Angelman's syndrome (Malcolm et al., 1991). In addition minisatellite loci provide extra polymorphic markers for the construction of genetic maps (Dennis-Keller et al., 1987).

**Genomic distribution and flanking sequences.** The recovery of minisatellites may be subject to some cloning bias, but it appears that an estimated copy number of 1500 minisatellite loci in the human genome (Braman et al., 1985; Jeffreys et al., 1987) are equally dispersed on all the autosomes (Nakamura et al., 1988; Armour et al., 1989, 1990; Vergnaud et al., 1991). In situ hybridisation and linkage mapping revealed that human minisatellite loci were not uniformly distributed throughout the genome, but showed a strong, though not exclusive, clustering in the proterminal regions at or near the ends of genetic linkage maps (Royle et al., 1988; Dennis-Keller et al., 1987). This is one drawback in the utilisation of highly polymorphic minisatellite loci in the construction of human genetic
linkage maps. In addition two examples of pairs of proterminal minisatellites (MS43A and MS43B, and P4g3 and an adjacent minisatellite sequence) showing very close linkage have been observed (Royle et al., 1988) suggesting that the subtelomeric regions of chromosomes may be enriched for repetitive sequences. Sequence analysis of the DNA flanking six hypervariable minisatellite loci (Armour et al., 1989) showed that minisatellite flanking sequences contain a high frequency of dispersed repetitive sequences such as Alu repeats, and L1 elements. In addition, putative retroviral LTR sequences were found flanking the human minisatellite MS32. The alignment of the LTR sequence is disrupted by the MS32 locus, but extends up to the 5' boundary of the minisatellite and resumes on the other side. This suggested that the tandem repeat block of the MS32 locus may have expanded from within a diverged member of the LTR family. From these studies it was demonstrated that the more variable minisatellite loci often cluster together in atypical regions of the genome, which are rich in both tandemly repeated sequences and dispermic repeats.

Coding minisatellites. Most of the minisatellites that have been cloned to date are non-coding (Wong et al., 1987; Armour et al., 1990, Buard and Vergnaud 1994), but there are a small number of minisatellite loci that show allele variability and are expressed, for example the per gene of Drosophila melanogaster which is involved in the control of circadian pupal development patterns and adult activity patterns (Costa et al., 1992). The number of repeat units in per have been shown to increase with geographical latitude, suggesting that this region may be involved in circadian activity control in response to temperature change. In addition one highly variable (80% heterozygosity) human expressed minisatellite is the hypervariable MUC1 locus which encodes the human epithelial mucin (Swallow et al., 1987). Other human coding minisatellites have been reported though these are less variable. They include other mucin proteins (Ginn et al., 1989) and the involucrin gene (Simon et al., 1991).

Properties of minisatellites. Hypervariable minisatellites can have allele length heterozygosities of up to 99% making them the most polymorphic loci yet identified in the human genome (Wong et al., 1987; Vergnaud et al., 1991). The differences in array length between different alleles are due to allelic variation in the in their tandem repeat copy number. For this reason minisatellites are sometimes referred to as hypervariable regions (HVRs) or variable number tandem repeat loci (VNTRs). Variation in repeat unit copy number can be assayed by restriction digestion with a restriction enzyme which cuts outside the tandemly repeated region, followed by Southern blot hybridization using the minisatellite sequence as a probe at high stringency (e.g., Wong et al., 1986). Southern blot pedigree analysis also revealed a high rate of mutation at these loci which forms the basis of the extreme level of allelic diversity at these loci with respect to tandem repeat copy number (Jeffreys et al., 1988a).

Variant repeat units. In addition to allele length variation, sequence analysis revealed variation between individual tandem repeat units at all of the hypervariable human minisatellites characterised to date (Bell et al., 1982; Capon et al., 1983; Owerbach & Aagaard, 1984; Wong et al., 1986, 1987). Variation in the pattern of interspersion across an allele of these variant sequence repeat units has been observed (Owerbach & Aagaard, 1984; Wong et al., 1986, 1987). Assaying this variation is a powerful alternative to distinguishing alleles by differences in length and has proved highly informative in studies of minisatellite allelic diversity and the mechanisms that maintain variability (Jeffreys et al., 1990, 1991a; Armour et al., 1993; Neil and Jeffreys., 1993; Desmarais et al., 1993; Anot et al., 1993; Buard and Vergnaud, 1994). Variant sequence repeat units are referred to as minisatellite variant repeats (MVR).
**Analysis of MVR interspersion at the MS32 locus.** MS32 is a hypervariable human locus that has a 29bp repeat unit with two major variant repeat units (see figure 1.1 A and B). The presence of a G/A polymorphism creates or destroys a *Hae* III site in some variant repeats, whilst both variant repeat types have a *Hin* I site. The pattern of interspersion of the *Hae* III⁺-repeat unit types across an MS32 allele was initially assayed using DNA from a PCR amplification across the entire MS32 repeat unit array with flanking primers, then a partial *Hae* III digestion of this product in conjunction with *Hin* I digestion of a separate aliquot of product (Jeffreys et al., 1990). This technique utilised the fact that MS32 repeat units are almost always cleaved by *Hin* I, but that only those variant repeat units with a ‘G’ at the A/G variant position also contain a site for *Hae* III restriction. Single MS32 alleles small enough for amplification in their entirety by PCR were isolated by amplification from genomic DNA to levels detectable as a discrete band on an agarose gel stained with ethidium bromide (Jeffreys et al., 1990). These alleles were then electroeluted onto dialysis membrane and reamplified with MS32 flanking primers, one of which carries an *Eco* R I extension allowing subsequent *Eco* R I cleavage and end-labelling of PCR of products by fill-in labelling. Aliquots containing an end-labelled MS32 allele were then partially digested with *Hin* I to cleave every repeat unit or by *Hae* III to cleave only those repeat units with the ‘G’ variant. The gel was exposed to autoradiography and the patterns of bands revealed an MVR map of the allele (figure 1.1 C). The *Hin* I cleaved lane will have a band corresponding to each repeat unit in the array, the *Hae* III digested lane will have bands only at the sizes corresponding to those repeats with the ‘G’ polymorphism. Repeat units cleaved by *Hae* III were given the name a-type repeat units and those cut only with *Hin* I were called t-type repeat units. PCR amplified MS32 alleles can be mapped from either end, depending upon which of the flanking primers carries the *Eco* R I site, allowing alleles of 5kb to be mapped completely.

**Putative function of minisatellite loci?** The initial observation of high mutation rates at minisatellite loci, coupled with their repetitive sequence led to speculation that high levels of variability at these loci were the result of unequal exchange of misaligned minisatellite alleles. These events were proposed to result from the predicted involvement of minisatellite loci in promoting chromosome synopsis and/or meiotic recombination. This function was initially suggested by the observation of a common G/C-rich consensus sequence ‘core’, identified in the repeat unit sequence of many of the minisatellite loci that hybridised to the multi-locus probes 33.15 and 33.6. This ‘core’ sequence of 11-16bp, showed homology to the X recombination signal of the bacterium *E. Coli*, leading to the hypothesis that the minisatellite ‘core’ sequence acted as a recombinogenic signal in the human genome (Jeffreys et al., 1985b). The telomeric and sub-telomeric regions of chromosomes regions have been shown to be G/C-rich and to act as sites for the initiation of chromosome synopsis and meiosis (Solari et al., 1980; Laurie et al., 1985). Minisatellites have been shown to cluster in these regions and in addition minisatellite core probes have been shown to hybridise preferentially to chiasmata in human bivalents (Chandley et al., 1988). These data lent weight to the theory of minisatellites being involved in some aspects of the human recombination process and it was suggested that they represented hotspots of recombination in these genomes, evolving as a consequence of their own, possibly sequence directed recombination proficiency.

Alternatively a second scenario is that minisatellites may have evolved as a highly unstable and coincidentally G/C-rich repeated sequences due to as yet unknown mechanism, which may possibly involve local recombination in the sub-telomeric regions. Another common feature of G/C-rich minisatellites is a purine/pyrimidine strand asymmetry. Studies of plasmids containing a number of 12bp G/C-rich repeats with a pronounced strand
Figure 1.1 A. Schematic of the MS32 locus. 5' and 3' flanking DNA (plain line), region of retroviral LTR homology (hatched box), repeat units cut by Hae III (open boxes), repeat units not cut by Hae III (shaded boxes). MS32 flanking primer sites are represented by arrows, and restriction sites are: S, Sau 3AI, F, Hinf I, H, Hae III. Repeat units cut by Hae III are called a-type repeats and those uncut by Hae III are called t-type repeats.

B. The 29bp MS32 consensus repeat unit sequence, showing the constant Hae III site and the polymorphic Hae III site.

C. Schematic profile of the internal repeat unit map of a single MS32 allele. Amplification is performed with primers 32C and 32D. Either one of these primers can be used with a 5', 10nt Eco RI linker extension. After amplification the PCR product is cleaved with Eco RI and end labelled. Two separate aliquots of labelled PCR product are partially digested with either Hinf I or Hae III, electrophoresed in adjacent lanes on agarose then the dried gel is exposed for autoradiography. The Hinf I digested product has a ladder of bands corresponding to each repeat unit in an allele, but the Hae III digested products only show bands corresponding to a-type repeats. By using either 32CE or 32DE, in the initial amplification, internal mapping can be performed from either end of the locus. This figure was adapted from D. Neil, 1994, Ph.D Thesis.
asymmetry from HSV1 have indicated that these repeat sequences can adopt unusual DNA conformations \textit{in vivo} (Wohlrab \textit{et al}, 1987) leading to physical stress on the double helix and potential strand bending resulting in double stranded breaks. Therefore it is possible that some of the elevated mutation rate at these loci is the result of the increase in double stranded DNA breaks at G/C-rich loci.

Support for the influence of the 'core' sequence upon minisatellite function and in terms of generating polymorphism at these loci was reduced following the discovery of highly polymorphic A/T-rich minisatellite sequences (Vergnaud \textit{et al}, 1991) with no homology to the G/C-rich 'core'. Other G/C-rich minisatellites had also been identified with greatly reduced homology to the 'core' sequence (Wong \textit{et al}, 1987; Nakamura \textit{et al}, 1987). Therefore if this 'core' does have a sequence specific effect upon minisatellite mutation processes it is limited to a subset of loci. Additionally if minisatellites were to play such a fundamental role in cell biology then they would be expected to behave in similar ways in different species. PCR and sequence analysis have revealed human minisatellite locus homologues in other species of primates, but they are often small and monomorphic (Gray 1991; Armour \textit{et al}, 1991). This implies that the primate minisatellite sequences do not all participate in the same molecular turnover processes as their human counterparts. This would argue against although not exclude, a conserved role for every minisatellite, but does not rule out the possibility that some minisatellite loci evolved to high repeat unit copy number through their proximity to local recombination hotspots.

**Mouse minisatellites**

The human minisatellite probes 33.15 and 33.6 can cross hybridise to mouse DNA and detect variable loci giving a DNA fingerprint similar to that seen in humans (Jeffreys \textit{et al}, 1987a). In contrast to human minisatellite loci murine minisatellites do not appear to cluster in the sub-telomeric or telomeric regions of chromosomes, which may be associated with an internal initiation of synopsis in the mouse (Jeffreys \textit{et al}, 1987a; Julier \textit{et al}, 1990).

The first minisatellite to be isolated and characterised from the mouse genome is called Ms6-hm. It was detected in mouse DNA by the human probe 33.6 and subsequently cloned, sequenced and mapped to an interstitial location on chromosome 4 (Kelly \textit{et al}, 1989). It has a pentameric repeat unit sequence (GGGCA)$_n$ which has expanded from within a member of the mouse MT (mouse transcript) SINE family (Heinlein \textit{et al}, 1986) and is flanked by another diverged MT repeat sequence and a mouse B2 element (Kelly \textit{et al}, 1991). The Ms6-hm probe hybridises to other loci in the mouse genome including the tetrameric (GGCA)$_n$ locus Hm-2 which was isolated, cloned, sequenced and localised to the distal region of chromosome 9 (Gibbs 1992; Gibbs \textit{et al}, 1993). A third mouse minisatellite locus, Pc-2, with a 7bp repeat unit (GGCAGGA)$_n$ was isolated and characterised by Suzuki \textit{et al} (1993) and shown to map to chromosome 6.

The repeat unit lengths of these mouse minisatellite loci are short when compared to the human minisatellite loci, but the array sizes at the Ms6-hm, Hm-2 and Pc-2 loci range between 3-25kb (Kelly 1990; Gibbs 1992; Suzuki \textit{et al}, 1993) and place them firmly within the minisatellite grouping of loci, resolvable by agarose gel electrophoresis. All three loci showed very high rates of germline mutation; Ms6-hm (2.5% per gamete), Hm-2 (3.6% per gamete) and Pc-2 (3.3% per gamete). In addition two of these loci showed a high level of somatic instability producing mutational mosaicism; Ms6-hm (2.8% of mice), Hm-2 (20.4% of mice) (Kelly \textit{et al}, 1989; Gibbs \textit{et al}, 1993).
Tandem repeat sequence turnover, processes?

Tandemly repeated DNA sequences are generally highly polymorphic components of the human genome in terms of total array length and of variant repeat unit interspersion. It is assumed that the high rate of variation is provided by an elevated mutation rate at these loci when compared to other regions of the genome. By analysing patterns of polymorphic variation and studying individual de novo mutation events it is possible to make some initial inferences as to the processes that may be operating to generate and maintain variation at these loci. The germline mutation rate at the most polymorphic tandem repeat loci, in particular the hypervariable minisatellites, is high enough that de novo germline mutation events can be identified by pedigree analyses, allowing mutation rates to be quantified and individual mutation events to be analysed (Jeffreys et al., 1988a, 1990, 1991; Vergnaud et al., 1991; Mahtani & Willard, 1993). To date the most intensively studied polymorphic loci are the minisatellites, which are the main focus of the studies in this laboratory and comprise the majority of the analyses in this thesis.

A number of mutational mechanisms that may alter the number of repeats in tandemly repeated blocks of DNA have been proposed. These include interallelic processes, like unequal recombination and gene conversion (Smith, 1976; Dover, 1982), and intraallelic mechanisms, such as replication slippage (Tautz et al., 1986; Levinson & Gutman, 1987), unequal sister chromatid exchange (USCE) (Smith, 1976) and deletion by intramolecular recombination. Although many of these processes may be ubiquitous to all classes of tandem repeated sequences it is possible that some mechanisms are more common to particular locus lengths and/or repeat unit sequence type and in addition that cis or trans acting factors may modulate the effects of these processes on tandemly repeated loci.

Putative minisatellite mutation processes. Initial analyses of minisatellite distribution and repeat unit sequence provided circumstantial evidence of a role for minisatellite loci as hotspots for unequal homologous recombination (Jeffreys et al., 1985a, 1988 and previous section on putative functions of minisatellites). However, subsequent analyses of individual minisatellite mutation events failed to show a single example of exchange of flanking markers, as would be expected in simple homologous recombination (Wolff et al 1988, 1989 and chapter 6). Jeffreys et al. (1988a) observed that gains and losses of minisatellite repeat units appeared to occur with an approximately equal frequency, suggesting that intramolecular recombination processes, which would tend to result in a reduction in repeat array length, were not the major mechanism involved in minisatellite mutation. The recent development of a PCR based system for assaying internal variation in minisatellites, MVR-PCR (Jeffreys et al 1991b and chapter 3), will enable the direct assessment of the prevalence of other processes such as replication slippage, USCE and gene conversion events in minisatellite mutation. Early results (Jeffreys et al., 1991b) provide evidence for the involvement of all three aforementioned mechanisms in the turnover of repeat units at minisatellite loci, often resulting in a complex mutational profile and have revealed a polarity of variation at three minisatellite loci suggesting the putative involvement of the minisatellite flanking DNA in the mutation processes.

This work

In the initial part of this study a new PCR based technique to identify the interspersion patterns of variant repeat units at the minisatellite locus MS32, called MVR-PCR, was used to analyse MVR patterns for a group of MS32 alleles of Asian origin. It was proposed to compare the type of MS32 variant repeat interspersion patterns seen with this population, with alleles from other population groups studied in this laboratory. This data would go
towards the compilation of an MS32 allele database. The information in this database would enable the study of
minisatellite evolutionary processes and give an indication of the patterns of variability seen at the MS32 locus.
This in turn led to clues as to the influence of the 5' flanking DNA upon minisatellite mutation with an observed
polarity of mutation at the locus.

Mice provide an accessible system in which to study the genetic behaviour of individual loci, with the advantages
of genetically defined inbred strains and mouse breeding systems. Potentially they provide a system for enabling
the study of minisatellite behaviour in response to genotypic factors such as mouse strain, genomic position of the
locus and flanking sequence type. A model system could also be used to analyse the susceptibility of minisatellite
loci to environmental effects on their mutation processes. The endogenous mouse minisatelites were first
investigated as a potential model system for minisatellite analysis which could then be extrapolated to humans.
However, early work on the first two minisatellites identified in the mouse genome (Ms6-hm and Hm-2) indicated
that these loci frequently mutated postzygotically. This type of mutation had not been observed in the endogenous
human loci indicating that different mutational processes may be operating on these loci and that the mouse
minisatelites would not provide a good direct model for minisatellite analysis that could be related to the situation
at the human minisatellite loci. In addition repeat unit lengths were short (tetramer/pentamer), and as yet no mouse
minisatellite variant repeat units had been identified (Kelly 1990; Gibbs 1992). Therefore it was not as easy to
follow the processes of minisatellite mutation and variability as it was at human loci using the MVR-PCR
technique.

It was therefore decided to create transgenic mice carrying a human minisatellite locus. The best characterised
human minisatellite locus is MS32 (D1S8) and this was the locus chosen for transgenesis. Microinjection was
chosen as the method of introducing the minisatellite DNA into the mouse genome (see chapter 4). With the recent
flurry of interest in the instability of tandemly repeated sequences in the human genome, instigated by the
discovery of a number of human disorders caused by expansions of tandem repeats, this experiment would provide
the first evidence of the behaviour of a foreign tandemly repeated sequence when inserted into the mouse genome.
We were interested in whether a tandemly repeated DNA would be able to insert into the mouse genome, without
deleterious effects and if so what type of structures it would form. There was also the possibility that a tandemly
repeated array such as MS32 which has a G/C-rich repeat sequence 'core' might show some targeting to other G/C-
rich sequences in the mouse genome, known to exist from mouse DNA fingerprints. If transgenic lines could be
successfully established then the aim was to analyse what mutation processes were occurring at these loci, if they
maintained the mutation properties seen at the human MS32 locus, such as a polarity of variation, and what effect
the mouse flanking DNA sequences had upon mutation at the transgenic loci. This study forms the basis of the
research in this thesis.

Finally, I looked briefly at some human colon cancer cell line DNAs with a characterised mutation in their DNA
mismatch repair processes. I wanted to ascertain if the instability seen at STR loci in this DNA was shared by
human minisatellite loci, to give some indication of the relative importance of mismatch repair processes in
controlling instability at human minisatellite loci.
Chapter 2

Materials and Methods

The data in this thesis was collected using a number of well established molecular biology techniques that are adequately described elsewhere. This chapter provides an overview of the techniques used, with references for more detailed descriptions.

Materials

Chemical reagents. All chemical reagents unless otherwise specified were supplied by Fisher, Loughborough. Antibiotics, BSA, IPTG, PEG, Ficoll 400, TEMED, dithiothreitol (DTT), Spermidine trichloride, and salmon sperm DNA were supplied by Sigma Biochemical company, Poole. Deoxyribonucleotides and dideoxyribonucleotides were supplied by Pharmacia, Milton Keynes. Agarose (Seakem™ HGT, and NuSieve™) was supplied by ICN Biochemicals Ltd., High Wycombe. Ammonium persulphate and 40% acrylamide/bis-acrylamide (19:1) solution, were supplied by Bio-RAD, Watford. Radiochemicals were supplied from Amersham International plc, Little Chalfont.

Oligonucleotides. Oligonucleotides for PCR amplification were produced in the Department of Biochemistry, University of Leicester using an Applied Biosystems 380B DNA synthesizer and reagents from Cruachem. The sequences and annealing temperatures of all the primers used are detailed in table 2.1.

Enzymes. Restriction enzymes were supplied by Gibco-BRL, New England Biolabs and the Boehringer Corporation. T4 ligase and REact™ buffers were obtained from Gibco-BRL, T4 polynucleotide kinase, T7 DNA polymerase and the Klenow fragment of DNA polymerase I were obtained from Pharmacia.

Molecular weight markers. λ DNA digested with Hind III, and ϕX174 RF DNA cut with Hae III were used.

Bacterial strains and media. Escherichia coli strain XL1Blue competent cells from Stratagene, were used in cloning experiments. Genotype: endA1, hsdR17 (rK-,mK+), supE44, thi-1, recA1, gyrA96, relA1, (lac-), [F', proAB, lacIqZAM15, Tn10, (tetO)].

All media were obtained from Oxoid, Basingstoke, except yeast extract and tryptone which were supplied by Difco, east Molesley. When required, media were seeded with Ampicillin, Tetracycline, IPTG and XGAL to a final concentration of 50μg/ml.

Cloning vector. The vector pBluescript SK+, from Stratagene, was used for sub-cloning PCR fragments, using the Hind III and Eco R I sites and the Sma I and Hind III sites.
### Table 2.1 primer sequences.

#### 2.1 a, MS32 flanking primer sequences

- *Eco* RI linkers are shown in bold.

<table>
<thead>
<tr>
<th>Primer</th>
<th>5’ - 3’ Sequence</th>
<th>Anneal Temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>32-A</td>
<td>TCACCGGTAATCCACCAGACCTT</td>
<td>68°C</td>
</tr>
<tr>
<td>32-B</td>
<td>TAACATAGCTCTCCATTTCCGAGTT</td>
<td>68°C</td>
</tr>
<tr>
<td>32-C</td>
<td>CTTCCGTTCTCTCCAGCGCTAG</td>
<td>68°C</td>
</tr>
<tr>
<td>32-CE</td>
<td>TCACGTTCTGTTCTCCTGCTTCTCCACGGCTA</td>
<td>68°C</td>
</tr>
<tr>
<td>32-D</td>
<td>CGACTCAGAGAGGATTCCTCCGTTCTCCTAG</td>
<td>68°C</td>
</tr>
<tr>
<td>32-DE</td>
<td>TCACCGGTAATCCGCTAGCTCAATGAGACATGGAGCC</td>
<td>68°C-70°C</td>
</tr>
<tr>
<td>32-O</td>
<td>GAGTAGTTTGGTGGGAAGGGTG</td>
<td>68°C</td>
</tr>
<tr>
<td>32-OR</td>
<td>AGACCACACAAACACTACCTC</td>
<td>68°C-70°C</td>
</tr>
<tr>
<td>32-E</td>
<td>CTTCCGTTCTCTCCACGGCTAG</td>
<td>68°C</td>
</tr>
<tr>
<td>32-ER</td>
<td>CTAGGGCTGAGGAGAACGAGGAAG</td>
<td>68°C</td>
</tr>
<tr>
<td>32-H2A</td>
<td>TGAGGCAATGAGGACAGGACGAAG</td>
<td>67°C</td>
</tr>
<tr>
<td>32-H2AR</td>
<td>GTGAGGCAATGAGGACAGGACGAAG</td>
<td>67°C</td>
</tr>
<tr>
<td>32-H1B</td>
<td>TTTGGCTGAGGACTCAACTG</td>
<td>60°C</td>
</tr>
<tr>
<td>32-01HR</td>
<td>CCTGTTCTTACACCCGCTTG</td>
<td>60°C</td>
</tr>
<tr>
<td>32-O2RH</td>
<td>CCTGTTCTTACACCCGCTTG</td>
<td>60°C-70°C</td>
</tr>
<tr>
<td>32-NR</td>
<td>AGTAGCAGGAGGGTGTGGC</td>
<td>60°C</td>
</tr>
</tbody>
</table>

#### 2.1 b, MS32 MVR-PCR primers

- 3’ polymorphic positions are shown in bold.

<table>
<thead>
<tr>
<th>Primer</th>
<th>5’ - 3’ Sequence</th>
<th>Anneal Temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>32-TAG</td>
<td>TCAAGAGTCTTCATGTCGCC</td>
<td>68°C-70°C</td>
</tr>
<tr>
<td>32-TAG-A</td>
<td>TCAAGAGTCTTCATGTCGCC</td>
<td>68°C</td>
</tr>
<tr>
<td>32-TAG-T</td>
<td>TCAAGAGTCTTCATGTCGCC</td>
<td>68°C</td>
</tr>
<tr>
<td>32-TAG-3A</td>
<td>TCAAGAGTCTTCATGTCGCC</td>
<td>68°C</td>
</tr>
<tr>
<td>32-TAG-3T</td>
<td>TCAAGAGTCTTCATGTCGCC</td>
<td>68°C</td>
</tr>
<tr>
<td>32-TAG-CA</td>
<td>TCAAGAGTCTTCATGTCGCC</td>
<td>70°C</td>
</tr>
<tr>
<td>32-TAG-GA</td>
<td>TCAAGAGTCTTCATGTCGCC</td>
<td>70°C</td>
</tr>
<tr>
<td>32-TAG-TC</td>
<td>TCAAGAGTCTTCATGTCGCC</td>
<td>70°C</td>
</tr>
<tr>
<td>32-TAG-AG</td>
<td>TCAAGAGTCTTCATGTCGCC</td>
<td>70°C</td>
</tr>
<tr>
<td>32-TAG-N</td>
<td>TCAAGAGTCTTCATGTCGCC</td>
<td>68°C-70°C</td>
</tr>
</tbody>
</table>

#### 2.1 c, Transgene specific primers

<table>
<thead>
<tr>
<th>Primer</th>
<th>5’ - 3’ Sequence</th>
<th>Anneal Temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>CCACACAGGGCCTGGTTGCTCC</td>
<td>68-70°C</td>
</tr>
<tr>
<td>E2</td>
<td>GGGGGCGCTTCCTCCACGTCCTT</td>
<td>68-70°C</td>
</tr>
<tr>
<td>IAP-I</td>
<td>GGGGGGTTCCTTCCACGTCCTT</td>
<td>62°C</td>
</tr>
<tr>
<td>IAP-II</td>
<td>GGGGGGTTCCTTCCACGTCCTT</td>
<td>62°C</td>
</tr>
<tr>
<td>105B-1</td>
<td>GGGGGGTTCCTTCCACGTCCTT</td>
<td>64°C</td>
</tr>
<tr>
<td>105B-2</td>
<td>GGGGGGTTCCTTCCACGTCCTT</td>
<td>64°C</td>
</tr>
<tr>
<td>11A-1</td>
<td>GGGGGGTTCCTTCCACGTCCTT</td>
<td>68°C</td>
</tr>
<tr>
<td>11A-5</td>
<td>GGGGGGTTCCTTCCACGTCCTT</td>
<td>70°C</td>
</tr>
<tr>
<td>11A-3</td>
<td>GGGGGGTTCCTTCCACGTCCTT</td>
<td>70°C</td>
</tr>
</tbody>
</table>

---

Chapter 2, table 2.1
2.1 d, Other minisatellite primers

<table>
<thead>
<tr>
<th>Primer</th>
<th>5' - 3' Sequence</th>
<th>Anneal Temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS1-A</td>
<td>GCTTTTCTGTGATGAGCCTTTAGT</td>
<td>67°C</td>
</tr>
<tr>
<td>MS1-B</td>
<td>AAGAAGCTATAGCACACCATTAGG</td>
<td>67°C</td>
</tr>
<tr>
<td>MS31-A</td>
<td>CCCTTTGCAGCTGACGGTGAGCAGG</td>
<td>67°C</td>
</tr>
<tr>
<td>MS31-B</td>
<td>CCAGACGGCCACCCCGGGGGGCGGAGG</td>
<td>67°C</td>
</tr>
<tr>
<td>MS205-A</td>
<td>CTGTGCAGGTCAGGCACGTCGGCAG</td>
<td>67°C</td>
</tr>
<tr>
<td>MS205-B</td>
<td>AGCCGAGCCCAGACCTCCCCACATTG</td>
<td>67°C</td>
</tr>
<tr>
<td>MS621-A</td>
<td>CCGTGGCTTTTAGGCGCAGAGCAG</td>
<td>67°C</td>
</tr>
<tr>
<td>MS621-C</td>
<td>GAGCTGCGACAGACGTGCAAGAGC</td>
<td>67°C</td>
</tr>
<tr>
<td>pA3-C</td>
<td>CTGAAGCTCTCACTGCTGCTGTCGAGAG</td>
<td>67°C</td>
</tr>
<tr>
<td>pA3-D</td>
<td>GCTGGAGAATAATTTTGTGTCGCACTC</td>
<td>67°C</td>
</tr>
</tbody>
</table>

2.1 e, Short tandem repeat primers

<table>
<thead>
<tr>
<th>Primer</th>
<th>5' - 3' Sequence</th>
<th>Anneal Temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wg0e7-A</td>
<td>GCACACTGGAAGTGCGCGAAGTAAC</td>
<td>65°C</td>
</tr>
<tr>
<td>Wg0e7-B</td>
<td>AGGCCATTGGAAGTTGCTGCTAAAATATG</td>
<td>65°C</td>
</tr>
<tr>
<td>Wg1e4-A</td>
<td>GATGGGAGACTCTGCACTCAAGAAG</td>
<td>60°C</td>
</tr>
<tr>
<td>Wg1e4-B</td>
<td>AGCTCTATACGATTGTGTTTGG</td>
<td>60°C</td>
</tr>
<tr>
<td>Mk2c8-A</td>
<td>TGAAGGCAAGAAACACATATATATTG</td>
<td>62°C</td>
</tr>
<tr>
<td>Mk2c8-B</td>
<td>GGCAGAAAGACAGACGACACTCT</td>
<td>62°C</td>
</tr>
</tbody>
</table>

2.1 f, pBluescript and M13 primers

<table>
<thead>
<tr>
<th>Primer</th>
<th>5' - 3' Sequence</th>
<th>Anneal Temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>T7</td>
<td>AATACGACTCTATAG</td>
<td>Cool from</td>
</tr>
<tr>
<td>KS+</td>
<td>CGAGTGCGACGTTATAG</td>
<td>65°C to</td>
</tr>
<tr>
<td>M13 (17mer)</td>
<td>GTAAAACGACGGCAGTT</td>
<td>25°C</td>
</tr>
</tbody>
</table>
Human DNAs. Asian family DNA samples from blood, were provided by M. Webb from Cellmark Diagnostics in Abingdon. Colon cancer cell line DNA's (normal and tumour) were provided by B. Vogelstein, Johns Hopkins Medical School, Baltimore, MD, USA. Lymphoblastoid cell line DNAs were provided by H. Cann and J. Dausset of the Centre d'Etude du Polymorphisme Humain (CEPH), Paris, France.

Methods

Mice. Mice were housed in the biomedical services facility at the University of Leicester under a 12hr light/12hr dark cycle. Identification of mice was by toe clipping following the numbering system of Allen et al., 1987. Positive founder mice were mated with C57BL/6J/CBA/ca F1 mice. Subsequent matings were with mice of a C57/6J, CBA/ca, or MF1 mice.

Tissue recovery. Routine DNA analysis was performed on DNA prepared from a length of tail (1-2cm) removed whilst the mice were anaesthetised using halothane. Toe clipping was performed at this time. For some experiments other tissues were used. Liver, heart and brain were removed from a sacrificed animal by dissection. Sperm DNA was recovered by dissection and manipulation of the vas deferens tubes under a binocular microscope to extrude the sperm using forceps, into 500μl of 1x SSC.

DNA preparation.

Somatic tissue preparation. Somatic tissue was chopped and suspended in 2mls of solution A (100mM NaCl, 20mM Tris-HCl pH 8.0, 25mM EDTA pH 8.0) and 2mls of solution B (10mM Tris-HCl pH 8.0, 10mM EDTA pH 8.0, 1% SDS, 200μg/ml Protease K). These reactions were left overnight at 55°C. DNA was extracted from bone fragments and proteins by one phenol extraction, one phenol/chloroform (phenol:chloroform:isoamyl alcohol in a ratio of 25:24:1) extraction and one chloroform extraction followed by isopropanol precipitation. The DNA pellet was cleaned with 80% ethanol then dissolved in 200μl of 1x SSC.

Germline tissue preparation. Sperm DNA was prepared by centrifugation of the sperm and 1x SSC mixture to pellet the sperm. The pellet was initially resuspended in 450μl of 1x SSC and 50μl of 10% SDS to lyse all non-sperm cells, then centrifuged for 5 minutes. The supernatant, including somatic cells was discarded. The pellet was fully resuspended in 500μl of solution A. To disrupt the sperm cell heads, 5μl of β-mercapto ethanol was added and mixed immediately. Then 500μl of solution B was added. These reactions were shaken gently in a 55°C water bath for 4-6 hours. DNA was extracted from the reaction by one phenol extraction, one phenol/chloroform extraction and one chloroform extraction, followed by ethanol precipitation.

DNA Quantitation.

Genomic DNA and PCR products. A 1μl aliquot of the DNA sample was electrophoresed on an agarose gel alongside a known amount of DNA. If the DNA to be assayed was undigested or digested to discrete bands, then a known amount of λHind III DNA was used as a concentration reference. If the DNA to be assayed was fully digested genomic DNA then an aliquot of known concentration of a genomic digest was used as a concentration marker. The amount of DNA present in the sample was estimated by ethidium bromide staining of the gel followed by visual comparison of the two samples.
For experiments where exact DNA concentration was important, in addition to gel analysis, fully digested genomic DNA was assayed using the DNA fluorimeter (Hoeffer instruments, San Francisco). 1μl of a known concentration sample between 20-1000ng, was added to 1ml of filtered 1x TNE buffer (0.05M Tris-HCl, pH 7.4, 0.1M NaCl, 0.005M EDTA, seeded with 1μl Hoechst™ dye, number H33258, per 10mls) to set the reference reading and then 1μl of each sample to be assayed was added to fresh 1 X TNE buffer and the concentration reading measured.

**Oligonucleotides.** The concentrations of ethanol precipitated oligonucleotides were estimated by measuring the optical density at the absorbance, λ260 of three dilutions of the oligonucleotide (1/1000, 1/500, 1/300) in a Cecil Instruments CE 202 Ultraviolet Spectrophotometer. The average of these readings was then used to calculate the concentration based on the approximation that 1 O.D. unit is equivalent to 33μg/ml of oligonucleotide.

**DNA Manipulation.**

**Restriction digestion.** Restriction digestions of total genomic DNA, PCR products and clones was performed in the appropriate REact buffer (Gibco-BRL) using conditions recommended by suppliers. Total genomic mouse tail DNA digestions were typically overnight with 1 unit of enzyme per pg of DNA. PCR product and clone digestion was typically performed for 2 hours with 1 unit per pg of DNA.

**Agarose gel electrophoresis.** Agarose gels in the concentration range 0.5-3.0%, and 15-40 cm in length, were used depending on the size of the DNA fragments to be separated and the degree of resolution required. 0.5-2% gels were cast from Sigma agarose or SeaKem HGT. Gels for resolving products smaller than 200bp used 3% NuSieve agarose. Gels were run using TBE buffer (0.089M Boric acid, 2mM EDTA, pH 8.0) or TAE buffer (40mM Tris-acetate, 20mM Na-acetate, 0.2mM EDTA, pH 8.3). Ethidium bromide was added to all buffers at a concentration of 0.5μg/ml.

**Recovery of DNA by electroelution/preparative gel electrophoresis.** To selectively obtain and concentrate DNA fragments of a given size prior to ligation into vectors, ligation of vectorotic linkers or amplification in a PCR, preparative gel electrophoresis was used. If the DNA to be recovered was detectable using ultra violet light, on an agarose gel after ethidium bromide staining, then a block of agarose containing the UV fluorescing band was cleanly cut from the gel with a scalpel. This DNA was electroeluted onto specially prepared dialysis membrane. The membrane was cut to a size double the width of the agarose block and longer than the depth of the gel from which the block has been removed. This membrane was then boiled in distilled water for 3 minutes. A gel of the same agarose concentration as that from which the block was cut was prepared. A well just wider than the agarose block and long enough to accommodate the dialysis membrane was cut into this second gel. The agarose gel after ethidium bromide staining, then a block of agarose containing the UV fluorescing band was cleanly cut from the gel with a scalpel. This DNA was electroeluted onto specially prepared dialysis membrane. The membrane was cut to a size double the width of the agarose block and longer than the depth of the gel from which the block has been removed. This membrane was then boiled in distilled water for 3 minutes. A gel of the same agarose concentration as that from which the block was cut was prepared. A well just wider than the agarose block and long enough to accommodate the dialysis membrane was cut into this second gel. The agarose block was then inserted into this well with it's largest face perpendicular to the eventual flow of electric current. The single layer of dialysis membrane was inserted into the well beneath and in front of the block of agarose to trap the DNA that migrates out of the block of agarose, when an electric current of 10-20V per cm was passed through the gel. Dependant upon the size of the fraction of DNA to be electroeluted this was between 1-2 minutes for small 100bp fragments and 15-20 minutes for fragments of several kb. The movement of the DNA was monitored by UV illumination from a hand-held UV wand. The DNA was recovered by reducing the voltage to 1-2V per cm of gel then swiftly removing the membrane with the DNA adhering into a microfuge tube with a pair of tweezers. The DNA was then centrifuged off the membrane, by trapping the corner of the membrane in the microfuge tube before
centrifugation, 15,000 rpm, for 1 minute. The DNA was then purified by standard ethanol precipitation (Sambrook 
et al., 1990).

If the concentration of the DNA was too low to be visualised by UV after ethidium bromide staining, for example 
when recovering mutants from a small pool PCR reaction (see PCR methods section) then an ethidium bromide 
free gel was run. Marker lanes were also run and after the allocated electrophoresis time were cut from the gel and 
individually stained with ethidium bromide. Under UV the marker bands were marked by the addition of 5μl of 
loading buffer. These marker lanes were positioned alongside the gel to indicate the regions of the gel where size 
cuts should be made. When dealing with non-visible DNA it was appropriate to take 3-5 adjoining size cuts from 
the region of the gel where the DNA is expected to be and to perform electroelution and ethanol precipitation on all 
these agarose bands. Subsequent PCR analysis revealed which size cuts contain the appropriate DNA sample. The 
speed with which the DNA migrates out of the agarose block can be monitored by reference to some TAE loading 
buffer soaked into the gel alongside the well containing the agarose block and dialysis membrane.

**Southern blotting and hybridisation.** Southern blotting procedures were standard and as described in Jeffreys et 
et al. (1991b). The membrane used for Blotting was Hybond N-FP transfer membrane (Amersham). DNA was 
crosslinked to this membrane using 45 seconds of exposure to UV light. Hybridisation was using a modified 
(Wong et al., 1987) phosphate/SDS-based hybridisation mixture (0.5M sodium phosphate, pH 7.2, 1mM EDTA, 
7% SDS; Church and Gilbert, 1984) at 65°C in a Hybrid rotating bottle hybridisation oven. Pre-hybridisation was 
for one hour at 65°C. 10ng of the appropriate gel purified double stranded probe DNA was labelled by random 
hexamer priming (Feinberg and Vogelstein, 1984) incorporating α-32p-dCTP. Probes were boiled for 3 minutes 
before use. Hybridisation was overnight. Membranes were washed under high stringency (0.1x SSC, 0.01% SDS) 
at 65°C for 1 hour in the hybridisation bottles. Autoradiography was performed as described (Sambrook et al., 
1989) with exposure times in accordance with the signal from the membranes.

**Sequencing reactions and polyacrylamide gel electrophoresis.** Double stranded DNA for sequencing was 
prepared by PCR amplification followed by preparative gel electrophoresis. Sequencing of double stranded PCR 
products was by the method of P.Winship, (1989) which uses DMSO (di-methyl sulphoxide) to maintain single 
stranded DNA after boiling and snap freezing the primer and DNA mixture. Winship sequencing gave 
unsatisfactory results with some DNA fragments and the additional method of Taq cycle sequencing was used for 
these PCR products (Murray 1989). Sequencing products were separated by polyacrylamide gel electrophoresis, as 
described by Sambrook et al, 1989, with 6M urea, 8% acrylamide gels. Sequencing gels were prepared with a 1x 
TBE concentration or a 0.5X TBE to 2.5x TBE concentration gradient from the top to the base of the gel.

**Cloning.** Two PCR products, 110C 5' flanking DNA and 110C3' flanking DNA, proved to be refractory to 
sequencing by either the Winship method or Taq cycle sequencing. These were digested with the appropriate 
enzymes, 110C 5' flanking fragment with Eco RI and Hind III and 110C 3' flanking fragment with Hind III and 
Aci I, and ligated into pBluescript SK+ restricted with Eco RI and Hind III, or cleaved with Hind III and Sma I. 
Ligations were with 50ng of gel purified insert and 250ng of gel purified vector in 20μl of 50mM NaCl, 20mM 
Tris-HCl (pH 7.6), 5mM DTT, 0.1mM ATP, with 2 units of T4 ligase (Gibco-BRL), overnight at 15°C. 1μl of 
each ligation was electroporated into competent XL1Blue cells and plated out on luria agar seeded with ampicillin, 
tetracycline, IPTG and X-GAL and incubated overnight at 37°C.
Single white colonies were picked into a microtitre plate with wells containing 100μl of HMFM (luria broth with, 195μg/ml ampicillin, 12μg/ml of tetracycline, 3.6mM K2HPO4, 1.3mM KH2PO4, 2mM Na citrate, 1mM MgSO4, 4.4% glycerol). Before the microtitre plate was snap frozen to store the clones, two replicate luria agar plus Amp/Tet/IPTG/X-GAL plates were made from the microtitre plate using a steel, pronged 'hedgehog'. The second replicate was onto a plate with a hybond-N circular filter between the media and the cells. These plates were grown overnight at 37°C. The hybond filter with the grown cells was removed from the plate and placed upon a piece of 3MM paper soaked with 2xSSC/5% SDS solution to lyse the cells. The DNA was then fixed to the filter for 3 minutes in the microwave, then the filter was washed in 5xSSC/0.1%SDS, before hybridisation with gel purified 110C PCR products as a probe, then washing, as described for Southern blotting hybridisations.

Small scale DNA preparations from positively hybridising clones were performed (Stratagene pBluescript™ II protocol) and single stranded DNA was made from appropriate clones by the addition of 1μl of M13 phage (VCS-M13 interference resistant helper phage, Stratagene) to a culture of the clone at an OD of 0.7, then overnight growth at 37°C. Single stranded DNA was prepared from this culture (Stratagene pBluescript™ II protocol). 0.5μg of single stranded clone DNA was sequenced using the Sequenase protocol with primers KS+ and T7. Samples were separated by polyacrylamide gel electrophoresis.

Polymerase Chain Reactions.

General PCR (Saltari et al., 1988). PCR reactions were generally performed, unless otherwise stated in the text, with 1μM primer concentrations, and a PCR buffer of 45mM Tris-HCl (pH 8.8), 11mM (NH4)2SO4, 4.5mM MgCl2, 6.7mM 2-mercaptoethanol, 4.4μM EDTA (pH 8.0), 1mM dATP, 1mM dCTP, 1mM dGTP, 1mM dTTP, 113μg/ml BSA (10μg/ml solution from Pharmacia), and Taq polymerase (Amersham or Cetus) at a concentration of 0.015units per μl. Reactions were overlaid with a drop of paraffin oil. Cycling was carried out in a Perkin Elmer Cetus DNA thermal cycler, or model Geneamp 9600, (Perkin Elmer Cetus, Connecticut, USA.).

General amplification conditions were denaturation at 96°C for 1.3 minutes, annealing at X°C for 1 minute (for 'X' see table of primers, 2.1) and extension for 3 minutes at 70°C for Y cycles. [Denaturation and annealing times were reduced to 45s in the Geneamp 9600 machine because of the thinner walled reaction tubes.] This was followed by a single cycle chase of 1 minute at the annealing temperature and 10 minutes extension. Cycle numbers are given in the appropriate text and figure legends. At all times precautions were taken to ensure that the reagents and materials used in the PCR were kept separate from general laboratory chemicals and that each PCR reaction was set up in a laminar flow hood to prevent aerial contamination.

MVR-PCR. Two state MVR-PCR reactions were performed as described in Jeffreys et al (1991b) using 50ng input of human or transgenic mouse genomic DNA. Four state MVR-PCR was performed on 50ng of mouse DNA as described in Tamaki et al. (1993).

Vectorette PCR. Vectorette PCR (Arnold et al., 1991) was used to recover PCR products containing the mouse DNA flanking the insertion sites of the transgenes. Vectorette libraries were produced for each of the eight transgene insertions. Detailed descriptions of the products obtained are presented in results chapter five. The basic protocol is presented here.
Vectorette linker production. The sequences of the vectorette top and bottom strands are detailed in fig.2.1. 10μg of vectorette top strand and 10μg of vectorette bottom strand were annealed at 60°C for 30 min in 20μl 0.1M NaCl, 10mM MgCl₂, 10mM Tris-HCl (pH 7.6), to produce a complete vectorette linker. Neither vectorette oligonucleotide was 5’ phosphorylated.

DNA digestion, preparation, and ligation. Approximately 2μg of genomic DNA digested with a six cutter restriction enzyme were ligated to 0.2μg of annealed vectorette linker in 20μl 50mM NaCl, 20mM Tris-HCl (pH 7.6), 5mM DTT, 0.1mM ATP, with 2 units of T4 ligase (Gibco BRL) overnight at 15°C for sticky end ligations and 4°C for blunt ended ligations. The amount of vectorette linker in the reaction was increased to 2μg for total genomic DNA ligations cleaved with a restriction endonuclease recognising a 4bp cleavage site. Control ligations without DNA were also prepared and ligation monitored by removal of a 5μl aliquot and addition of 0.5μg of λ DNA cleaved with Hind III then agarose electrophoresis after ligation. Ligations (“vectorette libraries”) were diluted with 5mM Tris-HCl (pH 7.6) to 5μg/ml genomic DNA and stored at -20°C.

Vectorette PCR amplification. 10ng vectorette library genomic DNA was diluted to 10μl with water, denatured at 96°C for 3 min and then held at 80°C. 10μl of 2x concentrated PCR buffer, primers and Taq polymerase (Cetus Amplitaq) were added, to give a final concentration of 1μM transgene primer, 1μM vectorette primer and 0.06 units μl of Taq polymerase. PCR buffer was as described previously (Jeffreys et al, 1991b). Amplification was at 96°C 1.2 min, 68°C 1min, 72°C 3 min for 30 cycles on a Perkin Elmer Cetus thermal cycler. Vectorette PCR products were detected by agarose gel electrophoresis followed by Southern blot hybridisation or ethidium bromide staining and visualisation by long wavelength UV. Appropriate DNA fragments were recovered by electroelution onto dialysis membrane and re-amplified with vectorette and transgene primers for 25-30 cycles depending upon yield. PCR products were repurified by agarose gel electrophoresis and electroelution followed by ethanol precipitation then sequencing.

Sub-vectoretting of vectorette PCR products. The full sequence of vectorette PCR products over 500bp in length was obtained by the production of sub-vectorette fragments. Full or partial digestions with enzymes cleaving within the initial vectorette product were performed on 200-300ng of initial product. The digested DNA was phenol extracted and recovered by ethanol precipitation. 20ng of blunt ended vectorette linker was annealed to 10-15ng DNA fragments in a 20μl reaction using the previous ligation conditions. Re-amplifications were performed on 1μl of a 1:20 dilution of this sub-vectorette library for 25 cycles using the second vectorette primer plus either the transgene primer or the first vectorette primer. PCR products were repurified by agarose gel electrophoresis and electroelution followed by ethanol precipitation then sequencing.

Amplification of STRs. Simple tandemly repeated loci were amplified from 50ng of cell line DNA for 18-20 cycles, using one primer end-labelled (1.5pmol primer per subsequent PCR reaction) with [γ ³²P] ATP using T4 polynucleotide kinase. This primer and 10pmol of unlabelled primer were used with 0.1 units of Taq polymerase were used in a 10μl PCR. 3μl of these reactions were denatured by boiling then separated by polyacrylamide gel electrophoresis, followed by autoradiography. If the STR was derived from an expanded Alu polyadenylate tract then best results are achieved with the end-labelling of the non-Alu primer (Armour et al., 1994). Samples were run alongside a Sequenase sequencing reaction M13 phage DNA using M13 17mer primer (Sequenase sequencing Kit) to provide a size reference.
Figure 2.1 The oligonucleotides used in the preparation of the vectorettes.

Vec1 top - 64nt - AvaII degenerate

5'-g^5/gcagctggagatgtgcagatgatctagtttagtttaacctacctacccgaccgagctg-3'

Vec1 top - 64nt - Mbo I

5'-gtcagctggagatgtgcagatgatctagtttagtttaacctacctacccgaccgagctg-3'

Vec1 bottom - 60nt 3'-tcggaccttcacctgctcttgaggatctgtgtccaggacggtgctgtaac-5'

Vec1 primer - 26nt 3'-ctggatgtggagatctgtcagg-5'

Vec3 top - 64nt - Blunt ended

5'-gtcagctggacgctgtaattttaacctacctagctccgacagttccattgc-3'

Vec3 bottom - 64nt 3'-ctagctacggacggacggctgtaatccctccggtcttcgtaacg-5'

Vec3 primer - 24nt 3'-tcggatccgcaacctgagtcgg-5'

Vec1 = vectorette linker 1
Vec3 = vectorette linker 2
Small pool PCR (see also chapter 6). The purpose of this PCR which uses minisatellite flanking primers to amplify across the whole locus from multiple small aliquots of DNA, is to measure the mutation rate to new length alleles from single molecules of germline or somatic DNA. When an aliquot of these PCR reactions are electrophoresed on 40cm gels and Southern blotted before hybridisation with the minisatellite probe, new mutant molecules will appear as novel length bands (see chapter 6, figure 6.1). The DNA to be assayed is digested with an enzyme that does not cleave within the minisatellite or transgene locus and the concentration of the DNA is accurately measured. Assuming that 3pg is the amount of DNA representing one haploid genome, then reactions can be set up from a master mix aliquoted into several tubes, in which each tube has only the amount of DNA equivalent to 40, 80, 160, or 320 genomes worth of DNA. The PCR reactions are performed in the Geneamp 9600 (Perkin Elmer Cetus) for 96°C 45sec, 68°C 45sec, 70°C 3 min. The annealing temperature for the experiments in this thesis was always 68°C, but other annealing temperatures can be used.

Computing. DNA sequences were analysed by an IRIX Mainframe computer, operating the genetics Computer Group Sequence Analysis Software Package version 6.2, programs developed at the University of Wisconsin (Devereux et al., 1984). Digital MVR-PCR data was analysed with software written by Alec Jeffreys (Jeffreys et al., 1991b and unpublished) in VAX BASIC V3.4 and Microsoft QuickBasic, running on a VAX 8650 Mainframe and an Apple Macintosh Quadra 700 respectively.
Chapter 3

MVR-PCR at MS32 for a series of families of Asian origin

Summary

In addition to a high level of allele length variation observed at the human minisatellite MS32 by Southern blot pedigree analysis, a new technique has been developed which accesses the internal repeat unit sequence variation at this locus. This technique is called MVR-PCR and is an extension of a previous enzyme-based internal repeat unit variation mapping technique. MVR-PCR is a method which utilises the presence of a G/A polymorphic site in the 29bp MS32 repeat unit, to direct repeat unit specific primers in separate PCR amplifications, in conjunction with a constant primer rooted in the repeat unit flanking DNA. These reactions produce a ladder of products of increasing size representing extensions from the flanking DNA to each repeat unit of a specific type. This series of products is separated by agarose gel electrophoresis and visualised by Southern blotting and MS32 repeat unit probe hybridisation, followed by autoradiography. Initially MVR-PCR revealed higher levels of allelic variation and a higher heterozygosity than previously seen with Southern blot length analysis, for Caucasians from Northern Europe. In this chapter MVR-PCR analysis is extended to compare the heterozygosity and repeat unit type maps of alleles from a number of families of Asian (Bangladeshi) origin with the MVR-PCR patterns seen for the Caucasian alleles by allele alignment techniques. MVR-PCR for these Asian families revealed that the high levels of allele variation previously seen at the MS32 locus by MVR-PCR mapping are not limited to Northern European populations, but additionally extend to alleles of an Asian origin. The rate of homozygosity identified from the Asian families was higher than that expected from the predicted population size, due to the presence of one homozygous individual. No population specificity of Asian alleles was observed and pedigree analysis to obtain single alleles from diplodid code results did not reveal any mutation events within the Asian families.

Introduction

The MS32 locus. The locus was first cloned from a λ library constructed from Sau 3AI size fractionated human DNA (5-15kb) by cross hybridisation with the multi-locus DNA fingerprinting probe 33.15. MS32 detects a hypervariable locus with alleles ranging from 2-30kb, with an allele length heterozygosity of 97.5% based on detectable differences in allele length using standard Southern blotting analysis (Wong et al., 1987). In situ hybridisation and linkage mapping localised MS32 to an interstitial region on the long arm of chromosome 1 (1q42-43) and the locus name D1S8 was assigned (Royle et al., 1988). Sequence analysis of MS32 revealed that it is composed of a tandem array of a 29bp GC rich sequence. This region has expanded from within a region of DNA which shows homology to a retroviral long terminal repeat (LTR) (Wong et al., 1987; Armour et al., 1989).

The high heterozygosity of this locus is due to the existence of multiple allele lengths resulting from variation in the numbers of 29bp repeat units in an allele. These new length alleles are generated by a high germine mutation rate, estimated to be 1% per gamete by pedigree analysis (Jeffreys et al., 1991). Allelic repeat copy number at MS32 can vary between 12 and over 800 repeat units (Wong et al., 1987; Armour et al., 1989).

Chapter 3, page 1
Internal sequence variation at MS32. In addition to variations in repeat unit copy number, resulting in allele length polymorphism, most minisatellite loci show internal variation of repeat unit sequence and the repeat unit arrays at these loci are composed of an interspersed mixture of the different repeat unit types (Capon et al., 1983; Owerbach and Argaud, 1984; Jarman et al., 1986; Wong et al., 1987; Nakamura et al., 1987; Gray et al., 1991). Previous analysis of the human minisatellite MS32 revealed the presence of a single base substitution polymorphism (A/G), present in the 29bp repeat unit sequence, which creates or destroys a Hae III restriction site (Jeffreys et al., 1990)(fig. 3.1 A). The interspersion patterns of Hae III + and Hae III' repeat units across an MS32 allele represent a (Minisatellite Variant Repeat) MVR-map of the internal repeat unit array of an individual MS32 allele.

The pattern of interspersion of the Hae III+/- repeat unit types across an MS32 allele was initially assayed using DNA from a PCR amplification across the entire MS32 repeat unit array with flanking primers, then a partial Hae III digestion of this product (Jeffreys et al., 1990). Southern blot length analysis using MS32 had previously identified around 50 alleles (Jeffreys et al., 1990), indicating a hypervariable locus, but the enzyme-based MVR mapping technique provided the first insight into the exceptionally high levels of internal repeat unit variation at MS32. Some length alleles previously thought to be identical usually showed extreme variation in their MVR patterns indicating widely diverged ancestral origins (Monckton et al., 1991). When pairwise comparisons of alleles were made, using dot matrix analysis, it was found that there was a marked gradient of variability across the MS32 repeat unit array. Several groups of alleles shared blocks of repeat unit patterns (haplotypes) at one end of the array, but showed differences at the other end of the array. This suggested for the first time a polarity in the variation at the locus, and the possibility that one end of the array was mutating at a higher rate than the other (Jeffreys et al., 1990). This idea has been strengthened by analysis of seven mutant MS32 alleles identified during extensive pedigree analysis, where the changes in MVR-patterns are at the 5' end of the array (Jeffreys et al., 1991b). In Jeffreys et al., (1990), the hypervariable end of the repeat unit array was referred to as the 3' end. In subsequent papers and in this thesis, however, the hypervariable end is referred to as the 5' part of the array.

MVR mapping of MS32 using PCR. Whilst enzymically based MVR mapping proved very informative, it was limited in the size of alleles that it could assay, <6kb. Since small alleles are not common at the MS32 locus this meant that the dataset was biased. The technique was also laborious and technically challenging, and a more efficient approach to the recovery of MVR data from a wide range of MS32 alleles was required. These problems were overcome by the development of MVR-PCR, a completely PCR based assay for repeat unit type across an MS32 allele.

The theoretical basis for MVR-PCR is outlined in figure 3.1, A, B. In summary the method involves the use of variant repeat unit specific primers in conjunction with a flanking primer rooted in the adjacent single copy DNA to produce a ladder of PCR products corresponding to the position of each a-type and t-type repeat (Jeffreys et al., 1991b). The variant repeat specific primers have 20nt of repeat unit specific sequence and a 20nt 5' extension (Tag). The primers are identical except for the extreme 3' base which corresponds to the presence of an 'A' or 'G' at the Hae III position in the repeat. Between 66-72°C these primers will only extend from repeat units that they match perfectly to the 5' MS32 flanking DNA. 32-TAG-A primes from the Hae III+ (a-type) repeats and 32-TAG-T primers from the Hae III' (t-type ) repeats. Separate amplifications between each MS32 repeat unit
Figure 3.1A. The sequence of the 29bp MS32 repeat unit showing the *Hae* III polymorphic site and the sequences of the two repeat unit specific primers, 32-TAG-A and 32-TAG-T. These two primers are identical except for the 3' most base which is repeat unit specific. 32-TAG-A primes from repeat units with a 'G' at the polymorphic site and 32-TAG-T primes from repeat units with an 'A' at the polymorphic site. Both primers carry a 20nt 5' extension called TAG.

B. Principles of MVR-PCR amplification. At low concentrations 32-TAG-T (blue shading) will anneal to approximately one t-type repeat unit (blue shaded boxes) per molecule and extend out towards the 5' flanking DNA. Amplification from the flanking primer (black shading) creates a sequence complementary to TAG (red shading) when the 20nt extension is copied. The relatively high concentration of the TAG primer means that in subsequent rounds of amplification it will prime in preference to internal priming of the initial products by 32-TAG-T, to give a ladder of products extending to each t-type repeat in the array. An example of the profile type expected from single allele internal mapping is shown in 'i' and 'ii'. The ternary code pattern expected from genomic DNA with the alleles from 'i' and 'ii' is shown in 'iii'. Code 1 and code 2 bands are shaded black and codes 3, 4 and 5 are hatched, reflecting a lower intensity of signal. The last two positions in the diploid code are 4 and 5, reflecting the fact that allele 1 is two repeat units longer than allele 2. This figure was adapted from D.Neil, 1994, Ph.D thesis.
specific primer and a flanking primer rooted in the single copy MS32 DNA will generate two complementary sets of PCR products extending from the flanking primer to each repeat in the array. These are resolved by electrophoresis in adjacent lanes, Southern blotting and hybridisation with an MS32 probe, followed by autoradiography.

The Tag primer is included in the reaction to prevent progressive shortening of PCR products at each cycle due to repeat unit specific primers priming internally within PCR products. The repeat unit specific primers are used at low concentrations in comparison with both the flanking MS32 primer and the Tag primer. This means that in each cycle of PCR the repeat unit specific primers prime from each repeat unit to produce a product extending out to the flanking primer, but in the subsequent PCR cycle the flanking primer will copy this product and the 5' Tag extension on the repeat unit specific primer, thus enabling the subsequent amplification of this product by the more abundant Tag primer, in preference to the repeat specific primers. Although shorter products are amplified more favourably in later rounds of PCR, this is compensated for by increased hybridisation efficiency of longer products.

**Diploid coding.** MVR-PCR can be applied to both single allele DNA, recovered by agarose gel electrophoresis of an $Mbo$ I digest of total genomic DNA, then electroelution, or more conveniently, directly from total genomic DNA to produce a diploid ternary code derived from the superimposed patterns of the two alleles (figure 3.1 B). Each rung on a diploid code ladder may be coded as a type 1 (both alleles have an a-type repeat at that position, a/a), type 2 (both alleles have a t-type repeat at that position, t/t) or a type 3 (one allele has an a-type repeat and the other has a t-type repeat at that position, a/t).

**Null repeat units.** During MVR-PCR some repeat units fail to amplify, with either the a-type or t-type specific primers. These are referred to as 'null' or O-type repeats, and they indicate the existence of rare variants of the MS32 29bp repeat (Jeffreys et al., 1991b; Tamaki et al., 1992). In an MVR-PCR profile these null alleles present as gaps in the ladder for a single allele code, and in diploid coding they are identified by a reduction in intensity of the bands. For example, an a/a position will be twice as intense as an a/O position. This introduces three new coding states for diploid mapping; heterozygous a/O is referred to as state 4, heterozygous t/O state 5, and homozygous null O/O state 6. So far three null repeats have been sequenced and found to share the same polymorphism, a single base deletion (Tamaki et al., 1992). This defines another class of repeat units called N-type repeats which can also be assayed for using an N-type repeat specific primer (32-TAG-N) in the MVR-PCR and have been found to comprise the majority of O-type repeats (84%) (Tamaki et al., 1992). One other null repeat, J-type has been sequenced in two Japanese individuals, but is not commonly seen. The remaining O-type repeats are termed U-type, or unamplifiable by either N-type or J-type specific primers (Tamaki et al., 1992).

The presence of one short allele in a diploid code can be identified by the absence of type 3 (a/t) repeat positions and a characteristic long run of apparent type 4 and type 5 null repeats after the end of the short allele (Jeffreys et al., 1991b). These are not 'bona fide' null repeat units but really hemizygous type 1 and type 2 positions at the long allele. These can be reliably identified by pedigree analysis to elucidate single parental alleles (Jeffreys et al., 1991b).
**Individual allele codes.** The structure of individual MS32 alleles can be obtained by recovery of single allele specific DNA by conventional electroelution of restricted genomic DNA, from the appropriate region on an agarose gel. This is time consuming and requires at least 5μg starting DNA, which may not be available, particularly for forensic applications of MVR-PCR. It is also difficult when the alleles are close in size by Southern blot length analysis, and although some of these problems can be resolved by single molecule dilution followed by PCR recovery (Monckton et al., 1991) this only applies to small PCR amplifiable alleles. Alternatively if family groups are available, pedigree analysis of the diploid codes can reveal single allele codes. For a family with a single child, the binary codes of all alleles segregating within that family can be extracted sequentially along each repeat unit position of the MVR code. For each position the codes of the father, mother and child are checked in a look-up table to determine if any exclusions exist and if not to determine the repeat unit types transmitted from each parent to the child. For example if at a given position the paternal code is type 1 (a/a) and the maternal code is type 2 (t/t) then the child’s code (without exclusions) must be type 3 (a/t) with the a-type repeat from the father and the t-type repeat from the mother. Therefore the paternal alleles (1 and 2) must carry an a-type repeat at that position and the maternal alleles (3 and 4) a t-type repeat unit. In contrast paternal code (a/a), maternal (t/t), giving child (t/t) would indicate a paternal exclusion. In this way the repeat unit type can be elucidated for each position except when the father, mother and child are all heterozygous (a/t) at the same repeat position. Therefore with only one child available for analysis the elucidated single allele codes will have some ambiguous positions. Unambiguous codes can be obtained from two children who share only one allele in common (Jeffreys et al., 1991b). Recently 5 polymorphic sites have been identified in the MS32 5’ flanking DNA (Monckton et al., 1993, 1994). These can be used to assay for flanking heterozygotes and allele specific primers for MVR-PCR have been developed.

This work. In this chapter I present the diploid coding and MS32 5’ flanking DNA haplotype analysis from a series of seven families of Asian (Bangladeshi) origin. Single allele codes were retrieved by pedigree analysis and the results of a pairwise allele alignment of the Asian alleles with alleles already present in the MS32 allele database are discussed. Prior to the MVR-PCR coding of the Asian MS32 alleles, only alleles from Northern European Caucasians had been studied (Jeffreys et al., 1991b; Monckton 1992). The data in this chapter provided the first examples of MVR-PCR patterns at the MS32 locus in a different population group. The pedigree analysis performed for these Asian families also provided the potential for identifying mutation events at the MS32 locus, at which the mutation rate per gamete was estimated to be ~1%.
Figure 3.2 Diploid coding analysis for one Asian family. 9 = a-type and t-type repeat unit lanes for a control individual with a known MVR-PCR code for identifying the repeat unit positions on the autoradiograph. 8 = mother, 1 = father, 2-7 = children. Children 2, 3 and 4 share the same MVR-PCR diploid code. 5= negative control. The MVR codes are from 50ng of total genomic DNA with the flanking primer 32O for 18 cycles.
inherit the same MS32 alleles from their parents. In 3 of the 7 families, 3 different diploid codes were seen in the children and the rest of the families showed 2 different codes in their children. From 47 individuals a total of 31 different diploid codes were seen (figure 3.3). This is a reflection of the high level of variation at the MS32 locus. Positions which could not be scored reliably, for example very short products unable to hybridise efficiently with the probe, were scored as a "?". One individual, ‘G’, a father from one family gave what appeared to be a homozygous diploid code consisting solely of type 1, 2 and 6 positions. This homozygous pattern was seen when the MVR-PCR was repeated using 32B and 32D as alternative flanking primers (see figure 1.1A, chapter 1).

Pedigree analysis for single allele extraction. The individual single allele codes from the seven families were elucidated by computer analysis of groups of family diploid codes. For families with multiple offspring with different combinations of parental alleles the incomplete parental haplotypes (incomplete at positions where mother, father and child are heterozygous, a/t) can be defined separately for each child and then combined to give an unambiguous allele code (figure 3.4). The minimum requirement to achieve unambiguous allele codes is both parental diploid codes and two children sharing one allele in common, which was available for each of the seven Asian families.

Mutation rate. No maternal or paternal exclusions were identified during the pedigree analysis of the seven Asian families. This is not an unexpected result since the MS32 pedigree mutation rate is 1% (Jeffreys et al., 1991b) and only 66 gametes (33 children) have been scored in these analysis. Therefore both the male and female mutation rates per gamete from this study are $<\frac{1}{14}$, $<\frac{1}{15}$, $<\frac{1}{9}$, $<\frac{1}{9}$.

Allele diversity. The diversity of the single alleles obtained by pedigree analysis is shown in figure 3.5. 28 alleles were elucidated from the 14 parents of the seven family groups. All these alleles were different with one exception which was seen twice in this sample. This gives a value of $\theta = \frac{1}{5}$ resulting in a predicted heterozygosity value (1-Heterozygosity, where $\theta = \frac{1}{14}$) of 99.72% for the Asian population (Ewens et al., 1972). The two identical alleles were seen twice in the same individual giving a heterozygosity value of 7.1% (1/14) which exceeds the predicted value for the predicted population size from 14 alleles of 0.28%, therefore it is likely that the two alleles in this individual are the result of a recent consanguineous mating.

There is one short allele of 16 repeat units (T, allele 3), which is reflected by the abundance of 4, 5 and 6 positions in the 3' end of the diploid code from this individual and two of her offspring. The pedigree analysis to identify the short allele in individual 'T' is shown in figure 3.4. A number of other null, O-type repeat units are seen in some of the alleles, one of which ('E', allele 2) was additionally mapped using the N-type specific null primer to reveal that the null repeats in this allele were N-type repeats, the most common of the studied null repeat unit types (Tamaki et al., 1992). There is additionally one allele ('K', allele 2) which is completely homogeneous for a-type repeat units.

Flanking haplotype analysis. Five polymorphic positions previously identified in the 5' flanking MS32 DNA (Monckton et al., 1993, 1994) were assayed from genomic DNA from both parents and 2-3 children carrying different combinations of parental alleles from each of the seven families. The theory, primer positions and expected product sizes for each assay are shown in figure 3.6, and the single allele haplotype results are shown in figure 3.5. The conditions used in each assay are as described previously (Monckton and Jeffreys, 1993; Monckton Chapter 3, page 5
<table>
<thead>
<tr>
<th>Individual</th>
<th>F</th>
<th>M</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A/B-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A/B-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A/B-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C/D-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C/D-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F/F-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F/F-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F/F-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G/H-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G/H-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J/J-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J/J-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J/J-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K/L-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K/L-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N/N-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N/N-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.3 Diploid codes from the Asian families. MVR-PCR on total genomic DNA with flanking primer 32O (sequence, table 2.1) for 18 cycles from 50ng genomic DNA. The 51 different diploid codes are presented in family groups; M = mother, F = father, C = child. Coding usually begins at the second or third repeat (code positions 1 and 2 respectively), since the first repeat unit (**) is weakly detected and cannot be scored reliably. There are 31 codes from 49 individuals since several children in families share diploid codes and duplicate codes are omitted.
<table>
<thead>
<tr>
<th>Individual</th>
<th>Alleles</th>
<th>MVR code position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Father</td>
<td>1 2</td>
<td>77713311111121313311131131133111312313412351331313311331313133</td>
</tr>
<tr>
<td>Mother</td>
<td>3 4</td>
<td>7112131111113111111145144444444444455554545545454464644646464646546465646</td>
</tr>
<tr>
<td>Child 1/1-1</td>
<td>1 3</td>
<td>211313111111131354554544444464645546464644444444454454544444445654</td>
</tr>
<tr>
<td>Child 1/1-2</td>
<td>2 4</td>
<td>21332213133113111131311232123132222313314133531133141315323</td>
</tr>
<tr>
<td>Child 1/1-3</td>
<td>3 5</td>
<td>277753231111311345454545455444444445646464546654646464646464646564</td>
</tr>
</tbody>
</table>

Figure 3.4 Reconstruction of the single allele codes of individual MS32 alleles by pedigree analysis. The paternal alleles are called alleles 1 and 2, and the maternal alleles referred to as alleles 3 and 4. The children show three different diploid codes, corresponding to three of the four possible combinations of parental alleles. The consensus haplotype was extracted by comparing the incomplete haplotypes deduced from the diploid code of each individual child, to cover all uncertain positions. Finally the diploid code of each individual was compared to the code predicted from the two constituent alleles, to ensure no mismatches were present. Allele haplotypes were extracted using software written by A.J.Jeffreys in VAX BASIC V3.4. In this family the maternal allele 3 is a short allele of 16 repeat units, indicated by long runs of type 4, 5 and 6 repeat units in the maternal and 1/1-1 and 1/1-3 children's diploid codes.
Figure 3.5 Single allele codes and 5' flanking DNA haplotypes from the Asian families. For each parent, the two single alleles derived from pedigree analysis are shown, father’s, alleles 1 and 2, mother’s, alleles 3 and 4. The flanking haplotypes from assays of five polymorphisms in the 5' MS32 flanking DNA (refer to figure 3.6) are shown for each allele; H2 = Hump 2 status (C/T), H1 = Hump 1 status (C/G), Hf = Hinf 1 status (+/-), O1 status (G/C), O2 status (G/C). Repeat unit types are shown as a-, t-, O-, or N-type repeats.
<table>
<thead>
<tr>
<th>Parent</th>
<th>H2</th>
<th>H1</th>
<th>Hf</th>
<th>O1</th>
<th>O2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>C</td>
<td>G</td>
<td>-</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 1</td>
<td>C</td>
<td>G</td>
<td>-</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 2</td>
<td>C</td>
<td>G</td>
<td>-</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 3</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 4</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 1</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 2</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 3</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 4</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 1</td>
<td>T</td>
<td>C</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 2</td>
<td>T</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 3</td>
<td>C</td>
<td>C</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 4</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>H</td>
<td>C</td>
<td>C</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 3</td>
<td>C</td>
<td>C</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 4</td>
<td>C</td>
<td>C</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>I</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 1</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 2</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 3</td>
<td>C</td>
<td>C</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 4</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>J</td>
<td>C</td>
<td>C</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 3</td>
<td>C</td>
<td>C</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 4</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>K</td>
<td>T</td>
<td>G</td>
<td>-</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 1</td>
<td>T</td>
<td>G</td>
<td>-</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 2</td>
<td>T</td>
<td>G</td>
<td>-</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 3</td>
<td>C</td>
<td>C</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 4</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>L</td>
<td>C</td>
<td>C</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 3</td>
<td>C</td>
<td>C</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 4</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>M</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 1</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 2</td>
<td>T</td>
<td>G</td>
<td>-</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>N</td>
<td>T</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 3</td>
<td>T</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>allele 4</td>
<td>C</td>
<td>G</td>
<td>+</td>
<td>G</td>
<td>G</td>
</tr>
</tbody>
</table>

(homozygote)
The Hinfl assay tests for a C(Hf+) to T(Hf−) transition at position 143 (bp from the start of primer 32B) in the MS32 5' flanking DNA. PCR products from genomic DNA amplified between 32OR and 32B were assayed by restriction with Hinfl to test for the C/T polymorphism. All alleles give a constant 169bp band, Hf− alleles give a 197bp band, and Hf+ alleles which are cut by Hinfl I give two additional bands of 139bp and 58bp.

The Hump 2 polymorphism does not affect a restriction site and is therefore assayed by the use of polymorphism specific primers. A C to T transition at position 241 in the MS32 5' flanking DNA, is assayed by specific primers 32-H2C for the 'C' and 32-H2AR for the 'T' in conjunction with the regular flanking primers 32OR and 32B. All alleles give a constant 361bp (32B to 32OR) product, H2C alleles prime with H2C to give an additional 141bp product, and H2G alleles prime with H2AR to give a 259bp product.

With Hump 1, a C to G transversion at position 80 in the MS32 5' flanking DNA is assayed with a mismatch primer, 32-H1B which under low stringency conditions (60°C annealing temperature) forces the 3' misincorporation of a 'G' rather than the 'A' that is present in genomic DNA. This base change creates a Bsp 1286I site in alleles with a 'G' at position 80, but not for alleles with a 'C' polymorphism. The low annealing temperature used to force 32-H1B amplification does not work directly from total genomic DNA and requires seed DNA from a 32OR to 32B preamplification. H1C alleles give a 107bp product, and in H1G alleles this product is cleaved by Bsp 1286I to give two bands of 22bp and 85bp.

The O1 assay tests for a G to C transversion at position 412. It requires the use of polymorphism specific primers. Amplifications between 32B and O1RB, which forces a mismatched C into the final product are then cleaved with Bsr NI to give a constant band of 158bp. O1G alleles give an additional 199bp band and O1C alleles give additional 174bp and 25bp bands.

The O2 assay tests for a G to C transversion at position 422. It requires the use of another polymorphism specific primer, O2RH which in combination with 32B forces a mismatched C into the product. On digestion with Hae III a constant band of 158bp is produced, O2G alleles give an additional band of 209bp, and O2C alleles give two additional bands of 184bp and 25 bp.

The pedigree analysis of each family to elucidate single alleles has already given information as to which children carry which parental alleles. Therefore by logic the flanking haplotypes for each single allele can be elucidated and are shown with the single alleles (figure 3.5).
Chapter 3, figure 3.6
et al., 1994). The flanking polymorphism frequencies of the Asian family alleles are compared to data for Caucasian, Japanese and Afro-Caribbean alleles in Table 3.1. (polymorphism typings for these population groups were performed by D. Monckton, A. MacLeod, K. Tamaki, R. Neuman, T. Guram, N. Fretwell, and A. Jeffreys).

Although there is only a very small sample size of Asian alleles, it is clear that the flanking polymorphism frequencies in this population group do not vary considerably from those seen with the other population groups. In particular the Hump 1 and Hinfl, polymorphism frequencies appear consistent across all four population groups. The Hump 2 polymorphism shows more variation in frequency across the population groups, with a higher frequency of the H2C in Japanese and Asians. Two apparently population specific polymorphisms can be identified for Afro-Caribbeans, which are the only population group to show a high frequency of the O1C polymorphism (f=0.21) and the only group to show the O2C polymorphism.

Table 3.1 Frequencies of the five 5' flanking polymorphisms in four population groups.
(The numbers in brackets represent the total numbers of alleles typed)

<table>
<thead>
<tr>
<th>Locus</th>
<th>Polymorphism</th>
<th>Caucasian</th>
<th>Japanese</th>
<th>Afro-Caribbean</th>
<th>Asian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hump 1</td>
<td>G</td>
<td>.69 (55)</td>
<td>.80 (159)</td>
<td>.81 (70)</td>
<td>.82 (23)</td>
</tr>
<tr>
<td>freq. (nos.)</td>
<td>C</td>
<td>.31 (25)</td>
<td>.20 (41)</td>
<td>.19 (16)</td>
<td>.18 (5)</td>
</tr>
<tr>
<td>Hump 2</td>
<td>C</td>
<td>.59 (94)</td>
<td>.91 (184)</td>
<td>.57 (50)</td>
<td>.79 (22)</td>
</tr>
<tr>
<td>freq. (nos.)</td>
<td>T</td>
<td>.41 (66)</td>
<td>.09 (18)</td>
<td>.43 (38)</td>
<td>.21 (6)</td>
</tr>
<tr>
<td>Hinfl</td>
<td>+</td>
<td>.81 (129)</td>
<td>.81 (163)</td>
<td>.70 (61)</td>
<td>.82 (23)</td>
</tr>
<tr>
<td>freq. (nos.)</td>
<td>-</td>
<td>.19 (31)</td>
<td>.19 (163)</td>
<td>.30 (26)</td>
<td>.18 (5)</td>
</tr>
<tr>
<td>O1</td>
<td>G</td>
<td>.99 (792)</td>
<td>1.0 (166)</td>
<td>.79 (94)</td>
<td>1.0 (28)</td>
</tr>
<tr>
<td>freq. (nos.)</td>
<td>C</td>
<td>.0025 (2)</td>
<td>0</td>
<td>.21 (25)</td>
<td>0</td>
</tr>
<tr>
<td>O2</td>
<td>G</td>
<td>1.0 (800)</td>
<td>1.0 (166)</td>
<td>.92 (119)</td>
<td>1.0 (28)</td>
</tr>
<tr>
<td>freq. (nos.)</td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>.08 (9)</td>
<td>0</td>
</tr>
</tbody>
</table>

Discussion

**MS32 hypervariability.** The large numbers of different alleles present in the small Asian sample reflect the extreme level of allelic variation at the MS32 locus and show that the high level of heterozygosity seen initially (Jeffreys et al., 1991b) is not limited to the Caucasians. Subsequently similar patterns of variability have been observed for MS32 alleles from Japanese, Afro-Caribbean and Malaysian populations (Tamaki et al., 1994; Monckton et al., 1994; C.L.Koh, unpublished data). It is important for forensic applications that the levels of heterozygosity are similar for all population groups, which is the case for MS32 with the exception of a common short allele in the Afro-Caribbeans (see chapter 9). The potential number of diploid codes arising from different combinations of MS32 alleles is huge, a fact which highlights the potential importance of MVR-PCR in individual identification. Given the current world population of 5 x 10^9 individuals and the 1% per gamete mutation rate at the MS32 locus, the numbers of different alleles in the population can be predicted to be >10^3. MVR-PCR could
in theory distinguish between $7 \times 10^{23}$ different allelic states by comparison solely of $a$, $t$, and $O$ codes over the first 50 repeat units (Jeffreys et al., 1991b).

Pedigree analysis. The elucidation of single alleles by pedigree analysis of diploid codes is useful in paternity analyses, where paternal exclusions in the codes would indicate non-paternity. Caution must be exercised with data from a single locus (MS32) since the mutation rate at this locus is 1% per gamete, and paternal exclusion may be the result of mutation in some cases. This problem can be solved by additional MVR-PCR analysis at other minisatellite loci, for example MS31A and MS205 (Neil et al., 1993, Armour et al., 1993). However, this procedure does provide a useful source of de novo mutations at the MS32 locus, for analysis of MS32 mutation processes. No maternally or paternally derived mutation events were seen in the Asian families giving a mutation rate of 0/66 gametes. PCR methods of accessing de novo mutation have been recently developed (Jeffreys et al., 1994) but these use sperm DNA as a germline source and can therefore only measure mutation rates in the male germline. MVR-PCR diploid code pedigree analysis however, can measure both male and female germline mutation rates. Pedigree analysis data from the CEPH families has shown a Caucasian mutation rate of 7/720 gametes or ~1%. Four of these events are male derived and three are female derived, which does not indicate that there is a sex specific mutation bias (Jeffreys et al., 1991b).

Heterozygosity at MS32. Southern blot length analysis of minisatellite alleles can be inaccurate in distinguishing between homozygotes and heterozygotes where both alleles are of a similar size (Devlin et al., 1990). Internal repeat unit mapping of MS32 provides an objective method for distinguishing true homozygotes from pseudohomozygotes (Monckton et al., 1990). Genuine MS32 homozygotes will have diploid MVR-PCR codes restricted to $1$ (a/a), $2$ (t/t), and $6$ (O/O), for example Asian individual 'G' (figure 3.5). In the main MS32 diploid code database three other individuals gave true homozygote codes, from 500 individuals typed, giving an observed heterozygosity of 99.2%. This contrasts with the observed 97.5% heterozygosity from allele length analysis by Southern blotting (Wong et al., 1987). The observed heterozygosity from the Asian families is 93% (13/14 parental heterozygotes) though the small number of individuals typed make this a provisional estimate of the true heterozygosity which is biased by the presence of one homozygous parent out of fourteen parents. In comparison with the massive allele variation usually seen at the locus, it would seem that the majority of MS32 homozygotes represent recent consanguinity, since the actual numbers of homozygotes seen are higher than the predicted heterozygosity for the locus (Jeffreys et al., 1991b).

Homozygous diploid coding does not guarantee a true homozygote. There is the possibility that heterozygosity may not be seen if it arises towards the 3' end of the MS32 repeat unit array, if the 5' repeat unit haplotypes are identical, though in general the polarity of variation at the 5' end of the array would suggest that such alleles would be rare. There is also the potential for 'knockout' of the 5' flanking primer used in the MVR-PCR, due to changes in the 5' flanking sequence of one allele. MS32 alleles of different lengths not identified by PCR due to a 5' flanking DNA change could be identified by conventional Southern blot analysis. For Asian family individual 'G', the homozygous pattern is seen with three different flanking primers, 32B, D and O, therefore this individual is likely to be a true homozygote. The existence of a heterozygous individual with a large deletion of 5' flanking DNA, including all MVR-PCR flanking primer sites cannot be discounted, but no such individual has yet been identified and is likely to be extremely rare. Finally, if a 'null' allele was present in Asian individual 'G' then the
chance that the null allele is transmitted to any of this individual's five children (only three different codes were seen, figure 3.3) is $1/2^5$ or 0.03. No null alleles are transmitted from Asian parent 'G' and therefore this individual is a true homozygote.

5' flanking haplotype polymorphisms. Not only do these polymorphisms provide a detailed haplotype survey of the 5' flanking DNA of the Asian MS32 alleles, which may be useful in tracing ancestral origins of Asian population MS32 alleles, they also provide a system for allele specific MVR-PCR. Allele specific primers can be used as the MS32 flanking primer in MVR-PCR reactions from total genomic DNA (Monckton et al., 1993, 1994). For example 32D2 (figure 3.6) spans the $Hinf_1$ polymorphism and by using 32D2 as a flanking primer in MVR-PCR it is possible to amplify only alleles that carry the $Hinf_1$ polymorphism, knocking out $H^f$ alleles and resulting in a single allele MVR PCR code. The Afro-Caribbean alleles are the only group of alleles identified so far as having two apparently population specific flanking polymorphisms, the O1C and the O2C polymorphisms. This will be discussed further with relation to mutation processes in chapters 6 and 9.

Allele alignments. Although the majority of the MS32 alleles which have been MVR-mapped have unique repeat unit arrays, there are groups of alleles which show regions of relatedness (Jeffreys et al., 1990, 1991b and A. Jeffreys personal communication). Most of the mapped alleles are of Caucasian, Japanese and Afro-Caribbean origin, and some of these (~70%) fall into alignable groups, while the remainder do not align with any other alleles in the database. 17 of the 27 Asian alleles mapped in this study can be aligned with other alleles in the main allele database. 5 of these 17 alignable Asian alleles fall into two large aligned allele groups, each comprising around 10% of the total alleles mapped, with examples of alleles from Caucasian, Japanese, Afro-Caribbean, Malaysian, Mormon and Asian populations. These two groups of alignable alleles are thought to represent two relatively ancient MS32 haplotypes (Monckton, 1992). The remaining 12/17 alignable Asian alleles fall into 8 smaller groups of aligned alleles. The other alleles in these groups are either from mixed population sub-groups or consist solely of Japanese or Caucasian alleles. Three of these allele alignment groups are shown in figure 3.7. Groups A, B and C represent Asian alleles which align to otherwise Caucasian allele groups. All the alleles presented in this figure are different, but they show regions of MVR map similarity through which they are computationally aligned. As noted previously for Caucasians (Jeffreys et al., 1991b) the regions of alignment of the Asian alleles with other alleles from the database show a polarity of variation, concentrated at the 5' end of the alleles. This provides further evidence in the Asian population of a proposed 5' mutation hotspot, potentially regulated by mutation modulators acting upon the 5' region of the MS32 locus (Jeffreys et al., 1991b). Internal differences within regions of haplotypic similarity mostly comprise small repeat unit insertions or deletions or short patches of repeat unit type switching, possibly by gene conversion-like events.

In the main MS32 allele database, most alleles in aligned groups also share flanking haplotypes (Monckton, 1993b). Nevertheless some related alleles show differences in their flanking haplotypes, for example in figure 3.7, the Bangladeshi alleles in groups A and C have different Hump1 and Hump2 polymorphisms than the most closely related Caucasian alleles. In group C all the Bangladeshi alleles share the same flanking haplotype which differs from all the Caucasian alleles in this group. It is possible that this represents a population specific change in the 5' flanking DNA of alleles of this type, though the numbers of alleles of this type sampled are too small to substantiate this idea. There had been some previous evidence for population specific alleles in Caucasian and
Figure 3.7. **Examples of aligned Asian alleles.** Areas of alignment are detailed in red, non-alignable repeat units are in black. For each allele the ethnic origin, either (B) Asian (Bangladeshi), or (C) Caucasian and the flanking haplotypes, Hump1 (H1), Hump2 (H2), Hinf I (Hf), O1 and O2 are presented. The alphabetical code for the Bangladeshi allele (see figure 3.5) is given in brackets after the race label (B). The ends of short alleles and the unknown haplotype of long alleles are shown (<<). Groups A and B show one Bangladeshi allele aligned with a group of Caucasian alleles. The alleles are all aligned, but are also all different. The main region of alignment is at the 3' end, providing evidence for a 5' polarity of variation. Group C shows three Bangladeshi alleles aligned with some Caucasian alleles showing different flanking haplotypes in similar alleles from the two population groups. This change in flanking markers is also seen with the Bangladeshi allele in group A and its most closely related Caucasian alleles.
Japanese populations (Monckton 1992), but as more population groups are typed the evidence for this type of population specificity is reduced and none was observed for the Asian MS32 alleles, which aligned within some of the previously Caucasian specific groups, for example Group A, figure 3.7.

Conclusions
MVR-PCR analysis of this small sample of Asian DNA's has shown that the previously observed properties of high heterozygosity and polarity of variation are not limited to the Northern European Caucasian populations, but extend to alleles from an Asian (Bangladeshi) population sub-group. No population specific traits were seen for the Asian alleles, but the switching of 5' flanking Asian haplotypes in groups of aligned alleles indicates that changes can occur in the 5' flanking DNA. This data in conjunction with MVR-PCR results from a number of different population sub-groups (Japanese, Malaysian, Afro-Caribbean) adds weight to the proposal for a mutation hotspot acting upon the 5' region of the MS32 locus.
Chapter 4

Creation and structural analysis of transgenic mice carrying a human MS32 allele

Summary
As a first stage in dissecting the complex pathways of mutation leading to hypervariability at the minisatellite MS32, it was decided to create transgenic mice carrying an MS32 allele. The aims of this study were two-fold; first to study the insertional behaviour of a repeated DNA transgene and second, to ascertain if the mutation processes at any of the transgenic loci created reflect those of the endogenous human locus and therefore could act as a model system for the study of mutational processes at the MS32 locus. Microinjection of a PCR amplified DNA construct was chosen as the method for creating the transgenic animals, and a total of eight separate insertions of the MS32 transgene construct were recovered from four founder mice. In this chapter the creation and initial identification of mice positive for MS32 construct insertion are detailed and the initial structural analysis of these transgenic loci by restriction digestion and MVR-PCR mapping is presented.

Introduction
MS32. The best studied human minisatellite is MS32 (locus D1S8) which is comprised of a number of different sequence variants of a 29bp repeat unit (Jeffreys et al., 1990). The two major repeat unit types are called a-type and t-type repeats which differ by a single base. MS32 has an allele length heterozygosity of 97.5% (Wong et al., 1987) and a germline mutation rate of 1% per gamete (Jeffreys et al., 1988, 1991b). Allele alignment to identify groups of related alleles shows some variation within groups of otherwise closely related alleles (Jeffreys et al., 1991b; Monckton 1992). These alleles vary in a-type and t-type repeat unit patterns at the 5’ end of the locus. Additionally the MVR-PCR maps of seven mutations identified by pedigree analysis of the MS32 locus in CEPH families show changes at the 5’ end of the mutant allele when compared to the progenitor (Jeffreys et al., 1991b). These facts indicated that a polarity of variation at the MS32 locus existed with a mutation hotspot at the 5’ end of the locus. The biological mechanism which results in this polarity is unknown, but it is possible that the effect is modulated in cis or in trans by ‘elements’ operating on or near the 5’ end of the MS32 locus.

Creation of MS32 transgenics. In order to investigate the biological processes affecting MS32 mutation rate and mechanism, in particular to try to identify the regions of the locus involved in mutational mechanisms, transgenic mice were created harbouring an MS32 allele. Mice were chosen for the transgenesis since they represent a well characterised mammalian system and have no known sequence homology to the MS32 locus and therefore no homologous recombination of insert and mouse genome DNA would be predicted; instead random integration might be expected. The possibility of using MS32 transgenic yeast as a model for the study of MS32 integration and mutation processes was considered, but rejected for a number of reasons. The yeast genome is very streamlined with little non-coding DNA and therefore the integration of an MS32 transgene construct of a few kb is likely to cause a phenotypic change in the yeast which may have an effect on biological processes. An alternative was the
Introduction of a Yeast artificial chromosome (YAC) carrying the MS32 locus. Details of such a YAC have not been published and the construction of a YAC carrying the MS32 locus would be time consuming. YACs have been shown to carry repeated DNA sequences unstably (Neil et al., 1990) and therefore YACs do not currently represent a suitable vector for carrying MS32 DNA for analysis. However the main reason for choosing a mammalian system to study MS32 behaviour is that the yeast genome is very diverged from higher eukaryotic genomes and therefore the mouse was felt to be a more appropriate organism for study.

The technique chosen to create the mice was the direct microinjection of a PCR amplified linear MS32 allele construct into one-cell embryos (Gordon et al., 1980). Microinjection has been extensively used as method of introducing new genetic material into the mouse (Hogan et al., 1986) and usually results in stable chromosomal integration of the foreign DNA in 10-40% of the surviving embryos (e.g. Brinster et al., 1981b; Constantini and Lacy 1981; Lacy et al., 1983). The majority of integrations appear to be at the one-cell stage, since the foreign DNA is usually present in every cell of the transgenic animal, including the primordial germ cells. In 20-30% of integration events the DNA integrates at a later stage resulting in animals mosaic for the insertion (e.g. Constantini and Lacy 1981; Palmiter and Brinster 1986). The number of copies of the foreign DNA sequence that are retained per cell can range from one to several hundred and appears to bear little relationship to the initial number of molecules injected into the egg. The mechanism of integration of microinjected DNA is unknown. When more than one copy integrates, the transgenic DNA constructs may integrate in a tandem array, or at separate positions in the mouse genome (Palmiter and Brinster 1986; Hogan et al., 1986). There is no previous evidence of site-directed integration since microinjected foreign DNA constructs have integrated into different autosomes and into both sex chromosomes (Brinster et al., 1985).

An alternative to microinjection as a means of introducing new genetic material into mouse genome is the use of retroviral vectors to infect embryos (Jahner et al., 1985). Retroviral vectors are constructed containing the DNA for insertion, and infection of pre- or post implantation embryos results in the stable integration of the viral genome, including the chosen insert DNA, into the mouse chromosomes. The process of retroviral integration is well understood and results in single-copy integrations (Hogan et al., 1986). The disadvantages of this technique are that it requires the special construction of a retroviral vector carrying the insert DNA with an 8-10kb insert size constraint. The technique also usually results in mosaic integration of the viral and insert DNA since infection takes place after cell division begins, which requires breeding to obtain pure lines of transgenic animals. Perhaps more importantly for creating a minisatellite transgene, retroviral integration will result in the DNA insertion being flanked by the long terminal repeat sequences of the virus. The MS32 locus is thought to have expanded from human retroviral LTR sequences (Armour et al., 1989), and it is not possible to anticipate the potential interference of the vector LTRs on the biological behaviour of an MS32 transgene.

A third method for the generation of transgenic mice is the introduction of DNA into embryonic stem (ES) cells (Bradley et al., 1984). The DNA can be introduced into the ES cells by various means including retroviral infection, microinjection or electroporation. The major advantage of this technique is that cultured ES cells can be screened after the introduction of DNA for desirable characteristics, prior to manipulating them into an embryo by incorporating the cells into the blastocyst or aggregating them with eight-cell embryos (Lovell-Badge et al., 1985).
The disadvantages of this technique are the requirement for skilled preparation of the ES cells and the subsequent breeding program required to produce pure lines from the mosaic animals created.

Therefore microinjection was used to create MS32 transgenic mice because of the simplicity of insert DNA preparation by PCR, because it would probably produce non-mosaic animals which will transmit the new gene to 50% of the offspring, obviating the need for large breeding programs to obtain pure lines and because there would be no additional retroviral sequences flanking the insert DNA and potentially affecting MS32 mutation processes.

The DNA construct chosen for insertion is shown in figure 4.1. The construct is composed of a 71 repeat unit MS32 allele of Bangladeshi origin. This particular allele was initially chosen because it falls within the normal size range of allele lengths, but is still small enough to be efficiently amplified by PCR. The allele has an informative MVR code with a good interspersion of a-type and t-type repeat units, suitable for detecting subsequent MVR rearrangements such as mutation events. The construct DNA includes 212bp of 5’ repeat unit flanking DNA and 214bp of 3’ repeat unit flanking DNA and was initially amplified with MS32 primers 32A and 32D with attached 20nt synthetic oligonucleotides E1 and E2. The synthetic and regular MS32 repeat unit flanking DNA was included to provide characteristic primer and restriction sites (E1 is cut by Hind III and E2 is cut by Pst I) to aid in transgene insert structural analysis and because it was thought that the MS32 flanking DNA may exert some effects on MS32 mutational processes; the polarity of variation in the human locus at the 5’ end of the repeat unit array, suggested that the 5’ DNA may influence mutation rate and process.

The choice of mouse strain to provide embryos for microinjection depends on whether the genetic background of the mouse has a direct bearing on the subsequent use of the transgenic mice. For example, the analysis of particular immunoglobulin gene constructs or major histocompatibility (MHC) constructs will be greatly facilitated by the use of particular strains with a certain genetic background e.g. an immunodeficient strain such as SCID mice. In most cases, however the genetic background of the egg is not obviously important. For the purposes of creating MS32 transgenic mice, a hybrid, C57BL/6J/CBA/ca F2, was used. This is one of the strains typically used by the group in Cambridge who performed the microinjections, and has a C57BL/6J/functinal genotype, since these eggs are easy to culture in vitro (Allen et al., 1987).

This work. Tail DNA was prepared from the weaned progeny of the recipient females to identify animals positive for MS32 integration events by Southern blot analysis and hybridisation with an MS32 repeat unit probe. Multiple MS32 hybridising bands indicated that there was more than one insertion or multi-copy arrays of structurally changed insert DNAs present in three out of four of the positive founder mice. These animals were bred with negative C57BL/6J/CBA/ca F2 mice to ascertain if germline integration had been achieved and transmission was possible and also to test if different integration events could be segregated into separate transgenic lines. Structural analysis of the transgenes identified was performed by restriction digestion analysis and MVR-PCR mapping of the MS32 transgene repeat unit arrays in an attempt to shed light on mechanisms of integration and rearrangement of tandemly repeated transgenes.

A flow chart representing the overall plan of the MS32 transgenic project is presented in figure 4.2. The various chapters in which data from each stage of the project will be presented are also detailed.

Chapter 4, page 3
**Figure 4.1 A. Schematic of the original DNA construct for microinjection.** An MS32 allele was amplified by PCR from 150ng genomic DNA using the primers E1-A and E2-D for 15 cycles of 96°C 1.4 min, 66°C 1 min, 70°C 10 min. Primer sequences were: E1-A 5'CCCAAGCTTGCCCTAATGGCCgggaattcagacact3' and E2-D 5'TGCGCGGAATTCTGCAGTGGactcgcattgggaatt3'. These primers consist of a 20nt synthetic sequence (uppercase) followed by 20nt of sequence complementary to the MS32 flanking DNA (lowercase). PCR conditions were as described previously (Jeffuys et al., 1991b). A 150 dilution of the product was reamplified with primers E1 and E2 for 16 cycles under the same conditions until the product was detectable by agarose gel electrophoresis and ethidium bromide staining. 500ng of PCR product was electroluted onto dialysis membrane and purified on an NACS column (Allen et al., 1987). The primer sites are shown by arrows and some of the restriction sites are shown. This schematic is not shown to scale.

**B.** The two state MVR-PCR code of the MS32 allele used in the construct showing the interspersion of a-type and t-type repeat units.
**Figure 4.2.** Flow chart representing the overall plan of the MS32 transgenic mouse project, from creation of the mice, through structural characterisation by restriction analysis, MVR-PCR, and flanking sequence analysis, to the use of the transgenes in MS32 mutation studies. Steps b-d were performed by a research group in Cambridge. [(4), (5), (6), (7)] indicates the chapter in which data corresponding to this stage of the project will be presented.

- A - Preparation of MS32 allele DNA construct by PCR (4)

- B - Microinjection of the MS32 allele into one-cell embryos (4)

- C - Recovery of surviving two-cell embryos and replacement into recipient female mice (4)

- D - Preparation of tail DNA from the weaned progeny of the recipient female mice (4)

Identification of positive mice, Southern blot analysis and hybridisation with an MS32 probe (4)

Characterisation of the transgene insertions by restriction mapping analysis (4)

Breeding of three founder mice to segregate individual transgene insertions (4)

MVR-PCR analysis of the transgene insertions (4)

PCR analysis of mutation (6, 7)

Breeding of the individual transgenic lines and pedigree analysis by tail DNA Southern blotting to identify mutation events (6)

Recovery of mutation events and analysis by MVR-PCR (6, 7)

Sequence analysis of the transgene flanking DNA (5)

Effects of gamma irradiation on MS32 transgene mutation rate (7)

Chapter 4, figure 4.2
The experiments described in this chapter represent the efforts of a number of individuals. The microinjection procedures and preparation of tail DNA from founder mice were performed by Mike Norris and Sheila Barton in the laboratory of M. Azim Surani at the AFRC Institute at Babraham, Cambridge. Subsequent DNA analysis and direction of the breeding programme was by a research team at Leicester comprising Andrew Collick, Alec Jeffreys and myself.

Results

Transgenic founder mice identification. The MS32 allele for microinjection was PCR amplified from human genomic DNA using primers 32A and 32D flanking the minisatellite and carrying 5' extensions (E1 and E2) of synthetic oligonucleotides containing diagnostic restriction sites for subsequent transgene analysis (figure 4.1). Approximately 200 copies of the MS32 linear DNA construct PCR products containing the 71 repeat array, plus 212bp of 5' flanking MS32 DNA and 214bp of 3' flanking DNA were microinjected into 554 one-cell embryos of C57BL/6J/CBA/caF2 type. 375 embryos surviving to the two cell stage were transferred into 11 recipient female (CFLP) mice. All transgenic production techniques were as described previously (Allen et al., 1987). Subsequently 13 progeny survived to weaning, when tail DNA was prepared. Tail DNA from these mice was restricted with PstI, an enzyme that cleaves in the E2 linker at the 5' end of the construct. Southern blot analysis of these digests probed with a fragment of the MS32 locus containing only repeat unit sequence, revealed four of the mice to be carrying transgenic DNA by the presence of one or more bands hybridising to this probe. Three of the four mice gave multiple hybridising DNA fragments indicating that they were either multi-copy insertions or the result of more than one separate insertion into the same embryo. Mouse 102 only showed one hybridising band, suggesting at first that it contained only one single-copy insertion.

Transmission. By breeding these mice with non-transgenic C57BL/6J/CBA/ca F2 mice, we were able to show transmission of the transgenes and to establish six lines containing a single transgene locus. From founder male 110 three separate loci segregated to F1 mice: 110A (multi-copy/tandem insertion), 110C and 110D (single-copy insertions). Founder female 102 gave two single-copy inserts, 102A and 102B, whilst founder male 109 gave progeny with only one multi-copy insertion. Founder 105, a female, died before breeding preventing the demonstration of independent segregation of the two transgenes carried by this animal: 105A (multi-copy) and 105B (single-copy). Figure 4.3 shows a Southern blot analysis of the segregated transgenic lines, with the exception of founder 105, where two digests are shown to indicate the two transgenic insertions. The transmission data from the founder mice, 102, 109 and 110 is shown in table 4.1.

These data show that transmission of six MS32 transgene loci is possible and therefore that the integration is present in germ cells. If the DNA integration was at the one-cell stage then we would expect the transgene to transmit to 50% of offspring when mated with a negative mouse. For 109, 110A and 110D transmission rates are close to 50%. The transmission data from 110C shows a lower rate of 32.6%. Using the Fischer exact test, two tailed, P = 0.068. This is not a significant deviation from the expected transmission, however if this pattern was to be observed with a larger sample size then it is quite likely that this indicate a significant deviation from the expected 50% transmission. The data from founder female mouse 102 is complicated by the small numbers of
Figure 4.3, Southern blot of the segregated lines of MS32 transgenes.
Tail DNA was cleaved with the appropriate restriction endonucleases to give an individual pattern. These were identified during general restriction analysis. Multi-copy line 110A is cleaved with Pst I, line 109 with Hind III and Ssp I. Single-copy 110C with Eco R I, 110D, 102A and 102B with Ava II. The non-segregated 105 DNA contains both the 105A multi-copy and the 105B single-copy insertions. The 105-1 digest is with Eco R I. The 2.3kb band is equivalent to the truncated single-copy locus 105B and the larger band is therefore from the multi-copy locus. The multi-copy locus is therefore likely to contain only two rearranged copies of the original construct. The 105-2 digest is with Ssp I, the lower band again represents single-copy 105B and the larger band multi-copy 105A.
progeny that survived. In the initial breeding stages one litter was eaten by the mother and from the other four litters only 11 out of the 27 mice born survived for testing. From the founder 102 it is not possible to draw any conclusions about the transmission rate, however data from subsequent generations reveals transmission rates closer to 50%. For example, two litters from mouse 133, an F1 102A and 102B positive female, mated with a non-transgenic C57BL/6J/CBA/au F2 mouse, gave transmission rates of; 11/23 (48%) for 102A and 9/23 (40%) for 102B.

Table 4.1 Transmission numbers from the founder mice to the F1 generation for six transgenic lines. Founder mouse 110 transmits three separate transgene integrations 110A, 110C, and 110D; founder mouse 102 transmits two separate insertions 102A and 102B.

<table>
<thead>
<tr>
<th>Transgene locus from founder mouse</th>
<th>no. of F1 progeny tested</th>
<th>no. of positives</th>
<th>% transmission</th>
</tr>
</thead>
<tbody>
<tr>
<td>110A</td>
<td>46</td>
<td>24</td>
<td>52.2</td>
</tr>
<tr>
<td>110C</td>
<td>46</td>
<td>15</td>
<td>32.6</td>
</tr>
<tr>
<td>110D</td>
<td>27*</td>
<td>15</td>
<td>55.5</td>
</tr>
<tr>
<td>109</td>
<td>28</td>
<td>13</td>
<td>46.5</td>
</tr>
<tr>
<td>102A</td>
<td>11</td>
<td>3</td>
<td>27.3</td>
</tr>
<tr>
<td>102B</td>
<td>11</td>
<td>2</td>
<td>18.2</td>
</tr>
</tbody>
</table>

* fewer 110 F1 progeny were tested for 110D transmission since this locus was not initially identified; it is flanked by repetitive mouse DNA (chapter 5) uncleaved by Pst I and therefore a hybridising band corresponding to this locus was not seen with the Pst I digestions used for the first 110 litter analysis.

Linkage. We have no evidence of linkage between the three transgene loci integrated into founder mouse 110; transgenes 110A, 110C and 110D. Pairwise analysis using a computer program written by Jack Jeffreys of the transmission of 36 progeny of the founder mouse 110, showed that for loci 110A and 110C the maximum LOD score at \( \theta = 0.4 \), was \( z = 0.48 \). The same analysis for pairwise comparisons between loci 110A and 110D gave a maximum LOD score at \( \theta = 0.43 \) of \( z = 0.23 \), and between loci 110C and 110D at \( \theta = 0.43 \) of \( z = 0.22 \). Pairwise analysis of 27 progeny from a transgenic mouse carrying 102A and the 102B locus gave a maximum LOD score at \( \theta = 0.41 \) of \( z = 0.2 \). None of these LOD scores indicate significant linkage.

Physical mapping of the MS32 transgenes. The restriction maps of the eight transgenes were elucidated by full and partial (for the multi-copy transgenes and for 102A) restriction mapping with the enzymes shown and are presented shown in 4.4 and 4.5. Initially no MS32 Pst I fragments were detected for the single-copy transgenes 102B and 110D, suggesting a lack of neighbouring Pst I sites. Further analysis failed to detect the expected frequency of sites for most other restriction enzymes tested, both around these two transgenes, and around the single-copy transgene 102A and the multi-copy transgenes 105A and 109. The latter two transgenes can however be detected in a Pst I digest by the release of internal transgene blocks from the multi-copy integrant. All five of these transgenes showed sites for the enzymes Ava II and Mnl I close to the flanking ends of each transgene. Furthermore, partial digestion of Pst I cut 102A with Mnl I showed the presence of a restriction site every ~230bp over most of the 5' flanking DNA. Ava II gave a similar pattern though it did not cleave every ~230bp. These data
Figure 4.4 A, Schematic of the original construct for microinjection. 212bp of 5' and 214bp of 3' flanking MS32 DNA are included in the construct. Oligonucleotide primer sites are depicted by arrows. Restriction sites found in the MS32 flanking DNA are shown; Pst I (P), Eco RI (E), Hind III (H).

B. Structures of the five single-copy transgene inserts. The blocks of shading represent the regions of repeat units of the original construct maintained in each insertion. Thick lines represent human MS32 flanking DNA, thin lines represent mouse flanking DNA. Restriction sites in the flanking DNA are shown; Mbo I (Mb), Mnl I (Mn), Alu I (Al), Rsa I (Rs), Ava II (A), Bcl I (Bc), Bam H I (B), Hinc II (Hc), Dra I (D), Hae III (Ha). These maps are based on work performed primarily by Alec Jeffreys and additionally by myself.
Figure 4.5, Restriction maps of the multi-copy transgenes, 109 and 110A, plus a partial map of the multi-copy transgene 105A, 3' end. These maps are based on work performed by Phillippe Bois, Andrew Collick and Alec Jeffreys (Collick et al., in press). Restriction sites in the mouse flanking DNA are; Ssp I (S), Hind III (H), Pst I (P), Mbo I (Mb), Mnl I (Mn). Constructs maintaining the original number of repeat units, are labelled 2.5 kb long. For 109 and 110A there are expanded and deleted blocks of repeats, and the relative sizes of the MS32 hybridising bands on a Southern blot relating to these blocks are labelled in kb. The 105A 3' region is mapped but no information is available for the 5' end since the exact sizes of both blocks of repeats are unknown (?).
suggested that these five transgenes had integrated into some form of repeated DNA in the mouse genome, possibly mouse gamma satellite which has an array of 234bp repeat units, most of which contain one site for *Mnl I* and often for *Avi II* (Vissel and Choo 1989). In contrast, the single-copy transgenes 105B and 110C and the multi-copy transgene 110A showed a normal frequency of restriction sites in mouse flanking DNA. The single-copy transgene 102A shows loss of the 3' *Eco RI* and *Hind III* sites diagnostic of the synthetic sequence E2, suggesting that the integrant was truncated at the 3' end. Similar analysis showed that 102B was 5' truncated. 105B and 110D appeared intact by analysis of *Eco RI* and *Eco RI* diagnostic sites, but were shorter than predicted from the original construct length, suggesting loss of MS32 repeat units. 110C consisted of two adjacent MS32 repeat unit blocks separated by *Hind III*, *Pst I* and *Eco RI* sites, consistent with a rearranged integrant. The multi-copy transgenes 109 and 110A contain 6 and 11 blocks of repeats respectively, some of which show size changes relative to the original construct as detailed in the restriction maps. A full restriction map is not available for the multi-copy transgene 105A. Since no segregated lines could be established from the dead mother, multi-copy transgene analysis was complicated by the presence of the single-copy 105B insertion. A basic map of the 3' end of the 105A transgene is shown in figure 4.5. It was not considered a high priority to obtain a full map for a transgene integrant for which mutation studies by pedigree analysis were not possible.

**MVR mapping.** The internal structures of the minisatellite arrays of the five single-copy transgenes were analysed using four-state MVR-PCR (Tamaki *et al.*, 1993) which assays two polymorphic sites within MS32 repeat units to enable the interspersion patterns of the resulting four types of repeat unit (E, e, Y, y) to be determined. An example of a four state MVR-PCR from a transgene insertion is shown in Figure 4.6. MVR-PCR was performed using the flanking primer MS32-0, 5' to the array. With those 5' truncated transgenes which lacked MS32-O, reverse four state MVR-PCR was performed using the primer MS32-E in the 3' flanking human DNA (figure 4.1). The most 5' minisatellite array of the multi-copy transgene 110A was also MVR mapped, using a primer in the mouse DNA flanking the insert (chapter 5). MVR codes of the progenitor allele and each mapped transgene are shown in figure 4.7.

The first minisatellite array in the multi-copy transgene 110A carries an MVR code identical to the input MS32 allele. In contrast all five single-copy transgenes show different MVR codes, though all clearly align with the progenitor allele. In some cases null or ε-type repeat units were seen. These are repeat units that do not amplify with the E, e, Y, or y-type repeat specific primers (Tamaki *et al.*, 1992). There are no null repeats in the original allele and their appearance in transgenes 110C, 105B and 110D may be the result of Taq polymerase errors during PCR amplification of the original allele prior to injection (Dunning *et al.*, 1988). Transgene 110D contains full 5' and 3' flanking human DNA, but has lost 10 repeat units, 9 at the beginning of the allele and one in the centre of the array. Likewise 105B has lost 23 repeats from the 3' end of the array and has undergone a complex rearrangement over repeat units 20-24 where a segment of code reading 'yeyye' in the original allele has changed to 'eyye' in the transgene. As predicted by restriction mapping, transgene 102A has lost the 3' flanking human DNA together with the terminal 34 repeat units of the array. Transgene 102B is the apparent complement of 102A with no 5' flanking human DNA and the loss of the first 50 repeat units. Transgene 110C results from a circular permutation of the input allele and consists of the last 43 repeats followed by 5' flanking human DNA, then 5' flanking DNA followed by the first 31 repeats of the original array (figure 4.4). The two MVR maps combine to
Internal sequence variation:

<table>
<thead>
<tr>
<th>polymorphic sites</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'...TGACTCAGAATGAGACAGG</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>3'...ACCGGTCCCCACTGAGTCCT</td>
<td>E type repeat unit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3'...GCCGGTCCCCACTGAGTCCT</td>
<td>e type repeat unit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3'...ACTGGTCCCCACTGAGTCCT</td>
<td>Y type repeat unit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3'...GCTGGTCCCCACTGAGTCCT</td>
<td>y type repeat unit</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Original construct allele, 4 state MVR-PCR pattern

```
left悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠
```

Transgene 105B 4 state MVR-PCR pattern

```
left悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠悠
```

**Figure 4.6 A, Four State MVR PCR.** The 29bp MS32 repeat unit showing the two polymorphic sites, (C/T) and (G/A). In four state MVR-PCR the C/T polymorphism is used in conjunction with the G/A polymorphism, which forms the basis for two state MVR-PCR, to distinguish between four, (E-, e-, Y-, and y-) repeat unit types.

**B.** Autoradiograph showing the 4 state MVR-PCR banding pattern of the allele from the original construct (1) and the allele from transgene 105B DNA (2).

**C.** The original construct 4 state MVR-PCR and 105B 4 state MVR-PCR maps, read from the base of the autoradiograph in (B). The 25 repeat unit deletion in the 105B map is indicated by (-) and the 5 repeat unit 'switch' from the original construct pattern is underlined.
Figure 4.7. MVR-PCR codes of the M332 repeat unit arrays of the transgenic mice. Four state MVR codes (repeat unit types E, e, Y, y) are shown for the progenitor allele, for each single-copy transgene and for the first repeat unit block of the 110A multi-copy transgene. o = "null" repeat unit unamplifiable with E-, e-, Y-, or y- specific MVR primers; ? = repeat type not known. Differences between the transgene and progenitor are marked *. Deleted repeats are indicated by dashes (-).
give a full original construct plus a duplication of three repeat units (Yey, positions 28-31) at the point of integration of the MS32 allele.

**Discussion**

This work demonstrates for the first time that it is possible to create, by microinjection, mice transgenic for human minisatellites. Single-copy integrants were obtained together with novel multi-copy inserts containing tandem repeats of the MS32 array plus flanking DNA. Both multi-copy and single-copy integrants are expected from microinjection experiments (Hogan et al., 1986). A total of eight insertions of the MS32 construct DNA were recovered in four of the thirteen founder progeny. This is a good recovery of integration events when compared to other microinjection experiments (M. Norris, personal communication). This may be a reflection of reduced deleterious effects of a non-coding integrant on transgenic progeny, or due to a relatively extensive analysis of the founder mice and their progeny. With the obvious exception of the 105A and B transgenes for which no breeding from the founder could be performed, transmission of all the other six transgene inserts was achieved. Therefore germline integration of the transgene had been obtained for each insertion event. The transgenes all gave transmission rates that did not deviate significantly from those expected from a non-mosaic founder mouse indicating that the microinjection resulted in integration at the one-cell stage. Founder mice 102 and 110 harboured two or more different insertions. In mouse 110, there were two single-copy events, and one multi-copy integration consisting of a tandem array of constructs, but these integrants were not linked indicating completely separate integration events had taken place.

MVR mapping analysis showed that all five single-copy integrants had undergone rearrangement prior to, during, or post-insertion. Pedigree analysis has shown that the single-copy transgenes 102A, 102B, 110C and 110D are inherited stably, without change in the repeat copy number (see chapter 6). Abnormal structures in the single-copy transgenes are therefore most likely to have arisen either pre-insertion, where the linear constructs are free to recombine with each other and exchange sequence information, or during the insertion process.

Various types of insertional rearrangement are seen in the single-copy transgenes. For example, 105B has lost all of the 3' human flanking DNA except for the proximal 27bp (chapter 5), together with 26 repeat units from within the MS32 array. In addition, repeats 20-24 have switched MVR code which may be analogous to the products of complex gene conversion events seen in the male germline at the endogenous MS32 locus (Jeffreys et al., 1994) and may have arisen by mismatch repair at a heteroduplex formed during recombination of two progenitor molecules.

Another type of rearrangement is seen with transgene 110C which consists of a circular permutation of the original construct containing two blocks of MS32 repeats separated by a tail to head fusion of human flanking sequences. This structure suggests that transgene insertion occurred via a circular recombination intermediate (figure 4.8) to disrupt the MS32 repeat block into two halves plus a 3 repeat duplication at the site of insertion into mouse DNA. This structure strengthens previous evidence of circular DNA molecules being recombinationally active in transgene integration (Brinster et al., 1985, Hamada et al., 1993).
Figure 4.8, A putative insertion mechanism for transgene 110C
It has been proposed (Brinster et al., 1985) that the mechanism for integration of microinjected linear construct is rate limited by the occurrence of randomly generated chromosomal breaks in the DNA, that the ends of the injected DNA molecules initiate integrations at these breaks (to account for the fivefold increase in integration frequency of linear vs. circular DNA), and that homologous recombination occurs between injected molecules (resulting in tandem arrays of multiple single constructs), but not between the injected and chromosomal DNA sequences (possibly due to differences in chromatin structure between the two DNAs). There is some evidence for this as a mechanism. For example, injecting two different MT-human growth hormone genes with non-overlapping deletions into one-cell embryos resulted in some instances of homologous recombination between these genes, some of which produced functional MT-hGH (Palmiter et al., 1985). A similar experiment injected three overlapping non-functional genomic DNA fragments together constituting the whole human serum albumin gene, into one-cell embryos and analysed the resulting mice for expression of the transgene (Pieper et al., 1992). All mice analysed contained recombined fragments and 74% contained a reconstituted functional hSA gene. This data suggests that homologous recombination occurs with a very high frequency in murine zygotes, and provides a simple method for the generation of large transgenes based on genomic rather than cDNA sequences. The authors suggest that the head to tail tandem repeat integration patterns commonly observed when many copies of a single DNA fragment are microinjected into murine zygotes, probably stem from homologous recombination between insert DNAs. However this does not provide evidence of homologous recombination between insert and mouse genome DNA. The potential exists therefore that the multi-copy 110A, 109 and 105A transgenes integrations arose via, in part, homologous recombination between original DNA construct molecules. This type of process may have resulted in the changes in repeat unit block size between constructs in the array; both 109 and 110A have expanded and deleted repeat unit regions. However there is no direct evidence to suggest that this was the sole mechanism of tandem array production.

Conclusions

In this work we have shown that it is possible to create non-mosaic transgenic animals carrying a DNA construct comprised of a long stretch (2049bp) of 29bp repeat units, and to maintain the basic pattern of interspersion of repeat unit types seen in the original construct. There are some changes in the MVR-PCR patterns of the transgene inserts, probably resulting from the integration process, for example truncations of transgenes such as 105B, 102A and 102B, and a three repeat unit duplication of a single construct at the point of insertion, 110C. However the microinjection process is complex and not fully understood. To appreciate exactly the changes in the transgene structure revealed by MVR mapping and restriction analysis it is important to look at the junction sites between the transgene and mouse genome DNA. In chapter five the process developed for the recovery of mouse flanking DNA and subsequent sequence analysis is presented to further elucidate on the MS32 construct DNA integration processes.
Chapter 5

Transgene flanking sequence recovery by vectorette PCR and sequence analysis

Summary
Vectorette PCR permits the specific amplification of unknown DNA segments flanking a known DNA sequence. It enables the application of the PCR where sequence information is only available for one primer site. Vectorette PCR was used for the systematic mapping and retrieval of unknown mouse DNA flanking the MS32 transgenes, to provide information on integration processes. The retrieved DNA was sequenced to allow assessment of the effects of flanking DNA type on MS32 mutation rate and process at the separate transgenic loci. To enable the sequence of large vectorette PCR fragments to be obtained without cloning, the procedure of sub-vectorette production was developed. At least one flanking DNA sequence for each MS32 transgenic locus was recovered and sequence analysis of the mouse DNA flanking these transgenes showed that 5 of the 8 insertions were into mouse gamma satellite repeated sequence DNA. This suggested a non-random integration of the MS32 transgene construct into the mouse genome.

Introduction
In many transgene studies the aim is solely to achieve transgene expression, therefore the structural basis of insertion and the type of DNA flanking the transgene insert may not be important, and for this reason the DNA sequences flanking transgene insertions are rarely reported. In addition many studies analyse only one or two separate insertions, having selected these on the basis of their positive expression or some other required characteristic. Comparison of a small number of flanking sequences may not reveal any significant patterns in transgene integration. However for the MS32 transgenic project there were a number of reasons why the recovery and subsequent sequencing of transgene flanking DNA was important. The sequence data provide information about the processes of integration. Since DNA consisting of tandemly repeated sequence units had not previously been used in mouse zygote microinjection experiments, there were no clues as to how it might behave. Knowledge of the flanking sequence around the transgene inserts enables the study of any targeting of the MS32 constructs in the mouse genome. There was no known sequence homology between the MS32 repeat units and the mouse genome, but it was possible that the MS32 repeat units would integrate more favourably into certain types of mouse sequence. A total of eight separate insertion events were identified and to ensure no bias in the selection of these loci for study, the structural basis and flanking DNA sequence of each locus was analysed. This was a greater number of insertion events from the same starting construct than had been studied in this detail before, enabling the identification of any significant targeting effects.

As well as studying the integration behaviour of tandemly repeated MS32 construct DNA, another aim of creating MS32 transgenics was to investigate their potential for use as a tool to aid in the study of mutation processes at the
MS32 locus. For this reason it is important to know the type of sequence flanking each insert to assess if there are any positional or flanking sequence effects on mutation rates and processes at the different loci.

Mutation analysis for the truncated transgene insertions, for example 102A and 105B, is hampered by the loss of the MS32 flanking DNA primer sites which were incorporated into the design of the original construct before microinjection for subsequent mutation analysis of transgenic loci. Knowledge of the flanking sequence adjacent to the truncated regions of these insertions, facilitated the design of novel mouse sequence based primers for the mutation analyses. These primers could also be used to study the effects on the host DNA of construct integration, for example deletions, duplications or other rearrangements, by the use of these primers on genomic DNA from a non-transgenic mouse.

Flanking sequence recovery by PCR. Although transgene flanking DNAs have not been routinely analysed previously, some loci, in particular those which instigate rearrangements of the host DNA have been studied. In most cases the DNA is recovered by screening genomic DNA libraries with an insert probe (Woychik et al, 1985; Butner and Lo 1986). This is a time consuming process, and for the MS32 transgenes was not appropriate because eight different insertion events were studied. More recently people have used the PCR (Saiki et al, 1988) to recover fragments containing their transgene insertions from mouse genomic DNA (e.g. MacGregor and Overbeek 1990, Hamada et al, 1993), before sub-cloning these PCR fragments prior to sequence and structural analysis. In general the PCR is used for the specific amplification of DNA fragments between two regions of known sequence to which primers can be directed. However, sometimes the retrieval and subsequent sequence analysis of unknown regions of DNA is required. A number of different techniques have been developed to apply PCR to isolate unknown sequence provided that it is flanking a region of known sequence and that a primer can be designed in either orientation from the known sequence. Such techniques as inverse PCR, Panhandle PCR, Capture PCR and Vectorette PCR have been described as PCR methods for genome walking.

With the inverse PCR technique (Triglia et al., 1988) the DNA source is digested with a suitable restriction enzyme to give a fragment of amplifiable length. This fragment is then circularised by ligating the restricted ends. Thus a known DNA sequence can be made to surround sequences normally located on both sides of this DNA fragment. Two primers extending in opposite directions from the region of known sequence in this fragment are used in a PCR to amplify across 5' and 3' unknown sequences flanking the known sequence region, around the circularised molecule.

The Panhandle PCR technique (Jones and Whaisforter 1992) uses a DNA source (e.g. genomic DNA) digested with an enzyme which leaves a 5' overhang. This genomic DNA is treated with calf intestinal alkaline phosphatase and the ends are ligated to a 5' phosphorylated single stranded oligonucleotide, so that the 3' end of each strand of digested genomic DNA is modified. The protruding 3' ends are complementary to a region of known DNA sequence that is upstream to the unknown region of interest. Therefore during PCR performed in dilute conditions that promote intra-strand annealing, the single strands of genomic DNA that contain the complement to the modified 3' ends will form a stem-loop structure with a recessed 3' end. Therefore, the ligated oligonucleotide can prime template-directed DNA polymerisation on those genomic strands that contain the known sequence, resulting in known DNA being attached to the uncharacterised end of the unknown DNA. PCR can then be used to amplify the unknown DNA.
Capture PCR (Lagerstrom et al., 1991) is an additional method that can be used to retrieve unknown DNA flanking a known region. This method uses a biotinylated specific primer to amplify from a pool of DNA molecules previously modified by the addition of oligonucleotide linkers to all ends. Extension products are captured by binding to streptavidin to provide a pool enriched for the sequences of interest for amplification. These sequences are then reamplified with the oligonucleotide linker primer and a nested primer hybridising just downstream of the biotinylated primer.

For the MS32 transgenics project it was decided to further develop a PGR based approach to the recovery of DNA first reported for the recovery of sequences from YAC insert junctions (Riley et al., 1990) and subsequently referred to as vectorette PCR. Vectorette technology has also been used in the determination of exon structures and in the recovery of unknown bacterial sequences (Roberts et al., 1992, Arnold and Hodgson 1991). The vectorette technique was chosen in preference to the panhandle or capture PCR methods since the vectorette linkers and primers can be used universally (unlike pan-handle PCR) by modification of the 5' end of the vectorette 'bottom' strand. Vectorette PCR had also been reported as giving good specificity of recovery from complex fragment mixtures (Riley et al., 1990, A. MacLeod personal communication).

A 'hot start' PCR was performed when amplifying directly from the vectorette libraries, to ensure complete denaturation of the DNA before the addition of primers, buffer and Taq polymerase. In the absence of a 'hot start', unligated top strand vectorette oligonucleotide could, in principle, prematurely dissociate from the end of any genomic DNA fragment, generating a 5' single stranded vectorette extension which could be filled in by Taq polymerase to create a sequence complementary to the vectorette primer. Such fill-in would abolish the locus specificity of vectorette PCR. Hot start obviates the need for 5' phosphorylation of the top strand of the vectorette oligonucleotide.
Figure 5.1 A. Schematic of the vectorette linker. The sequences of the 'top' and 'bottom' oligonucleotides used to make the vectorette linker are detailed in figure 2.1.

B. Schematic of a vectorette library. The vectorette linker was ligated onto digested DNA from a whole genome or a size fractionated DNA source, to make a 'vectorette library' of fragments with a linker attached to both ends. Dilutions of this library were used in subsequent PCR amplifications. Only those fragments in the library which have the recognition sites for the fragment specific primer should amplify, providing a mechanism for retrieving single fragments from a complex mixture.
Figure 5.2, Schematic of basic Vectorette PCR strategy. For simplicity only one end of a fragment from the vectorette library is shown, with a ligated Sau 3 A 1 vectorette linker. Single lines represent the vectorette linker and boxes represent the Mbo I cut genomic DNA.
stranded sequencing methods. Most double stranded PCR sequencing methods will only give 200-300bp from each end of a fragment. Therefore to access full sequence information for the original vectorette fragment, sub-vectorettering procedures were developed to provide a series of overlapping DNA fragments for sequencing. Sub-cloning of two of the 110C vectorette products was necessary to obtain sequence information. Sequences obtained were analysed for sequence homologies by comparison with other known sequences in the EMBL database.

Results

Recovery of flanking mouse DNA by vectorette PCR. Standard restriction digestion and Southern blot analysis was performed on mice positive for MS32 transgene insertion (chapter 4). This identified appropriate restriction sites in the transgene flanking DNA for use in vectorette PCR to recover mouse flanking DNA sequences.

Transgene 110D. Restriction mapping showed this to be a single-copy insertion of the transgene construct with the loss of 10 repeat units, but with full 5' and 3' MS32 flanking sequences. There were a lack of common four and six cutter restriction sites in the flanking mouse DNA, but the enzyme Ava II cleaved frequently and was utilised in the preparation of a vectorette library. A size fraction of Ava II digested 110D positive mouse genomic DNA, known to contain the transgene fragment, was recovered by agarose gel electrophoresis and electrolelution and used in the preparation of the vectorette library. The size fractionation was performed to reduce the complexity of the DNA for use in the preparation of the vectorette library. After electrolelution, an aliquot of the recovered DNA was electrophoresed alongside some Ava II digested 110D mouse DNA to assay the concentration of the recovered DNA (figure 5.3 A). This DNA was used to prepare the vectorette library. Since Ava II cleaves a degenerate G'CJ^/tCC site, a 50% mixture of GAC and GTC 5' overhanging Vec1 top sequences were used in the preparation of the vectorette linker for this library. The first amplifications from this library were across the MS32 array and products were verified by diagnostic restriction analysis, Southern blotting and hybridisation with an MS32 repeat unit probe (figure 5.3 B and C). In addition to the transgene specific primer to vectorette primer product being isolated, the initial vectorette PCR will contain some non-specific products. The majority of the non-specific products will be vectorette primer to vectorette primer products. Therefore before re-amplification of the product for sequencing, the size fraction containing the correct product was recovered by electrolelution then amplified with a nested MS32 primer. This hemi-nesting strategy was utilised to aid specific recovery of transgene specific to vectorette products by making these products smaller than the vectorette to vectorette products. Smaller products out-amplified the larger non-specific products in the hemi-nested PCR. The products were amplified to a concentration level detectable by ethidium bromide staining on an agarose gel (figure 5.3 D), then electroleluted and ethanol precipitated for sequencing.

Transgene 102B - Overlapping vectorette products from a partial digest library. Restriction mapping showed this single-copy transgene to have intact 5' MS32 human flanking DNA, but no 5' flanking sequences. The mouse flanking DNA sequences again showed an unusual lack of most restriction sites other than for Ava II. Ava II was again chosen to construct a vectorette library using partially digested 102B genomic DNA. Vectorette PCR amplifications were used as a mapping tool to generate products extending out to a number of Ava II sites in the 5' and 3' flanking DNA (figure 5.4 A). Amplifications with the 32E and Vec1 primers gave four bands, representing products extending to four flanking Ava II sites. E1 and Vec1 primer amplification only gave products extending to the first three of these Ava II sites, presumably due to the favoured amplification of shorter products.
Figure 5.3. The steps in the strategy for the recovery of 5' and 3' mouse DNA flanking transgene 110D. A. Autoradiograph of a Southern blot of an aliquot of a recovered size selected fraction of 110D positive Ava II cut genomic DNA and assessment of the concentration of the recovered DNA in comparison with non-electroeluted digested DNA. The filter was hybridised in a phosphate/SDS hybridisation mix with 10ng of a 5kb long MS32 repeat unit probe labelled by ransom hexamer priming, at 65° overnight. In the non-electroeluted sample 4µg of total genomic DNA was electrophoresed. In comparison with this, the electroeluted fraction appears to represent the amount of transgene signal expected from a 1µg initial input of total genomic DNA. 2µl of the electroeluted sample were used in the preparation of an Ava II vectorette library.

B. Schematic of the 110D transgene locus and flanking DNA. The restriction sites shown are Ava II (A), Eco R I (E), Hind III (H) and Pst I (P). The primer sites utilised are Vec1, 320R, 32O, 32E and 32ER, and the extent of the vectorette products recovered is shown.

C. Autoradiograph of a Southern blot showing the diagnostic restriction sites in the initial vectorette products recovered, Vec1/32E is cut with Pst I and 32O/Vec1 with Hind III. MS32 hybridisation was as described in 'A'.

D. Agarose gel stained with ethidium bromide showing the nested PCR products 32ER/Vec1 and 32OR/Vec1. These products were electroeluted and used for sequencing.
Figure 5.4. The structure of the 102B transgene and associated vectorette PCR products.

A. E = Eco RI, P = Pst I, M = Mnl I, A = Ava II. Ava II vectorette linkers were attached to DNA partially digested with Ava II, to produce vectorette PCR products extending to Ava II sites in both the 5' and 3' mouse flanking DNA. (a, b, c represent the Vec/E1 products detailed in (B)) 32-TAG-A is an MS32 repeat unit specific primer.

B. Vectorette products extending from the 5' junction of the transgene. Using initial primer combinations Vec/32E and Vec/E1 there are three products extending out to three Ava II sites in the flanking DNA. There is also a product extending to a fourth Ava II site detectable in the Vec/32E amplification (not detailed in A). The filter was hybridised in a phosphate/SDS hybridisation mix with 10ng of a 5kb long MS32 repeat unit probe labelled by random hexamer priming, at 65° overnight.

C. The MVR-PCR pattern visible by ethidium bromide staining of an agarose gel, when the electroeluted band from the Vec/32E reaction was amplified with Vec/32-TAG-A (triplicate reactions). Duplicate amplifications on the same electroeluted product using the 32-TAG-T repeat unit specific primer were also performed to generate the corresponding set of MVR-PCR products extending to each t-type repeat unit. ΦX174 cleaved with Hae III. An arrow indicates the smallest product from the Vec/32-TAG-A reaction which was electroeluted and used in sequencing reactions.

D. Vectorette products extending 3' of the transgene insertion. A Southern blot of amplification products with Vec/32ER was probed with an MS32 3' flanking probe (141bp of the 3' flanking DNA). Three vectorette products were amplified extending to Ava II sites in the 3' flanking mouse DNA. The smallest product is cleaved by Mnl I which cuts in the MS32 flanking human DNA. MS32 hybridisation was as described in (B).
As there was no 5' human MS32 flanking DNA, a hemi-nested strategy for size reduction of the initial 5' vectorette PCR products (Vecl/E1 and Vecl/32E, figure 5.4 B) was developed using MVR-PCR. For this reaction the Vec1 primer was used as the flanking primer, and the repeat unit specific primers 32-TAG-A or 32-TAG-T as the transgene primers to produce a series of products extending from the closest Ava II site in the 5' flanking mouse DNA to each a- or t-type repeat unit in the transgene MS32 array (figure 5.4 C). The first repeat unit in the array is an a- type repeat and the appropriate band from the MVR-PCR (indicated by an arrow in figure 4C) was electroeluted and amplified to a level detectable by ethidium bromide staining on an agarose gel before electrolization and sequencing.

The 3' flanking mouse DNA was recovered by vectorette PCR using primer 32ER in the human 3' flanking DNA (figure 5.4 D). Junction products were verified by Southern blot hybridisation with a human 3' flanking sequence probe and digestion with Mnl I which has a diagnostic site in the 3' flanking DNA. A hemi-nested strategy using the reverse primer of 32A (table 2.1) with Vec1 was used to recover this fragment for sequencing.

Transgene IlOA- 5' flanking sequences and sub-vectorette production. Restriction mapping identified this transgene as a complex multi-copy array of the original construct. The enzyme Mbo I does not cleave in the original construct but cleaves in the mouse DNA 1.3kb upstream of the 5' end of the multi-copy array. A complete Mbo I digest was performed on the genomic DNA and Mbo I vectorettes ligated. The initial amplification using primer 32OR directed towards the 5' flanking mouse DNA produced a 1.34kb product detectable by ethidium bromide staining following agarose gel electrophoresis (A.Jeffreys). This initial vectorette product was too long for full double stranded PCR sequencing. The technique of 'sub-vectoretting' was developed to enable a series of smaller fragments to be produced for sequencing from the original vectorette product. The sub-vectorette procedure utilises restriction sites within the initial vectorette product as positions for ligation of a novel vectorette linker. For all the sub-vectorette procedures used in the transgenic mouse product, blunt ended restriction sites in the initial vectorette products were used, enabling the ligation of a 'universal' blunt end compatible vectorette linker, Vec 3.

Restriction mapping of the initial 1.34kb IlOA vectorette product revealed a number of blunt end restriction sites to which the universal blunt end novel sequence vectorette (Vec3) could be ligated. The 1.34kb DNA fragment was either fully digested with Alu I or Hinc II or partially digested with Rsa I ( since there were two Rsa I sites to access with Vec3) then phenol extracted and ethanol precipitated, before the novel Vec3 was ligated. These sub-vectorette libraries were re-amplified using 32OR and Vec3. The products retrieved are shown in figure 5.6 A and B. The products were amplified to a level visible by ethidium bromide staining on an agarose gel, electroeluted and ethanol precipitated before sequencing. Sequence analysis of the original vectorette products and the four sub-vectorette products produced a complete sequence of the 5' mouse flanking DNA.

Vectorette products for the other five transgenes. Vectorette libraries were constructed from transgenic mouse DNA digested with Mbo I, Bam HI, Bgl II or Ava II. 5' junctions were isolated using the vectorette primer and primer MS32-OR. For 5' truncated transgenes the 5' junction was amplified using primer MS32-E to yield products extending from the 3' MS32 human flanking DNA across the MS32 repeat array and into the 5' mouse flanking DNA. 3' junctions were likewise recovered using primer MS32-ER or, if necessary, MS32-O. All junctions were successfully recovered except for the 3' end of the multi-copy transgenes 110A and 109 and the 5' end of the multi-
Figure 5.5 A. Schematic of the 5' flanking region of transgene 110A. The initial Vec1/32OR vectorette product has been sub-vectoretted to give the products shown.

B. Amplification from the subvectorette libraries made from Rsa I, Alu I, Hinc II, digested initial vectorette. The products were visualised by agarose gel electrophoresis and staining with ethidium bromide.
copy transgene 105A. The regions of the transgene flanking DNA covered by vectorette products recovered by A. Jeffreys are shown in figures 5.6 (single-copy) and 5.7 (multi-copy). Complete sequence data was not obtained for the full length of all these vectorette products, since some products proved refractory to Winship double stranded sequencing; however at least one flanking DNA sequence was obtained for each transgene. Both 5' and 3' sequence was obtained for each of the single-copy transgenes. However, isolation of 5' and 3' vectorette products from the multi-copy transgenes proved more difficult, due to the more complex and less well characterised structures of these loci. The sub-vectorette procedure using the blunt end compatible vectorette 3 linker was performed for all of the transgene loci in figures 5.6 and 5.7. For the 102A locus the change in the unusual restriction pattern 3' of the insertion allowed the recovery of a large flanking vectorette product of 3.4kb, and sub-vectoretting of this fragment enabled the recovery of sequence information 3kb downstream of the 102A insertion.

Sub-cloning of transgene 110C vectorette PCR products. The 5' transgene 110C vectorette product, Vec3/320R (figure 5.6) and the 320/Vec3 (figure 5.6) 3' 110C vectorette products proved to be refractory to sequencing by either the Winship method or by Taq cycle sequencing. Therefore these products were subcloned into the pBluescript vector (Stratagene), and single stranded clones produced for sequencing. The strategy and results for 5' and 3' sub-cloning are presented in figures 5.8 and 5.9. For the 3' flanking sequence cloning strategy, the aim was to recover two overlapping sub-clones, one from the Hae III digested DNA which would have sequence from the junction between the last MS32 repeat unit and the flanking mouse DNA and an overlapping 3' clone. However the sub-cloning of the Hae III generated fragments (figure 5.9 A) was unsuccessful, although three positive clones were obtained for the downstream Hind III/Alu I DNA fragment (figure 5.9 B). This meant that the 3' junction between the truncated MS32 repeat units and the flanking mouse DNA was not obtained. The cloning of the 5' flanking fragment, including the transgene-mouse DNA junction, was successful.

Sequence homology of the flanking sequences. The sequences recovered from the transgene vectorette products were scanned against the EMBL database. The results of this scan are shown in table 5.1, and the sequence organisation around the transgenes is summarised in figure 5.10; where 5' and 3' human flanking sequences are shown they are complete, with the exception of transgene 109 which has lost 33bp of 5' flanking DNA at the mouse-human junction and 105B which has only the first 27bp of 3' flanking human DNA sequence. Sequencing confirmed that transgenes 102A, 102B, 105A, 109 and 110D had integrated into mouse gamma satellite DNA. The consensus gamma satellite 234bp sequence from the MS32 transgene flanking DNA is compared to the published gamma satellite consensus sequence (Horz and Altenburger 1981) in figure 5.11 and to the most closely related monomeric gamma satellite clone in the EMBL database (Vissel and Choo 1989) and shows a 98.3% identity with these sequences. This sequence comparison backed up the initial observation from vectorette and restriction enzyme mapping of the mouse flanking DNA where for some loci, a digestion periodicity of ~230bp was observed for the enzymes Msp I and Ava II, known to cut once in the 234bp gamma satellite repeat unit.

The 3' and 5' truncation of 102A and 102B respectively was confirmed; in both cases the junction consisted of MS32 repeats joined directly to gamma satellite DNA. The gamma satellite DNA 3' to 102A switches after 3.45 kb to a region showing 92% sequence identity to an LTR sequence from a mouse IAP element (figure 5.12; Aota et al., 1987). Restriction sites in this LTR are present in 102A genomic DNA, indicating that this junction fragment is authentic and not a vectorette PCR artefact. Mouse DNA sequences around 105B, 110C and 110A show no similarity with any known mouse sequence, except for an 80.5% sequence identity with the 3' end of a mouse B2 chapter 5, page 6
Figure 5.6. **Vectorette products isolated from the single-copy transgene flanking DNA regions.** Thick line vectorette products represent the primary vectorette products from the vectorette libraries. Thin line vectorette products with one primer as Vec3 represent the results of sub-vectoretting the larger primary vectorette products. * indicates sub-vectorette products where the MS32 repeat units have been "shrunk" by PCR over-amplification and the Hae III site of the first or second repeat unit in the array is used to ligate Vectorette 1 or 3 linkers. The restriction sites shown are the ones utilised in the vectorette strategy; *Mbo* I (Mb), *Mnl* I (Mn), *Alu* I (Al), *Bcl* I (Bc), *Bam* H1 (B), *Rsa* I (Rs), *Dra* I (D), *Hinc* II (Hc), *Hae* III (Ha). The primer sites shown are those used in vectorette production; Vec1 (vectorette linker 1 primer), Vec3 (universal sub-vectorette primer, vectorette linker 3), 32OR, 32O, 32E, 32ER. The repeat unit blocks are shaded to give an indication of the regions of the original construct MS32 repeat unit array maintained in each transgene insertion.
Figure 5.7, The vectorette products isolated from the multi-copy transgenes 109 and 105A. Thick line vectorette products represent the primary vectorette fragments isolated from the vectorette library. Thin line vectorette products represent sub-vectorette products. The restriction sites used in the vectorette production are; Mbo I (Mb), Mnl I (Mn), and Hae III (Ha). Other restriction sites shown are Pst I (P), Eco R I (E), and Hind III (H). The primer sites used for the vectorette isolation are Vec1 (vectorette 1 linker) Vec3 (vectorette 3 linker), Tag-C (MS32 repeat unit specific primer) and 32O.
Figure 5.8, Strategy for the sub-cloning of the 5' 110C vectorette product. The restriction sites used in the production of the vectorette fragment and in the subsequent sub-cloning into pBluescript vector (Stratagene - details in chapter 2), are; *Mbo* I (Mb), *Dra* I (D), *Hind* III (H), *Eco* R I (E). The primer sites used are Vec3 which contains a *Hind* III site and the primer 32E with an attached 10bp linker at the 5' end of the primer containing an *Eco* R I site. The XL1Blue competent cells and single stranded DNA are prepared as detailed in chapter 2. *The fragment for sub-cloning is reduced to 300bp in size by collapsing the MS32 repeat units by PCR over-amplification, and using a fragment with only 1-2 MS32 repeat units for sub-cloning.*
Cut the fragment with Ha III resulting in a 350bp product and many smaller products

- Cut the fragment with Hind III and Alu I and electroelute the 300bp fragment.
- Ligate into pBluescript SK+ cut with Hind III and Sma I
- Electroporate into competent XL1Blue cells
- Good transformation efficiency, 42 white colonies. Pick white colonies into a microtitre plate and make replicate plate and replica hybond filter
- Hybridise the hybond filter with a 3' flanking DNA probe, made from MS32 repeat unit free vectorette PCR products corresponding to the appropriate region (see fig.5.6.) 15 colonies hybridised positively with the probe (see right). DNA was prepared from five of these, and three were shown to have the corestly sized insert (300bp).
- Single stranded DNA was prepared from these three clones and they were sequenced using the Sequenase protocol

**Figure 5.9 Strategy for the sub-cloning of the 3' 110C vectorette product.** The restriction sites used in the production of the 3' vectorette and subsequent sub-cloning are; *Mbo I* (Mb), *Alu I* (Al), *Hae III* (Ha), and *Rsa I* (Rs), and the primer sites used are 320 and Vec3 (which contains a *Hind III* (H) site. The cloning, electroporation and sequencing protocols are detailed in chapter two. The strategy planned to have two overlapping 3' clones to cover around 450bp of sequence including the junction site between mouse and transgene sequence, however sub-cloning of the *Hae III* fraction, which covered the junction sequence was unsuccessful even when repeated.
<table>
<thead>
<tr>
<th>Flanking DNA name</th>
<th>Number of bp sequenced</th>
<th>EMBL accession number</th>
<th>Junction between mouse DNA and transgene DNA</th>
<th>Sequence homology (gamma satellite or single-copy)</th>
<th>Additional homologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>102A-5</td>
<td>400</td>
<td>X78000</td>
<td>yes</td>
<td>gamma satellite</td>
<td>Mouse IAP element LTR</td>
</tr>
<tr>
<td>102A-3</td>
<td>176</td>
<td>X78005</td>
<td>yes</td>
<td>gamma satellite</td>
<td>Mouse B element LTR</td>
</tr>
<tr>
<td>102B-5</td>
<td>704</td>
<td>X78005</td>
<td>no</td>
<td>gamma satellite</td>
<td>Mouse B element LTR</td>
</tr>
<tr>
<td>102C-5</td>
<td>157</td>
<td>X78007</td>
<td>yes</td>
<td>gamma satellite</td>
<td>Mouse B element LTR</td>
</tr>
<tr>
<td>105B-5</td>
<td>518</td>
<td>X78002</td>
<td>yes</td>
<td>gamma satellite</td>
<td>Mouse B element LTR</td>
</tr>
<tr>
<td>105B-3</td>
<td>449</td>
<td>X78092</td>
<td>yes</td>
<td>gamma satellite</td>
<td>Mouse B element LTR</td>
</tr>
<tr>
<td>10A-5</td>
<td>210</td>
<td>X78001</td>
<td>yes</td>
<td>gamma satellite</td>
<td>Mouse B element LTR</td>
</tr>
<tr>
<td>10A-3</td>
<td>364</td>
<td>X78001</td>
<td>no</td>
<td>gamma satellite</td>
<td>Mouse B element LTR</td>
</tr>
<tr>
<td>10C-5</td>
<td>620</td>
<td>X78009</td>
<td>no</td>
<td>gamma satellite</td>
<td>Mouse B element LTR</td>
</tr>
<tr>
<td>10C-3</td>
<td>537</td>
<td>X78009</td>
<td>no</td>
<td>gamma satellite</td>
<td>Mouse B element LTR</td>
</tr>
<tr>
<td>1100-5</td>
<td>255</td>
<td>X78004</td>
<td>yes</td>
<td>gamma satellite</td>
<td>Mouse B element LTR</td>
</tr>
<tr>
<td>1100-3</td>
<td>242</td>
<td>X78004</td>
<td>yes</td>
<td>gamma satellite</td>
<td>Mouse B element LTR</td>
</tr>
</tbody>
</table>

* The values for total numbers of bp sequenced include some human transgenic flanking DNA and repeats.
Figure 5.10. Schematic of the flanking sequence arrangement around the eight MS32 transgene loci. A. Schematic of the original construct. Hind III (H), Pst I (P), Eco RI (E). The repeat unit blocks are shaded to give an indication of the regions of the original construct MS32 repeat unit array maintained in each transgene insertion. B. Schematics of the structure of the five single-copy transgenes. Flanking sequence homologies, either single-copy or gamma satellite are indicated and by with arrows indicating the orientation of specific mouse sequences; IAP = Mouse Intracisternal A particle, 5’ LTR, B2 = Mouse B2 element. C. Schematics of the 5’ flanking sequence homologies of 110A and 109 and the 3’ flanking sequence homology of transgenic locus 105A. Each gamma satellite integration is in the same satellite repeat unit orientation except for the 109 5’ flanking sequence.
Figure 5.11, Sequence comparison of mouse gamma satellite clones; 'Consensus' is the original consensus sequence from Horz and Altenburger (1981), '10B' is the sequence of the monomeric gamma satellite clone from Vissel and Choo (1989) which is closest in homology to the consensus transgene flanking gamma satellite DNA sequence, 'TG consensus'. Only nucleotides that differ from the original consensus sequence are shown, all other sequenced nucleotides are represented by a dot. The length of the consensus gamma satellite repeat is 234 bp. Gaps (-) are included in the sequences to improve alignment with the transgene consensus sequence.
<table>
<thead>
<tr>
<th>Chapter 5, figure 5.12</th>
</tr>
</thead>
</table>

**Figure 5.12, Sequence comparison between the IAP consensus (IAP cons- black type) sequence (Aota et al., 1987) and the 102A downstream 3' flanking sequence (102A- blue type).** The gamma satellite/IAP junction is shown in the 102A sequence (gamma satellite sequence is shown as uppercase blue bold). The first 352bp of the IAP sequence represent the full sequence of the 5' LTR. The sequence of the 5' LTR at the junction with the gamma satellite is complete. The remaining 108bp of IAP consensus sequence (black bold type) corresponds to the gag gene of the IAP element. There are two separate regions of sequence from the 102A 3' vector expression that show sequence identity with the IAP element (the region of 102 sequence between these two sequences was not obtained). One of these regions shows identity with the 5' LTR from positions 1-76, and the downstream region of 102A 3' sequence shows identity with positions 273-447 of the IAP element, including the terminal IAP 5'LTR sequence and the start of the gag gene. The GATC at the end of the 102 sequence (positions 444-447) is the Mbo site used to make the primary vector library. Variant positions between the IAP consensus and the 102 3' flanking sequence are shown as a *, homologous regions are indicated by (III).
element (Krayev et al., 1981), upstream of 105B (figure 5.13). Sequence analysis of 110C also confirmed the circularly permuted structure of this transgene, which consists of MS32 repeats followed by a complete 3' flanking human DNA region joined to a complete 5' flanking human region followed by MS32 repeats.

Short regions of homology at transgene/gamma satellite junctions. Figure 5.14 shows the sequence of the 3' flanking mouse gamma satellite DNA adjacent to two of the transgene loci, 102A and 105A. Both gamma satellite integrations show 4-5 nucleotides of sequence shared between the minisatellite sequence and the gamma satellite sequence at the integration site. Small regions of homology have been observed in previous studies of non-homologous recombination involving transgenes (Rohan et al., 1990; Hamada et al., 1993). The position of insertion into the 234bp gamma satellite is different in both cases and in each of the other sequenced gamma satellite/transgene junctions (figure 5.15). The orientation of the gamma satellite with respect to the 5' end of the transgene is the same for 4 out of 5 loci, with the exception of the 5' sequence of the 109 locus.

Deletions of endogenous DNA at the integration site. Deletions of mouse DNA at sites of integration have been observed at other transgenes (Woychik et al., 1985; Kato et al., 1986) and we see evidence for this type of structural change around the 110D, 105B and 102A transgene junctions. The 5' mouse flanking DNA of the 110D transgene shows homology with gamma satellite DNA, but the 3' flanking DNA does not. This may be the result of a deletion resulting in the joining of the gamma satellite to flanking non-repetitive DNA. A deletion of mouse DNA at the junction site has also been observed with the transgene 105B, where PCR primers in the mouse DNA are able to amplify across the transgene efficiently, but cannot produce a product in mice negative for insertion (Figure 5.16A and B). This would suggest a deletion of mouse DNA too large to be amplified by PCR (8kb or more). With the insertions flanked both 5' and 3' by mouse gamma satellite DNA, for example 102A, there is also evidence of rearrangement during insertion since the gamma satellite repeat structure does not continue in phase after the transgene integration (figure 5.16 B and C). The mechanism of transgene insertion for 102A into gamma satellite may involve the loss of a number of full gamma satellite 234bp repeats plus 205bp of one repeat leaving a 29bp portion, or it may involve a duplication of 29bp flanking the site of insertion.

Discussion

In this chapter vectorette PCR has been demonstrated as a successful method of systematically isolating mouse DNA adjacent to transgene insertions as a prelude to DNA sequence analysis. In this study flanking DNA has been specifically isolated from a total genomic DNA mixture for 5' and 3' regions of the single-copy transgenes and either 5' or 3' for the multi-copy insertions, without the need for cloning via genomic DNA libraries.

The ability to recover specific vectorette PCR products depends on the complexity of genomic DNA, the frequency of restriction sites in the DNA used to prepare the vectorette libraries and the specificity of vectorette and transgene primers. The latter was optimised by designing primers with a relatively high GC content (54-58%), which permitted the use of a high annealing temperature (67°C) during PCR. In some cases, specificity was sufficiently high to enable the correct vectorette PCR products to be identified after the first round of PCR by agarose gel electrophoresis and staining with ethidium bromide (e.g. the 105B 5' Vec1/32OR product from a Bam H I vectorette library, Figure 5.6 and A.Jeffreys, personal communication.) In other cases, particularly if the vectorette library was constructed from genomic DNA cleaved with frequently cutting restriction enzymes (e.g. Mbo I).
Figure 5.13, Sequence homology between the consensus mouse B2 element sequence (Krayev et al., 1981) and the 5' 105A flanking sequence. The 105B sequence is in a 3'-5' direction upstream of the transgene insertion, therefore the 5'-3' B2 consensus sequence is in the opposite orientation to the 5'-3' transgene insertion. (| | |) indicates regions of sequence identity, * = mismatches. Gaps (-) are included to improve alignment.
Figure 5.14, DNA sequences at the junctions between mouse DNA and MS32 transgenes. A + B, Examples of homologies between MS32 repeat unit sequence (lowercase) and mouse gamma satellite sequence (uppercase) at transgene junctions involving fusion of minisatellite and satellite repeats.
Figure 5.15 Schematic of the positions of insertion of six of the MS32 transgene junctions into the 234bp mouse gamma satellite repeat unit. Junctions are classified as 5' or 3' and regions of shared sequence between the gamma satellite repeat and MS32 transgene are indicated. All the transgene-gamma satellite junctions shown have inserted into gamma satellite repeats in the same orientation, with the exception of the 109 locus, as indicated by dashed line arrows. The 102B 3' flanking sequence has also been shown to be gamma satellite, but the exact junction between the transgene flanking DNA and the mouse flanking DNA has not been elucidated.
Figure 5.16 A, Schematic of the 105B MS32 transgene locus and flanking DNA. The primer sites 105B-1 and 105B-2 in the mouse flanking DNA are indicated by arrows (not to scale).

B. The autoradiograph is of a Southern blot of products from a non-transgenic mouse and from founder mouse 105. Amplifications are with flanking primers 105-B1 and 105-B2 (table 2.1 for sequences) in a 10μl reaction for 20 and 25 cycles from 30ng starting genomic DNA. The Southern blot was probed with 10ng of a 265bp 5' mouse flanking DNA probes prepared from the Vec1/32OR 105B 5' flanking product (figure 5.6) in a phosphate/SDS hybridisation mixture and washed at high stringency. No products are visible in the lanes corresponding to the non-transgenic DNA input.

C. Disruption of gamma satellite repeat unit phase by transgene 102A insertion. Gamma satellite repeats are represented by arrows. The white tip of the arrow represents 29bp of 'duplicated?' gamma satellite repeat.
Chapter 5, figure 5.16
relatively complex sets of PCR products were frequently obtained. Two approaches were used to circumvent this problem. First, knowledge of the size of the restriction fragment containing the transgene enabled restricted genomic DNA to be size fractionated by agarose gel electrophoresis, followed by the recovery of a size fraction containing the transgene fragment. This type of fractionation was performed in the recovery of transgene 110D flanking DNA (figure 5.3 A), to reduce the complexity of the vectorette library 10-100 fold and substantially improve the specificity of the vectorette PCR. The second approach is to perform a first round of vectorette PCR amplification, then to detect products by Southern blot hybridisation, recover by gel fractionation and reamplify using a second nested transgene primer as performed for transgene 102B. This hemi-nesting strategy can also simplify the profile of vectorette PCR products. The technique was successful in isolating large fragments, for example the junction fragment 3' of the 102A locus which is 3.7kb in length.

The use of partially digested genomic DNA in vectorette libraries also provides a general method of mapping restriction sites adjacent to a transgene, as shown for transgene 102B; mouse gamma satellite has a 234bp repeating sequence and Ava II cleaves at a variant position in some gamma satellite repeat units producing a 234bp or 468bp periodicity of digestion. This periodicity is seen in the flanking Ava II restriction map determined from the 102B (and HOD) vectorette PCR products, suggesting strongly that the mouse DNA flanking these insertions was of gamma satellite type. This was confirmed by subsequent sequencing of these products. Such mapping data can not only provide clues as to the nature of the mouse flanking DNA, but can also generate primary vectorette products extending for any required distance into the flanking DNA, without the need for detailed restriction mapping prior to vectorette PCR. Similarly, vectorette PCR can be used to explore the structure of the transgene itself. For example, it is possible to perform MVR-PCR directly on primary vectorette libraries (data not shown), to create PCR products extending from a flanking restriction site to each a- or t- type repeat unit within the minisatellite array. This makes it possible to determine transgene structures in the absence of suitable flanking human DNA, and gives direct information on the distance between the flanking mouse restriction site and the start of the MS32 repeat array.

Vectorette PCR can also be used as a tool to aid sequence analysis of large PCR products by sub-vectoretting. The general strategy used in this project was to partially digest large PCR products with frequently cutting restriction enzymes that yield blunt end DNA fragments, followed by the ligation of a second vectorette and the specific amplification of a set of DNA fragments extending from one of the original primers (the first vectorette primer or a transgene primer) to each of the blunt end sites capped with the second vectorette. In this way sets of nested deletions could be rapidly generated and used for sequence contig construction. From experience it was found however, that this blunt ended approach was only suitable for sub-vectorette procedures since blunt ended vectorettes work very inefficiently in PCR amplifications from primary genomic vectorette libraries and therefore did not provide a useful means for isolating transgene/flanking DNA junctions from genomic DNA.

Flanking sequence homologies. Microinjection of DNA into the mouse pronucleus is a commonly used technique for introducing genetic material into mice. However, no previous evidence for non-random integration of linear DNA constructs into the mouse genome has been reported (reviewed in Palmiter and Brinster 1986). However in the present study 5 out of 8 of the MS32 transgene insertions were into mouse gamma satellite DNA which comprises only 5-10% of the mouse genome (Visuel and Choo 1989; Hastie 1989) and this suggests a non-random integration of these constructs (P = 0.0046, for a maximum of 10% gamma satellite DNA).
Gamma satellite DNA represents a different class of satellite array from the other alpha and beta satelltes (Waye and Willard 1986; Greig and Willard 1992) found in humans and the minor mouse satellites (Wong and Rattner 1988). The first consensus sequence for murine gamma satellite DNA from total uncloned Sau 96 I DNA monomers was published by Horz and Altenburger in 1981 (figure 5.11). Subsequently both Chaterjee and Lo (1988) and Vissel and Choo (1989) have sequenced monomeric gamma satellite repeats isolated from cell lines and have found a low level of sequence variation of individual repeats of between 0.9 - 9.1% from the original consensus sequence. The consensus sequence of gamma satellite repeats sequenced from the MS32 transgene flanking DNAs did not vary significantly from the original consensus sequence of Horz and Altenburger (1981); 4nt changes in a 234nt repeat unit giving a 98.3% sequence identity. One of the clones 10B sequenced by Vissel and Choo (1989) shows a strong sequence identity (97.8%) to the gamma satellite consensus sequence flanking the MS32 transgene loci. In contrast to other classes of satellite repeat the sequence homogeneity of gamma satellite appears well conserved; for example the sequence of the human alphoid satellite repeat has been reported to vary by as much as 20-40% between monomers (Waye and Willard 1986).

There are a number of possible explanations for this pattern of integration of the MS32 construct into the mouse genome. In two cases, 4-5bp of homology between the transgene repeat unit and the mouse gamma satellite are shared at the junction between the two sequences (figure 5.14). Since regions as small as 1bp have previously been suggested as having an effect on recombination leading to linear DNA integration (Rohan et al., 1990; Hamada et al., 1993), it is possible that these two integrations reflect some kind of sequence-specific targeting. However, this targeting is not site-specific since all five integrations into gamma satellite are at different positions in the 234bp repeat unit sequence and the orientation of integration is not consistent for all five insertions; the 109 locus has inserted with an opposite orientation into the gamma satellite repeat in comparison with the other four insertion events. An alternative explanation may be that gamma satellite is more likely to receive transgenes than other DNA regions. Perhaps the condensed chromatin structure of the gamma satellite DNA (Hastie 1989) is more likely than other regions of the genome to break under the forces of microinjection. If so, repair processes may be more likely to integrate a transgene into the gamma satellite DNA. A third possibility is that integration into single-copy DNA can result in dominant lethality, but that integration into gamma satellite causes no such deleterious effects. In this case lethal selection against integrants into single-copy DNA would result in an over-representation of gamma satellite integrants. A fourth possibility might result from the timing of the integration events in the murine zygote replication cycle. Previously in cell lines some repeated DNA sequences have been shown to replicate at a later stage in the cell cycle than non-repetitive sequences (Ten Haagen et al., 1990; Scott Hansen et al., 1993). Possibly a later integration of transgene constructs would result in more integrations into DNA sequences actively replicating at this time, for example gamma satellite DNA.

It is not clear whether preferential integration into gamma satellite DNA is a general phenomenon for all transgenes, since few studies have systematically addressed the nature of mouse DNA flanking sites of transgene integration. Other examples of satellite integration may not have been noted. Transfection of DNA into mammalian cell lines has yielded integrants flanked by centromeric DNA (Lo 1983; Wahl et al., 1984), with centromeric integration occurring at a frequency of 1/4 and 1/3 in these two studies. In some cases integrant expression was modulated by the surrounding centromeric DNA (Burner and Lo 1986). If transgene expression is repressed by the centromeric position of the majority of gamma satellite DNA (Vissel and Choo 1989; Hastie 1989), then screening mice for functional transgenes may overlook those integrations into this type of DNA, where
there may be a normal but inactive transgene. Even if transgene detection is by Southern blot analysis of genomic DNA rather than functional assays, the relative lack of restriction sites in gamma satellite DNA may make some transgenes undetectable, as was initially the case for the 102B and 110D loci. It may therefore be more appropriate to use enzymes such as AvaII, which cleave frequently in mouse gamma satellite, to identify satellite integrants.

The consensus sequence for mouse gamma satellite (Horz and Altenburger 1981) indicated that there were no sites for the restriction enzymes HindIII, BglII and BamHI. Using this information Vissel and Choo (1989) used pulsed field gel electrophoresis (PFGE) to assay for lengths of mouse gamma satellite arrays from total genomic DNA cleaved with these enzymes. They found bands of 240->2000kb in length which hybridised with a probe showing a 99% homology to the consensus sequence of Horz and Altenburger (1981). This indicated array sizes of between 1000 and 10,000 copies of the 234bp repeat units. The presence of satellite hybridising genomic DNA fragments of >2000kb indicated largely uninterrupted arrays of repeat units. Vissel and Choo (1989) also saw some array sizes of <300kb which suggested either variants in the gamma satellite repeat unit sequence resulting in the creation of one or more of the three enzyme sites used in the initial analysis (HindIII, BglII and BamHI) or the presence of non-satellite sequences interrupting the gamma satellite repeat unit array in an analogous manner to the frequently observed interruptions of human alphoid repeat arrays by for example I.L elements (reviewed in Singer 1982). In the present study we have seen mutations in the gamma satellite consensus sequence to produce MboI sites, which are not normally present but were utilised in the production of the primary vectorette libraries for 102A, 105A and 109 (figure 5.6 and 5.7). Additionally in this study the presence of a mouse IAP 5' LTR element approximately 3kb downstream of the 102A integration provides evidence of an interrupted gamma satellite repeat unit array.

Mouse IAP elements are found with a copy number of 1000 (per haploid genome) in the genome of Mus musculus and as a retroposable element they carry gag, pol and env genes between full length 5' and 3' LTRs (Kuff et al., 1983). Two regions of sequence identity between the mouse flanking DNA downstream of the 102A locus and the consensus mouse intracisternal A type particle sequence (Aota et al., 1987) were obtained; one at the junction between the mouse gamma satellite and the IAP 5' LTR and the second 168bp downstream which shows homology to the beginning of the gag gene. The level of sequence identity for both regions is about 90% since there is some variation in sequence between the two IAP clones, but this is not unexpected as the rate of base substitutions at IAP loci has been estimated to be 6-10 x 10^{-9} per site per year, higher than that observed for cellular genes (Aota et al., 1987). Since IAP elements are able to transpose throughout the genome it is not unexpected to find them in mouse gamma satellite, and in this study the insertion provided an MboI site downstream of the 102A insertion enabling the production of a primary vectorette library. There have been other examples of insertions of IAP elements into the mouse genome; Hawley et al (1982) describe the presence of an IAP element in two separate cell lines which has cause a defect in the mouse immunoglobulin kappa light chain gene.

The other three transgenes 105B, 110C and 110A (5') integrated into DNA sequences which showed no homology to other sequences present in the EMBL database. These were assumed to be single-copy (i.e. non-repetitive) in sequence. Most of the sequence flanking the 105B transgene locus shows no homology with other sequences in the EMBL database. However, upstream of the 5' terminus of the 105B integration the sequence shows an 80% sequence identity with a mouse B2 repetitive element (Krasyev et al., 1981). There is some variation from the
consensus sequence, but B2 elements in the mouse have been reported as showing high levels of sequence diversity (Hastie 1989) analogous to Alu repeat arrays in the human genome.

The use of primer sites in the mouse DNA flanking the transgene insertion 105B to amplify from non-transgenic animals did not result in a product (figure 5.16A). This indicated that a deletion or some other rearrangement in the endogenous mouse DNA has occurred as a result of transgene insertion. A deletion also appears to have taken place at the junction of the 110D transgene and mouse DNA, which conveniently appears to provide the first reported junction between mouse gamma satellite and non-repeated sequence (Allen et al., 1994). Deletions, transgene insertion promoted translocations and other rearrangements have previously been reported to occur post-microinjection in the mouse DNA (Woychik et al., 1985; Kato et al., 1986; Mahon et al., 1988; Krulewski et al., 1989 and reviewed in Pahniter and Brinster 1986). Another curious rearrangement in this study is seen with the transgenes 102A and 102B, which have arisen in the same founder mouse and contain 5' and 3' halves of the original construct. Both loci are flanked by mouse gamma satellite DNA but show no detectible linkage (chapter 4). It is possible that these two transgenes have resulted from a single original full length construct which has been disrupted by a rearrangement such as a translocation together with the loss of the central 13 repeat units. The absence of any trace of a full length construct in the founder mouse 102 or progeny DNA as shown by a PCR on 102 founder mouse and progeny DNA using the transgene flanking primers E1 and E2 (chapter 4), suggests that this disruption happened during integration. It has been proposed (Smith 1976; Vissel and Choo 1989) that the observed sequence homogeneity of mouse gamma satellite in comparison with other mammalian satellite arrays is maintained by unequal crossing over between blocks of almost identical gamma satellite on different mouse chromosomes. It is possible that the integration of the full length construct into the 102 founder mouse genome coincided with one of these recombination events and resulted in two halves of the original construct existing separately in the genome but with identical flanking DNA sequences.

The MS32 transgenic mouse lines in this study that have integrate into mouse gamma satellite provide a novel resource for studying the biology of these sequences and of the centromeres where they have been shown to cluster, though the exact positions of the MS32 transgenes are unknown, and it remains possible that they have integrated into ectopic blocks of gamma satellite rather than into centromeric gamma satellite.

Conclusions

In this study vectorette PCR has been shown to provide a highly flexible approach for the recovery of sequences flanking transgene insertions, from complex genomic DNA mixtures. Even prior to sequence analysis the vectorette and sub-vectorette products can play a role in confirming previous restriction maps and with vectorette libraries constructed from partial digested genomic DNA can clearly identify periodicitys in the flanking DNA, for example gamma satellite repeat arrays. The development of the sub-vectorette procedure provided an efficient method for sequencing larger products by double stranded sequencing methods. Surprisingly 5/8 of the MS32 transgene insertions were into mouse gamma satellite DNA, which suggested non-random integration, although sequencing of the junction regions between mouse and transgene revealed that this was not a direct sequence based targeting. The remaining transgenes integrated into single-copy DNA and any differences in the mutational behaviour of gamma satellite and single-copy DNA flanked transgenes will be analysed in chapter 6.
Chapter 6

Mutation analysis of the MS32 transgene loci

Summary

Allelic diversity at minisatellite loci is generated by a high rate of de novo mutation events to new length alleles. Initial studies analysing the linkage phase of markers flanking loci undergoing mutation indicated that simple unequal recombination was unlikely to be a common mechanism resulting in these events. More recent studies of the human minisatellite MS32 (D1S8) have looked at changes in the variant repeat interspersion pattern (by MVR-PCR) between mutant and progenitor alleles identified by pedigree analysis. These studies have revealed a polarity of mutation events and of variation in non-mutant alleles at the 5' end of the MS32 locus, indicating the possible existence of a mutational hotspot. This in turn led to the proposal for the existence of a cis-acting modulator of the MS32 locus mutation processes. Recent studies of MS32 alleles from male germ-line DNA using a single molecule PCR based method, small pool PCR (SP-PCR) for the identification and recovery of de novo mutations at the MS32 locus, indicated that the majority of de novo mutation events isolated in this way were the result of interallelic exchanges with a high incidence of gene conversion type events. As a first step to identify the important regulators of the MS32 mutation processes, it was important to ascertain if the transgenic MS32 alleles had maintained the mutational processes seen with the endogenous MS32 alleles, such as a polarity of mutation events, a bias towards gains in repeat units and examples of interallelic and gene conversion type events. In this chapter four single-copy and two multi-copy MS32 transgenic loci were analysed for evidence of mutation by pedigree analysis and by the development of the SP-PCR for use in transgene sperm analysis. Mutation frequencies of the single-copy loci were shown to be greatly reduced from those at the endogenous locus, but some mutation events were recovered by SP-PCR and analysed by MVR-PCR giving examples of repeat unit duplications, possibly resulting from unequal sister chromatid exchanges or from slippage events and of repeat unit switching for which there was no “donor” site, suggesting the occurrence of complex conversion type events. In addition one 110C mutant displayed a bona fide conversion event. The multi-copy loci showed high germline mutation rates and also showed a number of unusual mutation patterns, for example early embryonic mutation events, the frequent loss of a single restriction site and the common deletion of one entire multi-copy locus. Some mutation events identified by pedigree analysis were recovered and their variant repeat interspersion patterns analysed for evidence of the mutation processes involved. From the few events studied there was some evidence for repeat unit duplications which may result from slippage or from unequal sister chromatid exchanges.

Introduction

The existence of large numbers of minisatellite alleles of varying repeat copy number in human populations results in a high level of heterozygosity at these loci (Wong et al., 1986, 1987; Nakamura et al., 1987; Armour et al., 1989, 1990; Vergnaud et al., 1991; Desmarais et al., 1993). Most minisatellite loci exhibit no known phenotype, with the exception of some coding VNTR’s (e.g. the mucin protein (MUC1) VNTR, Swallow et al., 1987) and as such are thought to be selectively neutral (Kimura 1964, Ohta and Kimura 1973). This suggests that high levels of repeat unit length variation arise from an elevated mutation rate at these loci. The mutation rate at
some loci is high enough to be measured directly by pedigree analysis, to provide a measurement of germline mutation rate which is ~1% at the MS32 locus (Jeffreys et al., 1991b) and can be as high as ~15% for the CEB1 locus (Vergnaud et al., 1991). The detection of these mutation events and an estimation of their frequency in human populations is of direct relevance to the legal applications of minisatellite probes in paternity testing where a mutation event may falsely suggest the exclusion of an individual. More importantly for the purposes of this study the recovery of mutation events from both the endogenous and transgenic MS32 loci can help to shed light on the processes responsible for the elevated mutation rates at minisatellite loci.

**Minisatellite mutation.** Inferences about potential minisatellite mutation mechanisms were initially made by considering the general properties of minisatellite mutation events identified during pedigree analysis. Southern blot length analysis was used to study the inheritance of five separate minisatellite loci, MSI, MS8, MS31A, MS43 and p43 in the CEPH panel of families (Jeffreys et al., 1988). High rates of germline mutation were detected as heritable changes on a Southern blot profile. The mutation rates at these loci were found to increase with locus heterozygosity in accordance with the neutral mutation/random drift model of evolution (Ohta and Kimura 1973). Size analysis of mutations showed that small (4-10 repeat unit) changes were responsible for the majority of the mutation events and that there was no bias of gains or losses of repeat units (Jeffreys et al., 1988).

This would appear to rule out intramolecular recombination events as the major source of minisatellite mutation events since these would tend to cause losses of repeat units through the looping out of regions of the allele. However the data was consistent with the involvement of intraallelic mechanisms, such as sister chromatid exchange and replication slippage (Tautz et al., 1986; Levison and Gutman 1987) and with interallelic processes such as unequal recombination and gene conversion (Smith 1976; Dover 1982). In this study male and female germline mutation rates were apparently equal. Since there is a large discrepancy in the number of mitotic divisions that mature spermatoocytes have undergone in comparison with mature oocytes (~400 compared to ~24, Vogel and Nathenbug 1985) this observation argued against a major role for mitotic recombination or replication processes in the production of minisatellite mutations and conversely implied that mutation is restricted to one stage of gametogenesis, perhaps meiosis (Jeffreys et al., 1988b). However the majority of the mutations (maternal and paternal) were recovered from the locus MSI which had a five-fold higher mutation rate than that seen at the other loci (~5% vs. ~1%) and using data mainly from one locus with an unusually high mutation rate and therefore possibly an atypical mutation process, introduced a strong bias in the dataset. More recent data indicated that there may be a bias of gain mutations in the male germline at some minisatellites; for example Vergnaud et al (1991) have isolated a hypervariable minisatellite CEB1 (D2S90) which has a very high male specific mutation rate of ~15% per gamete as compared to the female mutation rate of ~0.3% per gamete. Additionally there appears to be a bias of paternal derived mutations at the minisatellite loci MS31 (D7S21) and MS205 (D16S309) (Jeffreys et al, 1994).

**Flanking marker analysis of mutants.** The direct analysis of isolated minisatellite mutation events can give more information on mutation processes at the locus. For example the hypothesis that simple interallelic unequal exchanges are involved can be tested for by looking for the exchange of markers in the DNA flanking the repeat unit array. By this type of study a single de novo mutation event at the human minisatellite YNZ22 was shown not to result in the exchange of flanking markers (Wolff et al., 1988), suggesting that the event did not involve interallelic exchanges of DNA. In a larger study of twelve mutation events at the human MS1 locus, Wolff...
et al (1989) again found no evidence of an elevated exchange of flanking DNA markers in the surrounding 10 cM interval. Vergnaud et al (1991) also found no evidence for an elevated rate of flanking marker exchange for mutation events at the CEB1 locus. Together these results indicate that simple unequal recombination events do not play a major role in minisatellite mutation at these loci. Flanking marker analysis only tests for evidence of unequal crossovers and cannot indicate the occurrence of a gene conversion event or of mutation events resulting from intraallelic mechanisms that may have been involved, for example sister chromatid exchange or replication slippage.

Minisatellite variant repeat (MVR) analysis of mutants. Changes in the variant repeat unit interspersion patterns between progenitor and mutant MS32 alleles can be analysed for clues as to the mechanisms involved in the mutation event. Enzymically based MVR mapping was first used to study the changes in the variant repeat unit interspersion patterns of deletion mutants (Jeffreys et al, 1990) in sperm and blood tissues using a single molecule approach. Just under 80 blood and sperm mutants were recovered from pools of size fractionated Mbo I digested DNA, giving an approximate 0.07% deletion rate per haploid genome with small deletions occurring more frequently than large deletions. There was no evidence of interallelic repeat unit exchange in the mutants but a pattern of clustering of deletion endpoints towards the 5' end of the alleles (designated 3' end in this paper and subsequently redesignated the 5' end) was observed for both blood and sperm mutants. The data did not suggest a single hotspot of deletion at the 5' end of the allele, but rather an excess of deletions terminating in the last 20-30 repeat units. In this study a low level of mosaicism was observed for 40% of sperm DNA mutants indicating a premeiotic origin for these events which would suggest that early hypotheses speculating that hypervariable human minisatellites are involved in chromosome paring and/or the initiation of meiotic recombination do not hold true for these mosaic MS32 deletion mutants. The nature of the mutation process resulting in the remainder of the blood and sperm mutants was unclear from this study, but it was proposed that unequal sister chromatid exchange and replication slippage could still be viable mechanisms. The study provided no information on the processes operating during oogenesis, nor about mutation events resulting in increases in allele length.

MVR-PCR to identify mutant MS32 alleles. The advent of MVR-PCR mapping of single alleles both from unrelated individuals and in pedigrees showed that the level of allelic heterozygosity at the MS32 locus was far higher than previous allele length estimates from Southern blot analysis (Jeffreys et al., 1991b). Allele alignments of MS32 alleles from CEPH individuals indicated that the variation between groups of closely related alleles is largely restricted to the 5' end of the alleles (Jeffreys et al., 1991b). This polar bias of variation can be assumed to result from an analogous polar bias of mutation events at the 5' end of the MS32 repeat unit array. Pedigree analysis of CEPH families revealed 7 mutation events (Jeffreys et al., 1991b), which all involved small gains of repeat units close to the 5' end of the array. Two of these mutations provided the first direct evidence for an interallelic exchange of repeat units. In the first of these, the first 11 repeat units from one allele were inserted into the first repeat unit of the other allele, with no loss of information from the recipient allele. Analysis of a 5' flanking base substitutional polymorphism (Hump 1, Monckton., 1992) at both alleles showed that it was not exchanged ruling out unequal recombination involving the exchange of 5' flanking DNA. In the second interallelic mutant 3 repeat units from the paternal allele appeared to have been inserted into the first two repeats of the maternal allele. Unfortunately this individual was homozygous for flanking markers in the 5' MS32 flanking DNA so unequal recombination could not be ruled out as a potential mechanism (D.Monckton, 1992).
In general the results indicated that simple unequal recombination was not the major mutation mechanism involved in the creation of hypervariability at the MS32 locus, but that some other interallelic process could be involved. From analysis of variant repeat interspersion patterns at MS32 (Jeffreys et al., 1990, 1991b, 1994) it appears that allelic variability results from two major mutation pathways. The first, deletions appear to occur at a similar frequency in somatic and germline tissues and can involve the loss of a few or of a large number of the repeat units in an allele. The deletion events studied remove repeat units from the internal regions of minisatellite alleles, but show a tendency to occur towards the 5' end of the MS32 locus. The other mutation process involves the gain of repeat units and shows much greater polarity with most events localised to changes in the first few repeat units at the 5' end of the array. These may be derived from interallelic exchanges involving a short stretch of repeat units, possibly in a gene conversion-like process or alternatively result in duplication events arising by replication slippage processes or by unequal sister chromatid exchanges. From pedigree analysis, gain mutations appear to be more frequent than deletions in the germline and it is proposed that these processes account for the majority of allelic diversity at the MS32 locus. This proposal makes sense for the hypervariable loci studied in this laboratory. If the majority of minisatellite mutations were losses of repeat units then minisatellite alleles would never evolve to the high repeat unit copy number characteristic of hypervariable minisatellite loci such as MS32.

Mutations identified using small pool PCR (SP-PCR). Although pedigree analysis can reveal both paternal and maternal derived mutation events the numbers of mutation events available for analysis are limited. With a pedigree mutation rate estimated at approximately one mutant allele in every hundred studied, large numbers of blood samples need to be processed to identify more than a few mutation events. However, a single molecule PCR based technique was developed to assay for de novo mutation events at individual MS32 alleles, called small pool PCR (SP-PCR). The basic strategy behind the SP-PCR technique is outlined in figure 6.1. This technique has been applied to a number of minisatellite loci (see chapter 8) and enables the calculation of mutation rates at the minisatellite locus, per number of molecules analysed as well as the recovery of mutant MS32 alleles for internal structural analysis by MVR-PCR. In the SP-PCR, each PCR reaction shows an intense signal from the progenitor allele plus the occasional appearance of products of an abnormal length. These extra bands are assumed to represent PCR products from single mutant molecules on the basis of several lines of evidence (Jeffreys et al., 1994). The signal from one mutant band is equivalent to the intensity of signals seen from progenitor alleles amplified in a single molecule PCR, consistent with the abnormal length product originating from a single mutant molecule. Additionally the signal from these abnormal length bands is quantal in nature; prolonged exposure of autoradiographs does not reveal any fainter bands as would be expected from PCR artefacts initiated at different stages in the PCR. The frequency of these mutation events is also proportional to the amount of input DNA in each PCR reaction. Results from studies on human sperm and blood DNA have revealed a decreased mutation rate in the blood DNA, where there are clean backgrounds in each PCR lane, indicating the specificity and efficiency of the technique (Jeffreys et al., 1994). In addition some of the mutants in the human studies were recovered and structurally analysed using MVR-PCR. Many of the resulting maps showed differences from the progenitor allele inconsistent with having arisen from PCR artefact processes, e.g. evidence for interallelic exchanges (Jeffreys et al., 1994). Mutation rates as low as $10^{-4}$ per gamete have been measured using the SP-PCR technique (A.J. Jeffreys, personal communication).
Figure 6.1 Basic strategy for Small pool PCR at minisatellite loci

1. Digest the DNA source with an enzyme that cuts 5' and 3' of the minisatellite locus for study (e.g. MS32).

2. Accurately measure the concentration of the DNA by comparative agarose gel electrophoresis, and fluorimetry.

3. Assuming that 3pg of DNA is the amount of DNA that represents one haploid genome (6pg for the hemizygous transgenic loci) perform PCR on multiple dilute aliquots of DNA, across the whole locus.

4. Electrophorese an aliquot of the PCR on a 40cm, 1.0-1.2% agarose gel, Southern blot and hybridise with a repeat unit probe (e.g. MS32).

5. Mutant molecules will appear as bands of a different length from the progenitor.

6. To recover mutants for analysis by MVR-PCR, electrophorese a second aliquot of the PCR reaction corresponding to the lane on the gel where the mutant band was identified, under identical conditions but without ethidium bromide, and recover the mutant DNA by electroelution.

7. Analyse the variant repeat unit structure of the mutant DNA by MVR-PCR.
Data from studying sperm DNA at 18 endogenous human MS32 alleles (Jeffreys et al., 1994) revealed a mutation rate per gamete of 0.81%. The mutation rate was independent of allele length. The majority (93%) of mutation events were clustered within a size range of 1-20 repeat units from the progenitor allele. Of these 74% were gain mutations, in line with preliminary pedigree mutation data that suggested gain mutants were more common than deletion events (Jeffreys et al., 1991b). These germ-line events showed a 5' polarity of mutation. In some cases (approximately one quarter to one third of mutation events) they resulted from a simple reduplication of the progenitor allele, probably from intrallelic processes such as replication slippage or unequal sister chromatid exchange. However the majority of the events (50%) were shown to involve the interallelic transfer of repeat unit blocks with repeat units from a 'donor' allele fused into the 'recipient' allele repeat unit array. The transfer of repeats appeared to be non-reciprocal in that both alleles were capable of acquiring repeat units from the other, however no examples of flanking marker exchanges were observed indicating that recombination events were not involved in the mutation events studied. In many cases anomalous repeat units or repeat unit blocks were acquired with no apparent origin in either allele. These are classic indications of complex gene conversion type events occurring during the MS32 mutation process at these alleles.

The tendency for mutation events from pedigree and SP-PCR analysis to be polarised at the 5' end of the repeat unit array indicated the potential for the 5' flanking DNA to be involved in the mutation processes. Recent evidence has revealed the presence of a 5' base polymorphism upstream of the start of the repeat unit array within the 320 primer site, referred to as the 01 polymorphic site (see figure 3.6, chapter 3. There can be either a 'C' or a 'G' nucleotide at this position. Mutation analysis of human sperm by SP-PCR has revealed that this polymorphism is associated with changes in MS32 mutation rate; alleles that carry a 'G' at this position show a mutation rate within the normal range, but alleles with a 'C' at this position show a reduced mutation rate (Monckton et al., 1994). It is likely that the sequence region of this polymorphism is involved in the regulation of mutation processes at the MS32 locus.

This work. The creation of a model of MS32 mutation in a mammalian system was one way of expanding our knowledge of the mechanisms of mutation operating at the MS32 locus and to attempt to identify the important regulators of these mutation processes. The production and analysis of mice transgenic for the human minisatellite MS32 was a first step towards the creation of such a model. Structural analysis of the MS32 transgenes indicated that the basic locus structure was different for each of the insertion events. The mouse DNA flanking each insertion was either single-copy or mouse gamma satellite and individual locus mutation analysis was performed in order to assess the effects of transgene structure and flanking DNA on MS32 transgene mutation rate and process.

In this chapter the transmission of six of the eight MS32 transgene inserts is analysed. Two of these loci, 109 and 110A had a multi-copy structure, and four 102A, 102B, 110C and 110D had a single-copy structure. The initial breeding strategy bred hemizygous transgene positive mice with non-transgenic mice, allowing the transmission and mutation rates of the MS32 transgene at each locus to be measured by pedigree analysis. Southern blotting of digested tail DNAs prepared from the progeny of these matings enabled the identification of transgene-positive mice and of mice where a mutation had occurred which could then be bred for further analysis if necessary. The initial mutation analysis was performed on mice hemizygous for the transgene locus and therefore was a measure of intrallelic mutation events. Later breeding strategies bred hemizygous vs. hemizygous transgenic mice in order to
produce mice homozygous for the transgene loci and to compare the mutation rates of these with those for hemizygous mice. Mice that are homozygous for transgene loci have two alleles and can therefore provide a measure of the rate of interallelic events at the transgene loci. Three of the mutations identified by pedigree analysis of the multi-copy locus 110A were isolated and analysed for variant repeat unit changes by MVR-PCR.

In addition to pedigree analysis, mutation analysis of the loci was performed by PCR, using an adaptation of the SP-PCR (Jeffreys et al., 1994). Where transgenes were 5' or 3' truncated and did not have the appropriate MS32 flanking DNA to provide primer sites for SP-PCR, new primer sites were developed in the mouse DNA flanking the insertion. Some of the transgene mutation events identified by SP-PCR for the single-copy loci were recovered and the changes in their internal variant repeat unit arrays examined using MVR-PCR. The frequency and type of mutation events seen at the transgenic loci are compared to the type of event seen at the endogenous human locus.

The results presented in this chapter draw on the work of a number of individuals; in particular the large breeding program of MS32 transgenic mice was designed by A. Collick, and the mouse colony maintained in Leicester by A. Collick and in Cambridge by M. Norris. Over 1500 tail DNAs were prepared for pedigree analysis and subsequently analysed by a number of individuals; A. Collick, M. Norris, P. Bois, and F. Scaerou and myself. My personal analysis was of 520 progeny from a mixture of hemizygous x non-transgenic and hemizygous x hemizygous crosses for different loci. 352 of these animals were transgene-positive and this data is presented with the data from the whole group research to ascertain total mutation rates for each locus.

**Results**

**A. Mutation analysis of the single-copy MS32 transgene loci**

Rates of mutation and stability of transmission. The transmission and mutation data for the six lines of transgenic loci are presented in table 6.1. For mutation analysis the most informative combination of restriction enzymes for each transgene locus was used. Line 110D DNA was cut with Ava II or Eco R I to release a single transgene fragment. Line 110C DNA was cleaved with Mbo I to release one fragment or with Pst I to release two separate hybridizing fragments corresponding to the two separate blocks of MS32 repeats. Lines 102A DNA and 102B DNA were cut with Ava II or Eco R I to release a single hybridizing fragment.

No mutation events were seen by pedigree analysis for any of the four single-copy transgene loci (for example see figure 6.2). Minor variations in band intensity were sometimes seen due to variations in the amount of DNA loaded. Occasionally this made hybridizing bands run at a marginally different position to the progenitor locus. However the non-mutant status of these bands was further checked by internal repeat unit mapping (MVR-PCR using the flanking primers 32O or 32E and associated repeat unit specific primers, see chapter 4) and this did not reveal any changes in the allele length or internal variant repeat maps of the locus in these animals DNA (data not shown). For the 110D locus 216 transgene-positive progeny were analysed for length change mutations by Southern blot pedigree analysis. No mutations were seen. Therefore, the \( \mu \) (mutation rate per gamete) is \( \frac{0}{216} \) (\( P > 0.95 \)). Total numbers of single-copy loci analysed give a \( \mu \) (mutation rate per gamete) of \( \frac{c}{e} \) (\( P > 0.95 \)).

Chapter 6, page 6
Table 6.1. Pedigree data for the six separate transgenic lines.

<table>
<thead>
<tr>
<th>Locus</th>
<th>mating</th>
<th>numbers of positive progeny analysed</th>
<th>maximum number of alleles</th>
<th>total number of mutation events</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>102A (S/C)</td>
<td>hemi x non-TG</td>
<td>39</td>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>hemi x hemi</td>
<td>39</td>
<td>52</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>78</td>
<td>91</td>
<td>0</td>
</tr>
<tr>
<td>102B (S/C)</td>
<td>hemi x non-TG</td>
<td>16</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>16</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>110C (S/C)</td>
<td>hemi x non-TG</td>
<td>24</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>hemi x hemi</td>
<td>28</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>52</td>
<td>61</td>
<td>0</td>
</tr>
<tr>
<td>110D (S/C)</td>
<td>hemi x non-TG</td>
<td>136</td>
<td>136</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>hemi x hemi</td>
<td>80</td>
<td>107</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>216</td>
<td>243</td>
<td>0</td>
</tr>
<tr>
<td>109 (M/C)</td>
<td>hemi x non-TG</td>
<td>129</td>
<td>129</td>
<td>12 (germline) 6 6</td>
</tr>
<tr>
<td></td>
<td>hemi x hemi</td>
<td>74</td>
<td>99</td>
<td>8 (4 embryonic) 4 4</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>203</td>
<td>228</td>
<td>20 (9.8%) 10 10</td>
</tr>
<tr>
<td>110A (M/C)</td>
<td>hemi x non-TG</td>
<td>69</td>
<td>69</td>
<td>8 (3) 5</td>
</tr>
<tr>
<td></td>
<td>hemi x hemi</td>
<td>96</td>
<td>128</td>
<td>0 (0) 0</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>165</td>
<td>197</td>
<td>8 (5.0%) 3 5</td>
</tr>
</tbody>
</table>

(S/C) = single-copy transgene locus,
(M/C) = multi-copy transgene locus,
Hemi = hemizygous mouse, non-TG = transgene negative mouse.
'Maximum number of alleles' assumes 1/3 of progeny from hemizygous x hemizygous crosses will be homozygotes with two alleles.
Total mutation rates are shown in brackets for the multi-copy loci 110A and 109.
Total mutation numbers for the multi-copy loci are broken down into germline or embryonic events.
Figure 6.2 Autoradiographs of Southern blots of tail DNA from progeny of single-copy transgene positive parents.
The Southern blots were probed with a 5kb MS32 repeat unit probe in a phosphate/SDS hybridisation mix overnight at sixty five degrees.
A. *Ava II* (Av) digests of 110D progeny; no length change mutations are visible. The *Ava II* band is 2.34kb in length.
B. *Mbo I* (Mb) digests of 110C progeny; no mutations. The *Mbo I* band is 4kb in length.
C. *Ava II* (Av) digests of 102A and 102B positive progeny. No mutations were visible. The *Ava II* band for the 102A locus (A) is 1.8kb and for the 102B (B) locus is 1.5kb.
which is the equivalent of <0.8%. The mutation rate in humans is 0.8-1% (Jeffreys et al., 1991, 1994) so it is not possible to say if the mutation rate is reduced for the single-copy loci as a whole. Assuming that some mutation events will occur even if the rate is lower than that seen with the endogenous MS32 locus, then very large numbers of progeny would need to be analysed to obtain these mutation events by pedigree analysis. Few 102B mice were bred since this is a truncated transgene flanked 5' and 3' by mouse gamma satellite DNA and as such does not provide an easily accessible model for mutation analysis by PCR, and was not deemed a high priority locus for study. The low numbers of animals bred at the 110C locus reflect breeding problems that were experienced with mice carrying this locus in later stages of the breeding strategy, resulting in fewer progeny for analysis.

Therefore another technique of analysing large numbers of individual MS32 transgene alleles was required to ascertain the mutation rates at the single-copy loci. The SP-PCR technique had previously been successfully used to recover single molecule mutation events, enabling the calculation of mutation rates per individual allele and the subsequent analysis of mutation events from single sperm by MVR-PCR (Jeffreys et al., 1994). The technique was adapted to enable the determination of mutation rates at three of the single-copy MS32 transgenic loci; 110D, 110C, 102A. Additionally the technique facilitated the recovery of some mutant MS32 single-copy transgene alleles for internal structural analysis by MVR-PCR and comparison with mutation events at endogenous MS32 alleles.

SP-PCR at the MS32 transgene locus 110D. Since this insertion has full 5' and 3' MS32 transgene construct flanking sequences, SP-PCR was performed on sperm DNA from a mouse carrying the 110D locus, using the transgene flanking primers E1 with E2 and 32A with 323D (figure 6.4 A). The sperm DNA was cleaved with Ava II which cuts in the mouse flanking DNA either side of the transgene and the concentration of the DNA was assessed by comparative agarose gel electrophoresis and by fluorimetry. To enable the calculation of an absolute mutation rate for the number of individual sperm in a reaction the SP-PCR efficiency using this DNA was determined by performing multiple dilutions of the sperm DNA into PCR reactions and correcting for single molecule PCR efficiency, as established by Poisson analysis of single molecule dilutions of sperm (Jeffreys et al., 1994). The transgenic DNA was diluted to single molecule levels with 5mM Tris-HCl (pH 7.5) and 0.1|xM carrier primer 32A. Sixty 7|fl PCR reactions each containing 5.12pg of sperm DNA were performed using primers 32A and 32D to amplify across the entire 110D locus for 25 cycles of 96° 45s, 68° 1min, 70° 3min. The PCR products were resolved by agarose gel electrophoresis and products detected by Southern blot analysis and hybridisation in a phosphate/SDS mixture at 65° overnight with a 5kb MS32 repeat unit probe. The autoradiograph shown in figure 6.3 B represents 30 of these PCR reactions from 110D positive mouse sperm DNA, and includes 16 positives. Of the other 30 reactions performed 15 positives were identified. This gave a failure rate of 29/60 reactions = 0.483 per reaction. From Poisson, e^-m = 0.483, giving a value of m = 0.727. Therefore from this experiment there are 0.727 amplifiable molecules per 5.12pg of DNA, or 0.85 per 6pg of DNA. This gives a PCR efficiency of 85%.

Neither combination of transgene flanking primers E1 plus E2 or 32A plus 32D, resulted in a Southern blot profile indicating a level of locus mutation comparable with that seen at the endogenous locus (Jeffreys et al., 1994). A total of 5 putative mutation events were observed in 110D sperm DNA representing an estimated 68,000 individual sperm. This indicated that the male germline mutation rate at the 110D transgene locus was 5/68,000 or 0.007%.
Figure 6.3. Small pool PCR analysis of sperm DNA from a transgene 110D positive mouse.

(first page of figure)

A. Schematic of the MS32 transgene locus 110D. Restriction sites shown are Ava II (Av), Pst I (P), Hind III (H) and Eco R I (E). Primer sites shown are E1/E2, 32A and 32D and the nested primers 32E and 32O.

B. Poisson analysis of DNA concentration at the single molecule level. Poisson analysis was performed as described in the text. The products of sixty reactions were separated by electrophoresis through 1% agarose, Southern blotted and hybridised with a 5kb MS32 repeat unit probe in a phosphate/SDS hybridisation mixture at 65°C overnight. The autoradiograph shown here represents 30 of these PCR reactions, and includes 16 positives. Of the other 30 reactions performed 15 positives were identified.

C. Small pool PCR experiment amplifying from multiple dilute aliquots of 110D sperm DNA. Each separate reaction contains enough sperm DNA to represent 320 individual sperm. Amplifications were with the transgene flanking primers E1 and E2 for 25 cycles of 96 45s, 68 1min, 70 3min, in a 7.5μl reaction. 2μl aliquots of each PCR were electrophoresed on 40cm, 1% agarose gels for 16 hours at 200V. Southern blots were hybridised with a 5kb MS32 repeat unit probe at 65°C overnight in a phosphate/SDS hybridisation mixture. One presumptive mutation event (mut 1.), a gain of 8 MS32 repeat units is indicated.

(second page of figure)

D. E. MVR-PCR analysis of transgene 110D sperm mutant 1. DNA representing this mutation (mut.1) was recovered by electroelution onto dialysis membrane and analysed by 2-state MVR-PCR and 4-state MVR-PCR to identify any changes in the transgene 110D variant repeat unit map. In addition to indicating the position of an 8 repeat unit duplication, MVR-PCR analysis also revealed the presence of a region of repeat unit type switching (see F). Transgene flanking primer 32O was used for the MVR-PCR reactions. Amplifications were for 15 cycles from the electroeluted DNA fraction and 18 cycles from progenitor 110D genomic DNA (prog.). Southern blots were hybridised as in (A).

F. The 4-state MVR-PCR variant repeat unit maps of mutant 1 and of two other mutants recovered. The dashes (-) in mutant 1 are to improve alignment of the progenitor MVR map with the mutant MVR map. The first region of underlined repeat units indicates an eight repeat unit duplication and the next underlined region a four repeat unit type switch. The dashes (-) in mutant 2 represent an eleven repeat unit deletion. The underlined repeat units in mutant 3 represent a twenty two repeat unit duplication.
Chapter 6, figure 6.3
Chapter 6, figure 6.3

PROGENITOR yeEEEEYoEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEyEeEY
This mutation rate was approximately one hundred times less than the average rate measured for the endogenous locus of 0.8% (Jeffreys et al., 1994). The Southern blot profile of one of these mutation events is shown in figure 6.3 C. The signal from this mutant is similar to the signal from figure 6.3 B, where each band in the poisson analysis is representative of 1-3 individual sperm. This indicates that the mutant event seen in figure 6.3 C is of the appropriate signal intensity to represent PCR products of a single mutant sperm molecule. This mutant resulted from a gain of 8 repeat units and the autoradiographs of the Southern blots of the 2-state and 4-state MVR-PCR reactions from the electroeluted mutant DNA fraction are shown in figure 6.3 D and E. Mutant DNAs were recovered from two of the other mutants, one gain mutant resulting from a 22 repeat unit duplication and one 11 repeat unit deletion mutant. The 4-state MVR-PCR maps of all these mutants are shown in figure 6.3 F. The other two mutation events identified for the 110D locus were small repeat unit gains of 2-3 repeat units. These were from pools of 320 genomes worth of DNA per PCR reaction and the concentration of progenitor DNA present was too great to allow for the recovery of pure mutant DNA for analysis by MVR-PCR.

SP-PCR of the MS32 transgene locus 110C. The structure of this transgene was circularly permuted (see chapter 4) and MS32 transgene flanking primers which can amplify across the entire locus were not available. However, the mouse DNA flanking these insertions was single-copy and therefore primer sites were designed in the 5' and 3' flanking mouse DNA extending toward the transgene locus (figure 6.4 A). Potentially this allowed for the analysis of the full construct using the mouse flanking primers or of the two separate blocks of repeat units individually by the use of a combination of mouse flanking primers and transgene specific primers (e.g. analysis of the 5' block of repeat units by a combination of 110C-5 and E1 primers and of the 3' block by a combination of E2 and 110C-3). However attempts to create an efficient specific primer rooted in the 5' mouse flanking DNA were unsuccessful, producing either no product or a series of non-specific products for two separate primer sequences tried (data not shown). It was decided therefore to concentrate on the analysis of the 3' block of repeat units downstream of the 5' MS32 construct flanking DNA (figure 6.4 A), since the 3' flanking primer 110C-3, which is 260bp downstream of the last MS32 repeat unit in the mouse DNA, gave a 1.32kb product with primer E2 as predicted.

SP-PCR was performed on sperm DNA digested with Mbo I, which cleaves either side of the 110C transgene locus, using the primers E2 and 110C-3 (figure 6.4 B). It was not possible to score for deletion mutations since beneath the progenitor band in the Southern blot autoradiograph there were additional hybridizing bands corresponding to up to 19 of the 29 MS32 repeat units in the 110C 3' repeat block map (figure 6.4 B). The greater the concentration of input 110C DNA, the greater the extent of the additional bands produced. The SP-PCR was repeated using a hemi-nested strategy of primer pair 32D and 110C-3 on Mbo I digested DNA. The same type of banding pattern was observed on the Southern blot autoradiograph (data not shown). To test if the other upstream block of MS32 repeat units (45 repeat units long) was affecting the block under analysis, the SP-PCR experiments with both E2/110C-3 and 32D/110C-3 primer pairs were repeated on Pst I cut DNA. Restriction digestion by this enzyme results in the separation of the 5' and 3' blocks of MS32 repeat units before PCR, thus preventing the upstream block from affecting the downstream block during the PCR. The same type of banding pattern was observed (data not shown). This indicated that the additional bands were probably the result of some form of non-specific binding of the 110C-3 primer rather than the transgene specific primers or interference from the upstream block of MS32 repeat units.
Figure 6.4. Small pool PCR analysis of the 110C locus.

A. Schematic of the 110C locus. The transgene DNA primer sites E1, E2, 32A (A), 32D (D), 32O (O) and 32E (E) and the mouse DNA flanking primers 110C-5 and 110C-3 are shown. The restriction sites shown are those used to digest the transgene 110C sperm DNA prior to SP-PCR analysis, *Mbo* I (Mb) or *Pst* I (P).

B. Autoradiograph of one transgene 110C SP-PCR experiment. In this experiment 12 tubes containing 110C *Mbo* I cut DNA equivalent to 40, 80 and 160 genomes worth (gw) of sperm DNA are shown. Products were separated by agarose gel electrophoresis on 1% agarose, then Southern blotted and hybridised with a 5kb MS32 repeat unit probe in a phosphate/SDS hybridisation mixture at 65°C overnight. Amplifications were for 25 cycles with the transgene primer E2 and the 3’ mouse flanking DNA primer 110C-3, to give a 1.32kb progenitor band. This primer combination assays mutation rates in the 3’ block of MS32 repeat units (see A). Two gain mutations are indicated. Only gain mutation events could be scored since the 110C-3 primer in combination with transgene primer E2 (as shown) or 32D (not shown) gave a series of additional MS32 repeat unit hybridising products corresponding to up to 19 MS32 29bp repeat unit deletions from the size of the progenitor. The greater the input of DNA the greater the number of additional bands detected, but there were 14 strong deletion artefact bands corresponding to 14 MS32 repeat units for each DNA concentration (40, 80 and 160gw).

C. Four state MVR-PCR maps of the original progenitor, the 110C 5’ and 3’ repeat unit arrays and the recovered 110C mutation event. The mutant analysed here is not one of the mutations visible on the autoradiograph in B, but represents a 22 repeat unit gain from the 110C 3’ repeat unit array. Two changes between the 3’ repeat array and the mutant are highlighted in bold; a two repeat unit type switch at positions 18 and 19 and a fifteen repeat unit duplication, positions 31-45. An additional eight repeat unit difference between the mutant and the 110C 3’ repeat unit array is underlined, positions 46-53.

D. A putative mechanism of mutation for the recovered 110C mutation. There is a donor site for the fifteen repeat unit duplication in the 110C mutation in the 5’ MS32 repeat unit array. If this is aligned with the position of donation into the 3’ MS32 repeat unit array then the two repeat unit type switch may also be explained by an interarray donation of a ‘yE’ from a position one repeat unit out of alignment, in the 110C 5’ array. The positions of putative exchange are highlighted in bold. It is not possible to suggest an origin for the underlined changed repeat units at the terminus of the 110C mutant.
A

5' block of MS32 repeats
43 repeats

3' block of MS32 repeats
31 repeats

HOC 5

HOC 3'

HOC mutant

B

12 tubes of 160gw
12 tubes of 80gw
12 tubes of 40gw
No-DNA control

12 tubes 30-41
12 tubes 18-29
12 tubes 6-17
1-5

Gain mutants

Progen, band

Additional bands/PCR artifacts

C

4 state MVR-PCR codes -

1 ........ 10 ........ 20 ........ 30 ........ 40 ........ 50 ........ 60 ........ 70
progenitor yEEEYeyEyEEEEEYeeyEyEeyEyeyEeeEeyEYEEYyYEEYyYyYyYyYyYyYyYyYyYyY
110C 5' array ----------------- yEeyEeeEyEYEEeEyEYEEYyYEEYyYyYyYyYyYyYyYyYyYyYyY
110C 3' array ????yEeyEeeEyEYEEeEyEYEEYyYEEYyYyYyYyYyYyYyYyYyYyYyY

110C mutant ????eY??EyEEEEYyEyEyEyEEYYEEEYyYyYyYyYyYyYyYyYyYyYyYyY

D

Putative 'interarray' interactions -

110C 5' array ----------------- yEeyEeeEyEYEEeEyEYEEYyYEEYyYyYyYyYyYyYyYyYyYyYyY
mutant ????eY??EyEEEEYyEyEyEyEEYYEEEYyYyYyYyYyYyYyYyYyYyYyYyY
110C 3' array ????yEeyEeeEyEYEEeEyEYEEYyYEEYyYyYyYyYyYyYyYyYyYyYyY

Chapter 6, figure 6.4
However, above the progenitor band on the Southern blot there was no trace of this type of multi-band pattern. Occasional single bands were observed and these were scored as gain mutation events. Two presumptive mutation events were seen from a total \( Mbo \) I cut DNA sample representing a total of 2960 sperm molecules in two separate PCR reactions (figure 6.4 B). These mutation events were visible as gain mutations of 13 repeat units (1.7kb) and of 10 repeat units (1.6kb). Two more presumptive gain mutations of 22 repeat units (1.92kb) and of 15 repeat units (1.8kb) were identified in further analyses using \( Pst \) I cut DNA, giving a total of four gain mutants in DNA corresponding to 9520 sperm molecules. This gave a male germline mutation rate 0.04% (95% confidence limits 0.01-0.21%) per gamete, which is twenty times less than the rate at the endogenous human locus (Jeffreys et al., 1994).

The 22 repeat unit gain mutant (autoradiograph not shown) was successfully recovered by electroelution and analysed by 4-state MVR-PCR and the variant repeat unit map is presented in figure 6.4 C. Interestingly it appears that the upstream block of MS32 repeat units has acted as a donor, in this 110C mutation event, for a block of 15 repeat units not present in the 3' MS32 repeat unit array, but present in the 5' MS32 repeat unit array. This mutant event was recovered from 110C \( Pst \) I digested DNA. Therefore this is likely to have been a \textit{bona fide} conversion mutation event since this upstream MS32 repeat unit array was not adjacent to the downstream array during the SP-PCR reaction.

**SP-PCR of the MS32 transgene locus 102A.** This transgene locus is 3' truncated and does not have any 3' MS32 flanking DNA. Therefore the primers E1, 32D and 32E are not available for small pool PCR (figure 6.5). Additionally the immediate 5' and 3' mouse flanking DNA was gamma satellite which meant that most of the mouse flanking DNA could not be used for primer sites, since gamma satellite is present in a high copy-number throughout the mouse genome and a primer based on this sequence would not be specific. The possibility of a primer at the MS32 repeat unit and gamma satellite junction was considered, but this was a transgene locus where there were 4bp of sequence homology between the gamma satellite and MS32 repeat unit (figure 5.14, chapter 5). This made the design of a specific primer overlapping this junction impossible. However approximately 3kb downstream of the 102A locus is a mouse IAP element 5' LTR and gag gene sequence. Two primers were designed in the IAP 5' LTR sequence, IAP I and IAP II and used in combination with the transgene primers E2 and 32O (figure 6.5 and figure 6.6 A ) to amplify from the 102A locus to the IAP across the mouse gamma satellite sequence from 30ng of starting genomic DNA. These primer combinations gave a pattern of additional bands corresponding to almost every gamma satellite repeat unit in the region between the terminal MS32 repeat unit and the IAP LTR sequence (12-13) though not apparently from each MS32 repeat unit. Similar "collapse" patterns were observed for a number of different mice carrying the 102A locus (F.Scaerou, personal communication) and the same pattern was observed in tail DNA as well as sperm DNA (figure 6.6 C). These bands may be PCR artefacts or they may represent an exceptional level of variation/mutational instability at the gamma satellite surrounding the 102A locus.

To investigate this unexpected pattern, SP-PCR experiments were performed using aliquots of \( Mbo \) I digested 102A sperm and tail DNA from the same mouse. \( Mbo \) I was used to digest the DNA since it cleaves in the middle of the downstream IAP gag gene sequence (see figure 5.13, chapter 5) and leaves the IAP I and IAP II primer sites plus the 102A 3' flanking mouse gamma satellite DNA array intact for PCR. Pools of 5, 10, 20, 40 and 80 genome
Figure 6.5. Structure of the 102A locus showing the primers used in small pool PCR analysis of mutation. The transgene specific primer sites shown are E2, 32D (D) and 32O (O) and the IAP specific primer sites IAP I and IAP II. Sequences for these primers are detailed in table 2.1. Restriction sites shown are Eco R I (E), Pst I (P) and Mbo I (Mb). The large arrows represent the mouse gamma satellite sequence flanking the 102A transgene insertion.
A. **PCR analysis using the IAP specific primers.** Amplifications were from 30ng of total genomic transgene 102A positive mouse tail DNA for 18 cycles using the primer pairs E2/IAP I, 32O/IAP I, E2/IAP II and 32O/IAP II. Two tail DNA samples from different 102A positive mice were tested for each primer pair. Products were separated by electrophoresis through 0.8% agarose, then Southern blotted and hybridised with a 5kb MS32 repeat unit probe, in a phosphate/SDS hybridisation mixture at 65°C overnight. In addition to the full length product extending from the transgene specific primer to the IAP specific primer, 12-13 other products at 234bp intervals were also produced. These correspond to the 13 gamma satellite repeat units between the truncated 102A 3' end and the mouse IAP element (figure 6.11). These products may be PCR artefacts or represent major instability of the mouse gamma satellite DNA in this region. No additional bands corresponding to an MS32 repeat unit 29bp periodicity were observed.

B. **SP-PCR on sperm DNA from a transgene 102A positive mouse.** Amplifications were using the primer pair E2/IAP II for 25 cycles in PCR reaction tubes containing aliquots of Mbo I cleaved 102A positive sperm DNA, representing 20, 40 and 80 sperm molecules/ genomes worth (gw) of DNA per tube. Aliquots of the PCR products were separated by gel electrophoresis through 8% agarose, Southern blotted and hybridised with a 5kb MS32 repeat unit probe as in A. The major progenitor band is indicated (<prog>). Three presumptive ‘satellite’ mutation events are visible at 1.5kb (2x) and 1.7kb (+).

C. **The effect of input DNA concentration on PCR artefacts.** Amplifications from sperm and tail DNA from the same mouse (which was the same mouse DNA sample used in the SP-PCR in 'B.') were performed using the primer pair E2/IAP II with varying input concentrations of 102A positive DNA from 3ng (equivalent to an input of 500 sperm in the SP-PCR) to 24ng. For both tail and sperm DNA the ‘additional’ band pattern is concentration dependant indicating that it is most likely the result of a PCR artefact. A ‘hot start’ PCR was performed from 30ng of tail and sperm DNA to try to remove the additional bands from the profile, but this manipulation of the PCR appeared to have no effect in the artefacts.
Chapter 6, figure 6.6
equivalents of DNA were used in the small pool PCR (figure 6.6 B, data for 5 and 10 molecules is not shown). None of the PCR reactions showed this 'collapse' pattern. This is surprising because in a reaction tube containing the DNA equivalent of 80 sperm or tail DNA molecules, at least some abnormal length bands corresponding to the gamma satellite periodicity, whether bona fide mutation events or the result of PCR artefacts, would be expected if the instability in this DNA is as high as suggested by the results from 30ng of tail DNAs. To investigate this phenomenon further the IAP II and E1 primers were used to amplify from increasing amounts of input DNA from 3ng (equivalent to 500 single molecules) to 30ng to check if this pattern of PCR artefacts is related to the amount of input DNA (figure 6.6 C). Some additional bands became visible with a DNA input of 8-12ng, equivalent to 2600-4000 molecules of input DNA. Therefore it appears that these additional PCR products are a strange type of PCR artefact that is dependant on the concentration of the input target DNA molecules. A 'hot start' PCR was attempted in the hope that by keeping the two strands of DNA separate before primer annealing the levels of gamma satellite initiated artefacts would be lower (figure 6.6 C). This manipulation of the PCR did not seem to have any effect on the type of PCR artefacts seen. This suggested that the artefacts were not due to mispriming of the IAP based primers at each gamma satellite repeat unit, but due to some other process as yet not understood.

With the SP-PCR reactions using IAP II and E2 there were three apparent deletion mutations observed in sperm DNA (figure 6.6 B). However the deletions are large, >2kb and therefore must have involved the loss of some gamma satellite repeat units. Time was not available to recover these mutations and assay by MVR-PCR to see if the MS32 repeat unit pattern had altered. Therefore no mutation rate was measured for the 102A sperm DNA. No mutant bands of any kind were seen with the tail DNA (data not shown). Additionally had time been available it would have been interesting to further study the processes of mutation acting upon the gamma satellite repeat units between the 102A transgene insertion and the IAP LTR sequence.

### Homozygosity.

From a hemizygous × hemizygous cross, one third of all positive mice should be homozygous. For the single-copy transgenes 102A, 110C and 110D putative homozygotes were identified initially by the relative intensities of the hybridizing bands (the amount of DNA in each digest was carefully controlled by assaying each digest by fluorimetry before loading onto the gel). Putative homozygotes by dosage were then tested by breeding with non-transgenic mice; a true homozygote should give all positive offspring. For 102A, two homozygotes were identified by this method (F.Scaerou). For 110C the putative homozygote that was tested turned out to be a hemizygote. The other animals carrying this locus were quite old and subsequently we experienced difficulties in obtaining progeny from matings with these animals. Therefore during the course of this study no other putative 110C homozygotes were tested. The situation at the 110D locus was unusual. From separate hemizygous × hemizygous crosses a total of 17 transgene positive mice were identified. All 17 putative homozygotes were tested by breeding with negative animals then testing the progeny for any transgene negatives. One third of these 17 mice, 5-6 animals, would be expected to be homozygous for the 110D locus. In fact only one mouse was shown to be a possible homozygote with 8 out of 8 progeny testing positive for the 110D locus. There is a very low probability of obtaining such a result; the probability that at least 16 out of 17 progeny from hemizygous × hemizygous cross would be heterozygous is \((2/3)^{17} + (2/3)^{16} \times 1/3 \times 17 = 0.0096\), and this may reflect selection against homozygosity at this locus.
B. Mutation analysis of the multi-copy loci

Rates of mutation and stability of transmission. The transmission and mutation data for the six lines of transgenic loci are presented in table 6.1. For mutation analysis the most informative combination of restriction enzymes for each transgene locus was used. *Pst*I cuts the multi-copy 110A locus into five fragment sizes, four of which are single-copy (two flanking fragments and two internal fragments), and the fifth multi-copy, derived from the internal copies of the original allele (see figure 6.7 A). *Hind*III and *Ssp*I cut the multi-copy locus 109 into five DNA fragments, including one fragment consisting of two separate MS32 tandem repeat arrays (see figure 6.8 A). The mutation rates seen at the multi-copy loci were higher than those seen at the endogenous locus; for 109 the rate was 9.8% per offspring and for 110A the mutation rate was 5.0% per offspring.

As well as the apparent elevation in mutation rate, these events also show a number of interesting properties. In addition to presumptive germline mutation events which had been seen previously with the endogenous MS32 locus (Armour et al., 1989; Jeffreys et al., 1990; Jeffreys et al., 1991b), a number of embryonic mutation events were also seen. These were identified as having an additional band in the multi-copy transgene locus band pattern (figure 6.7 B, mutant 295 and 6.8 B, mutant 1594). Further breeding of these mutants showed that they were transmitted in addition to the progenitor locus, indicating that the events had occurred prior to germ cell differentiation in the early embryo. Germline mutations were identified as those mutants with an additional band in the multi-copy profile but which upon breeding only transmit the mutant form of the transgene. Sometimes germline mutations can be identified by a corresponding lost band from the multi-copy profile, unless the mutation is in one of the multi-copy 2.5kb bands. It remains possible that these presumptive germline events may have resulted from a very early embryonic mutation event, perhaps at the two-cell stage resulting in both germline and tail tissue DNA carrying the mutant form of the transgene, but other tissues carrying the progenitor form. However it is not possible to prove if this is the case for the presumptive multi-copy germline mutation events by pedigree analysis and for the purposes of this study these mutations are referred to as presumptive germline events.

The figures presented in table 6.1 give no indication of a bias in the number of mutation events seen in hemizygous x non-transgenic matings as compared to hemizygous x hemizygous matings for the 109 locus. However although there are 8 mutation events in hemizygous x non-transgenic matings for the 110A locus, there were none from hemizygous x hemizygous matings. Interestingly all of the 110A mutations identified by Southern blot analysis were from tail DNAs from mice bred in the Cambridge colony. No mutations were seen in mouse bred in the Leicester colony.

A third observation from the multi-copy transgene analysis was the recurrence of a particular type of mutation event at the 109 locus. This involved a change in a specific region of the flanking DNA between the second and third repeat unit arrays (from the 5' end) in the transgene locus, which resulted in the loss of a *Hind*III site. The lost *Hind*III site is identified by the gain of a larger 8.7kb band and the loss of the two smaller 6.2kb and 2.5kb bands in a *Hind*III and *Ssp*I digestion Southern blot profile (figure 6.9). This type of mutation event was classified as a 'missing site' or type II 109 mutation event and the frequency of its occurrence was very high: 50% of the germline 109 mutations identified and 30% of the embryonic mutations (table 6.2).
Figure 6.7 A. Schematic of the 110A progenitor and two mutant loci. The restriction sites shown are P(Pst I), H(Hind III) and R(Rsa I). (A - D) refer to fragments visible on the Southern blot autoradiograph. These bands are not of the common 2.5kb size. Blocks corresponding to changes in the mutant maps are shaded black, and estimates of size increases are included. The restriction mapping of the progenitor 110A locus and mutants was performed by P. Bois using a complex partial mapping system; DNA was cleaved with Rsa I which cuts 5' and 3' of the transgene, then a partial Pst I digest was performed. These partial products were cleaved by Mbo I. This enzyme only cuts at the 5' end of the transgene locus allowing identification of the order of the transgene blocks across the array and therefore the identification of which block in the array has undergone the expansion in mutants 292 and 295. The full 5' flanking mouse DNA extending to the Pst I site is not shown, but (/) represents a longer flanking sequence than shown.

B. Autoradiograph of a Southern blot of two 110A mutants. The 110A mutants were identified during my 110A tail DNA analyses. Mutant 295 carries an embryonic mutation event resulting in an expansion in the 8th block in the array (from the 5' end). Breeding from mouse 295 resulted in the transmission of both progenitor and mutant forms of the transgene locus 110A, indicating a true early embryonic event. Mutant 292 shows an additional band of 2.7 kb (an expansion of 200bp in the 5th block of the array). Progeny of mouse 292 carried only the 292 mutant form of the 110A locus. For this reason it was assumed to represent a germline mutation event, although it may represent a very early embryonic event. Mutant bands are indicated by arrows. This autoradiograph was kindly provided by P. Bois. This Southern blots were hybridised in a phosphate/SDS mixture with a 5kb MS32 repeat unit probe at 65°C overnight.
Chapter 6, figure 6.7
**Figure 6.8 A. Restriction maps of these 109 mutants:** The progenitor 109 locus was restriction mapped by A. Jeffreys and A. Collick. The restriction sites shown are S (Ssp I), H (Hind III) and P (Pst I). Bands (A-E) are the five different bands that are visible in the Hind III and Ssp I profile of the 109 locus.

**B. Autoradiograph of 109 progenitor locus and embryonic mutation 1594 cut with Hind III and Ssp I.** Bands relating to the restriction maps are indicated. The embryonic mutant band is shown by (*). The Southern blot was probed with a 5kb MS32 repeat unit probe in a phosphate/SDS hybridisation mix overnight at sixty five degrees. The actual position of the novel embryonic band in mutant 1594 is has not been mapped due to time constraints.

**C. Progeny of mutant 140** (initially identified by A. Collick) cut with Hind III and Ssp I. From the band patterns of 109 progenitor, 140 mutation (germline) and progeny of 140, mutation 140 can be shown to have lost bands (A + B) from the 5' end of the multi-copy array. Further restriction analysis showed that the 5' Ssp I site was still present, and therefore that the deletion probably did not extend to this position in the 5' mouse flanking DNA (A. Collick). The Southern blot was probed with a 5kb MS32 repeat unit probe in a phosphate/SDS hybridisation mix overnight at 65.
Figure 6.9 A. Restriction map of the 'missing site' or type II 109 mutation. This is a recurrent mutation first identified by the loss of a HindIII site and restriction mapped by A. Collick. This mutation occurred twice in the mice that I typed and 8 times in total across the 109 transgene pedigree. The loss of the Hind III site results in the loss of bands (B + C) to give one larger (8.7kb) band.

B. Autoradiograph of a Southern blot showing an example of a 'missing site' mutation. Mutant mouse 487 is the result of a germline 'missing site' mutation. The mutant was passed to the 487 x progenitor 109 locus progeny, mouse 552. The 487 mutation is a germline event, but this 109 'type II' mutation has also occurred embryonically (table 6.2). The Southern blot was probed with a 5kb MS32 repeat unit probe in a phosphate/SDS hybridisation mix overnight at sixty five degrees.
Type I signifies mutations most likely to have involved changes in the MS32 repeat units or possibly in the 5' or 3' flanking DNA. Type II mutations are events that occur in a small area of the flanking DNA between two repeat unit blocks and result in the loss of a Hind III restriction site without an apparent change in repeat unit copy number.

Heterozygotes for the multi-copy loci were easily identified by breeding two hemizygous parents carrying different mutations of the 109 or 110A locus. The banding pattern for Southern blot analysis of these homozygotes could therefore be predicted. A number of heterozygous animals were identified for both the 109 and 110A loci (e.g. mouse 1145 in figure 6.10). Another type of mutation event observed at the 109 locus was a complete deletion of the full transgene construct array. This type of event was first identified when breeding mice heterozygous for two different 109 mutations with another heterozygote or a hemizygote animal. All the progeny of such a cross should carry at least one mutant 109 allele. In figure 6.10 a 109 (709/Type I) heterozygote was mated with a 109 hemizygote (745 -type) which incidentally also carried 110D locus (from attempts to breed all the transgene loci into one mouse line), and eight of the progeny mice are shown. Progeny 'c' to 'h' show different combinations of the parental alleles but progeny 'a' and 'b' show full deletions of the 109 alleles ('b' does carry the 110D locus, but shows no evidence of a 109 locus).

MVR-PCR analysis of mutations identified by pedigree analysis. During this study it was assumed that mutations identified by pedigree analysis from changes in the fragment sizes of bands in the multi-copy loci profile were due to changes in MS32 repeat unit copy number (with the exception of the missing Hind III site mutation event at locus 109). To prove that MS32 repeat units within multi-copy loci can undergo changes, three transgene 110A presumptive germline mutation events were recovered and analysed by MVR-PCR. The 110A mutants 292 and 239 revealed from my Southern blot analyses as an additional 2.7kb band in the 110A fragment profile (figure 6.7), also gave additional bands when analysed by PCR. Amplifications were using MS32 flanking DNA primer pairs 32D/A and 32O/E. In addition to the regular bands amplified from the progenitor molecules, an additional 200bp larger hybridizing product was produced (figure 6.11). An aliquot of the PCR used to make this Southern blot was electrophoresed under the same conditions, to enable a series of agarose fractions potentially containing the mutant PCR product to be electroeluted onto dialysis membrane. Electroeluted products were ethanol precipitated then an aliquot of the recovered product was checked by Southern blot analysis to identify the appropriate electroelution fraction containing the mutant band and to assay for any contamination by progenitor allele (data not shown). The appropriate electroelution fractions were analysed by 2-state MVR-PCR (Jeffreys et al., 1991b; Chapter 3). The mutant MVR-PCR maps are shown in figure 6.11 B + C. Mice 292 and 239 share the same mutation which is a seven repeat unit duplication.

An additional germline mutation P30 (figure 6.12) was also analysed by 2-state MVR-PCR directly on total genomic DNA, using a flanking primer rooted in the 5' flanking mouse DNA to map the first block of repeat units in the multi-copy transgene. This is the block that appeared to be affected by the P30 mutation. Surprisingly the
Figure 6.10. Full deletion of the 109 locus.

A. Schematics of the 109 progenitor locus and three mutant loci. The mutant loci shown are the 109 type II mutant, which results in the loss of a Hind III site, and the 109 deletion mutations 745, which is 5' deleted and 709 which is 3' deleted. The 745 and 709 mutations were first identified during pedigree analysis by A. Collick. The restriction sites shown are Ssp I (S), Hind III (H) and Pst I (P).

B. Full deletion of the 109 locus in two out of eight progeny. Autoradiograph of Southern blot pedigree analysis of progeny from a hemizygous 109 positive animal carrying the 745 mutant locus and the transgene 110D locus (mouse 1079) crossed with a 109 heterozygote mouse, carrying a 709 mutant 109 locus and a type II 109 mutant locus (mouse 1145). Tail DNA digests were separated through 0.8% agarose, Southern blotted and hybridised with a 5kb MS32 repeat unit probe. Six of the eight progeny from this cross, progeny 'c' to 'h' show various combinations of the mutant 109 alleles, as indicated in the table beneath the autoradiograph, (+) = positive for 109 or 110D transgene allele, (-) = negative for the transgene.

However the progeny 'a' and 'b' do not show any 109 locus, indicating that a full deletion of one or more of the 109 mutant alleles has occurred in these progeny from the cross. Mouse 'b' is transgene 110D positive.
Figure 6.11 A. Autoradiograph of Southern blot analysis of PCR amplified 110A mutants. 30ng of total genomic tail DNA from the mutant mice 292 and 239 was amplified for 18 cycles in a 15 μl reaction with the MS32 flanking DNA primer pairs 32D with 32A and 320 with 32E. PCR reactions were performed as described in chapter 2. 5 μl of the reaction was electrophoresed on 0.8% agarose, then Southern blotted and hybridised with a 5kb MS32 repeat unit probe at 65° overnight in a phosphate/SDS hybridisation mixture.

Most of the MS32 repeat unit arrays in the multi-copy transgene 110A are 2.5kb in length (E1 to E2, figure 6.1). These amplify to give a constant non-mutant band of 2.18kb with 320/32E and of 2.49kb with 32A and 32D. The seventh block of MS32 repeat units from the 5' end of the multi-copy transgene 110A is reduced in size to 2.37kb (E1 to E2, figure 6.1). Amplification from this repeat unit block results in a band of 2.01kb with both mouse 292 and 239, using the primer pair 320/32E. However, with the primer pair 32A and 32D a 2.3kb band is only seen for mouse 292 and not for mouse 239. This is most probably due to a mutation event destroying one or other of these two primer sites (though not the nested 320 and 32E sites) in the flanking DNA of this repeat unit array in mouse 239. During the course of this study no time was available to investigate the nature of this mutation event. The mutant bands of 2.7kb from the 32A with 32D amplification were recovered by electroelution onto dialysis membrane and analysed by 2-state MVR-PCR.

B. Autoradiograph of two state MVR-PCR maps of the 110A progenitor and mutants 292 and 239. Reverse 2-state MVR-PCR was performed using 32E as a flanking primer and 32-TAG-3A and 32-TAG-3T as repeat unit specific primers (for sequences see table 2.1) for 18 cycles from the electroeluted DNA fraction. Close to the start of the MVR-PCR profiles shown here there is a duplication of seven repeat units. The position of this duplication is indicated. MVR-PCR using 32E gives the variant repeat unit pattern from the 3' end of the MS32 allele. Therefore the seven repeat unit change is actually close to the 3' end of the MS32 repeat unit block array.

C. MVR-PCR maps of the 110A mutants. The 2 state MVR-PCR maps of the 292 and 239 mutants as shown in (B). The MS32 variant repeat unit changes are shown to be the same for both mutant mice.
Chapter 6, Figure 6.11

A

32 O/32 E  32 D/32 A

239  292  239  292

3.0  3.0

2.5  2.7

2.37  2.5  2.49

2.18  2.3

2.01 - 2.0

B

Progenitor
Mutant 292
Mutant 239

7 repeat unit duplication

C

Progenitor  
Mutant 292  
Mutant 239
Figure 6.12  A. Autoradiograph of a Southern blot of germline 110A mutant P30.
This mutant was identified during Southern blot pedigree tail DNA analysis by P. Bois and this autoradiograph is kindly provided by P. Bois. Tail DNA was digested with Pst I and electrophoresed on 0.8% agarose before Southern blotting and hybridisation with a 5kb MS32 repeat unit probe in a phosphate/SDS hybridisation mixture at 65° overnight. The P30 mutant showed an expansion of the 5’ flanking band in the digest (band A in figure 6.2) from the 8.7kb size in the progenitor to 9.45kb in the mutant, a gain of ~750bp.

B. Schematic of the mutant P30 110A locus. The restriction sites shown are Hind III (H), Pst I (P), and Rsa I (Rs). An expansion in the first block of MS32 repeat units in the multi-copy array (black shading) is indicated to account for the P30 mutation. The full length of the 5’ mouse flanking DNA to the Pst I site is not shown, but (//) represents a longer DNA stretch than shown in the diagram.

C. Autoradiograph of a four state MVR-PCR of the P30 mutant. To elucidate the structure of the P30 mutation, MVR-PCR was performed on 30ng of genomic tail DNA using a primer rooted in the 5’ mouse flanking DNA 600bp upstream of the first transgene block in the multi-copy array, primer 110A-1 (for sequence see table 2.1). Primer 110A-1 was used as the MVR-PCR flanking primer in conjunction with the four state forward MVR-PCR primers, 32-TAG-CA (identifies E-type repeat units), 32-TAG-CG (e-type repeats), 32-TAG-TA (Y-type repeats) and 32-TAG-TA (y-type repeats). Amplifications were for 20 cycles of 96° 1 min, 70° 4 min. The PCR products were electrophoresed on 1.2% agarose, Southern blotted and hybridised with a 5kb MS32 repeat unit probe in a phosphate/SDS hybridisation mixture at 65° overnight.

D. MVR-PCR variant repeat maps for P30 and 110A progenitor loci. There was no change between the MVR-PCR maps from progenitor 110A tail DNA and mutant P30 tail DNA. This would suggest that the expansion in size of this restriction block is an event in the 5’ flanking DNA.
results showed no change from the progenitor MVR map for the first block of repeat units. This suggested that the size gain in this band, estimated to be 0.75kb, may be the result of changes in the 5' flanking DNA. More restriction analysis is required to investigate this result but the time was not available in this study.

**SP-PCR for the multi-copy loci 109 and 110A.** Due to the difficulty of identifying bona fide germline mutation events from pedigree analysis, SP-PCR was performed on sperm DNA from transgene 109 positive and transgene 110A positive mice. Mutations identified directly from germline DNA would be bona fide germline events. The mutation rates observed by SP-PCR at the multi-copy loci could then be compared to those at the single-copy loci 110C and 110D. The high rates of presumptive germline mutation at the multi-copy loci would suggest that an equivalent high rate of locus mutation would be observed using SP-PCR. Initially the SP-PCR strategy used for each of these transgenes utilised the MS32 transgenic flanking DNA primer sites E1 with E2. Therefore for each locus there were a number of target sites for amplification since almost every individual original construct in the array appeared from restriction mapping analysis to possess primer site E2 (cut by Pst I) and therefore most probably the adjacent 32D primer site and to possess primer site 32A (cut by Hind III) and the adjacent E1 site. The autoradiographs of Southern blots of these experiments showed an extremely complex banding pattern (figure 6.13 A for transgene line 110A, the pattern at the 109 locus was of the same type, data not shown) with bands at 29bp intervals (one MS32 repeat unit) above and below the progenitor bands. These bands were most likely PCR artefacts, rather than reflecting a huge instability at this locus, greater than that observed by pedigree analysis. To examine this possibility the annealing temperature of the PCR was increased from 68° to 70° and 72° in an attempt to reduce the level of non-specific products. These PCR products gave the same type of banding pattern (data not shown). In addition reactions were performed with the nested primers 32A and 32D and with E1 only and E2 only, to assess if the E1 and E2 linkers were the cause of this result. Amplifications with the primer pair 32A and 32D gave the same type of pattern, El and E2 alone gave no products (data not shown).

Since this pattern did not appear when these primers were used at the single copy locus 110D, it was decided to digest the multi-copy constructs with enzymes that would separate the individual constructs in the array using Hind III and Pst I, which cut at either end of the initial construct and Mbo I which will cleave in the mouse DNA flanking the insert. SP-PCR on this digested DNA gave an autoradiograph pattern with a reduced level of additional bands below the progenitor signal (figure 6.13 B) for both transgene 109 and 110A (data only shown for transgene 110A). No additional bands were observed above the progenitor signal. However, it was still not possible to measure directly the mutation rate at these loci since there was no way of distinguishing if the deletion events seen in figure 6.13 B, were bona fide mutation events or PCR artefacts from the small number of genomic DNA molecules that had not been fully digested. At this level of analysis a single uncut full locus molecule may give a mutant artefact pattern.

Small pool PCR was attempted on sperm DNA from a 109 mutant 709/745 type (figure 6.10) heterozygote mouse, using sperm DNA digested with Hind III and Pst I and Mbo I. Although this DNA was digested the level of artefact/mutant bands seen below the main progenitor alleles were greater than the number seen with the hemizygote progenitor 109 allele (data not shown). However this may be a reflection of an increased number of artefacts from two loci rather than just one locus in the hemizygous mouse. Because of the problem of incomplete
A. SP-PCR of a transgene 110A hemizygote. Amplifications were with the transgene specific primers E1 and E2 on 110A DNA cut with Mbo I which cleaves the 110A locus into two halves; in addition to flanking Mbo I sites a novel Mbo I site exists within the 9th block (from the 5' end) of MS32 repeat units. Multiple aliquots of digested DNA were amplified in 7μl PCR reactions containing concentrations of 110A sperm DNA equivalent to 40, 80 and 160 individual sperm molecules for 96° 45s, 68° 1 min, 70° 3 min for 25 cycles. The products were separated by agarose gel electrophoresis (1% agarose concentration), Southern blotting then hybridisation with a 5kb MS32 repeat unit probe in a phosphate/SDS hybridisation mixture at 65° overnight. The main progenitor 110A locus is indicated (P). This represents an amplification from the multi-copy 2.5kb bands in the 110A array (figure 6.1). Interestingly no products are produced from the 2.37 and 2.67kb MS32 repeat unit arrays, suggesting that the E1/E2 primer sites in the DNA flanking these repeat unit arrays is changed. Beneath the progenitor band there are numerous additional bands accounting for 64 of the MS32 repeat units. The collapse pattern may extend further but these products were not kept within the boundaries of the Southern blot.

B. SP-PCR on 110A DNA cut with Hind III, Mbo I and Pst I. Amplifications and subsequent procedures were as described in 'A.' The level of additional bands is lower, but there are still some bands of abnormal length (all deletions) for which it is not possible to assess if they are bona fide mutation events.

C. SpPCR on transgene 109 heterozygous DNA cut with Mbo I. Amplifications and subsequent procedures were as described in 'A.' The transgene 109 DNA used was from a heterozygous mouse carrying the two deletion mutations 709 and 745 (figure 6.5). Three progenitor bands are seen, 1 = the amplification of the 6.2kb band containing the 3rd and 4th MS32 repeat unit arrays in the construct. The transgene flanking DNA between these two blocks of repeat units has changed significantly to not allow the amplification of the E1 to E2 product across the 3.7kb array. 2 = the E1 to E2 product from the 3.9kb array, and 3 = the 2.5kb E1 to E2 product. The numbers of additional bands from the Mbo I cut DNA is less than the number seen with the 110A locus in 'A.' However this is likely to be due to the fact that there are more target repeat unit array molecules in the 110A construct to produce additional mutant bands.
digestion of the mouse DNA for analysis it is not possible to measure the rate of mutation directly for the hemizygous versus homozygous 109 mice.

To ascertain if this pattern of additional bands for the 109 and 110A loci was male germline specific, DNA was prepared from the liver and brain tissue of the mice used in the sperm DNA analysis. The same type of banding pattern was observed in these somatic tissues (data not shown) and therefore no somatic mutation rate could be measured by SP-PCR.

Discussion

A. Single-copy loci

Mutation frequencies by pedigree analysis. In this chapter it has been shown that some MS32 transgenes can be extremely unstable in mouse pedigrees. The main distinction between the mutation frequencies at the different transgenic MS32 loci studied was between mutation rates at the single-copy and multi-copy loci. So far no mutation events resulting in allele length changes have been identified by Southern blot pedigree analysis for the single-copy transgene loci, 102A, 102B, 110C and 110D. The mutation rate for the single-copy loci is < 0.8%. The mutation rate observed by Southern blot analysis for the endogenous human MS32 locus (Jeffreys et al., 1991b). Therefore it is not possible to say from pedigree analysis if the mutation rate at these four single-copy loci is greatly reduced from that seen at the endogenous locus.

Mutation frequencies measured by SP-PCR analysis. By using the SP-PCR technique we were able to assay many thousands of individual sperm for evidence of mutation events in order to attempt to define the male germline mutation rate of the 110D and 110C loci more accurately. SP-PCR was sensitive enough to identify mutation events in mouse sperm DNA at the hemizygous 110D and hemizygous 110C loci. Five presumptive mutation events (four MS32 repeat unit gains and one deletion) were recovered from transgene 110D sperm DNA corresponding to ~68,000 sperm, giving a mutation rate of 0.007% per sperm. Four presumptive gain mutation events were observed from transgene 110C sperm DNA in 9520 sperm molecules giving a mutation rate of 0.04% per sperm. Absolute mutation rates were determined by correcting for a PCR efficiency of 80% as established by Poisson analysis of single-molecule dilutions of sperm DNA as described in Jeffreys et al. (1994). The presumptive rates of mutation at these single-copy loci were reduced from that observed by small pool analysis of the endogenous MS32 locus in sperm DNA (Jeffreys et al., 1994) of 0.81% per sperm. These results are concordant with the fact that no mutation events were seen by pedigree analysis for the 110D and 110C loci.

There are a number of reasons why the mutational behaviour of these single-copy MS32 transgenes may differ from that seen at the endogenous locus. The first point to consider is that the mutational rates for the 110C and 110D locus are for hemizygous transgene loci, since no mice were identified homozygous for either of these loci during the course of this study. From SP-PCR analysis of the endogenous human locus in sperm DNA, interallelic processes have been identified as being the most common (~50%) mutational events occurring, although intrallelic processes may account for 25-30% of mutation events (Jeffreys et al., 1994). Obviously interallelic processes cannot occur in transgenic animals carrying only one allele and this limits the scope of mutation processes able to
act upon the transgenic MS32 loci to intraallelic events such as unequal sister chromatid exchanges and replication slippage. However, the 102A locus was examined by SP-PCR in sperm and tail DNA from a mouse shown to be homozygous for the 102A locus by breeding analysis (chapter 5). The only mutations seen were three deletions which appeared to be of gamma satellite repeat units. Therefore for the small number of molecules analysed (~6500 for tail and 6500 for sperm) no MS32 repeat unit changes were identified. Although DNA from a hemizygous 102A mouse was not analysed by SP-PCR it is not likely that the result from a mouse of this type would have shown a different mutation rate. Therefore it is not likely that the hemizygous nature of the transgenes is the only factor contributing to their reduced mutation rate.

It is possible that the mouse flanking DNA surrounding the transgene insertions may exert an effect on mutation rates. 102A and HOD are flanked by mouse gamma satellite which may have a repressing effect on mutation processes. The mutation rate for the single-copy 110C locus, flanked by single-copy DNA, appears to be five times the rate at the 110D locus. However if gamma satellite does have a negative effect on single-copy transgenes then this effect does not extend to the multi-copy locus 109 which is flanked by mouse gamma satellite and has a high mutation rate.

The 5' polarity observed with mutation events at the endogenous locus and the allele length independence of mutation (Jeffreys et al., 1994) suggested the involvement of cis-acting elements modulating mutation processes at the MS32 repeat unit array. Recently some MS32 alleles have been identified which show reduced variability in human populations and are associated with a 'G' to 'C' transversion in the 5' MS32 flanking DNA (Monckton et al., 1994). This G-C transversion is positioned 48 bp upstream of the tandem repeat array and 16 bp upstream of a diverged MS32 repeat which precedes the array (figure 3.6, chapter 3) SP-PCR analysis of single sperm demonstrated a significant reduction in MS32 mutation rate at these alleles. In addition the few mutation events recovered from these alleles did not exhibit the properties of polarity and frequent interallelic events observed with MS32 alleles mutating within the normal range (Monckton et al., 1994). The existence of this mutation rate polymorphism lends weight to the proposal for the action of elements in the flanking DNA near the minisatellite repeat unit array, in the regulation of tandem repeat instability. Such elements may possibly be required for the initiation of mutation. Analysis of this polymorphic site in transgenes that carry 5' flanking MS32 DNA, (102A, 110C and 110D) have shown that these transgenes have maintained the 'G' variant that was present in the 5' flanking DNA of the MS32 allele used to make the transgene construct for microinjection (data not shown). Therefore the lower rates of MS32 repeat unit mutation at the single-copy MS32 transgene loci compared to the rate at the endogenous human locus cannot simply be explained by the presence of a 'C' variant, associated with a reduced MS32 mutation rate, at this flanking polymorphic site. However, it is possible that a greater length of MS32 5' or 3' flanking DNA than the 200bp present in the original construct for microinjection (figure 4.1, chapter 4) and maintained in some transgene insertion events, is required to operate this regulation of MS32 mutation processes. In addition the mechanism of this cis-acting influence on MS32 mutation is not understood. It may be that elements which interact with the MS32 flanking DNA to modulate this process, for example DNA binding proteins, exist and that these elements are not conserved sufficiently in the mouse genome to perform the same function.
Recessive lethal phenotype of transgene HOD. Breeding of hemizygous x hemizygous 110D mice did not produce a homozygous animal for mutation rate comparisons. Statistically the failure to obtain a homozygote is highly significant. However, transgene 110D mouse flanking sequence analysis provided a clue as to why it is not straightforward to breed a homozygous 110D mouse. The 110D locus is flanked 5' by mouse gamma satellite repeat units and 3' by single-copy DNA (chapter 5) and therefore may have integrated at the junction between gamma satellite DNA and non-repetitive DNA. It is possible therefore that a deletion event at this junction may have occurred upon 110D locus integration, resulting in a recessive homozygous lethal deficiency. If these putative deleted genes are important in development it might be possible to identify at what stage, by observing the embryos of hemizygous x hemizygous 110D matings at different post-zygotic stages. This might identify at what stage these embryos were dying if the 110D locus is causing homozygous lethality.

MVR-PCR of mutations recovered from SP-PCR. Three transgene 110D mutants and one transgene 110C mutant were recovered from SP-PCR analyses in this study and analysed by MVR-PCR. These numbers are too small to make any generalisations about the mutation processes operating at these loci but it is useful to compare them with the type of event seen at the endogenous locus (Jeffreys et al., 1994). Structural analysis of mutant alleles identified in human pedigrees or recovered from single sperm has shown a complex pattern of tandem repeat instability. Mutations at the endogenous human locus frequently involve the transfer of MS32 repeat units from a 'donor' to a 'recipient' allele, often accompanied by complex rearrangements of these repeat unit blocks. No flanking marker exchanges have been observed during the analysis of these mutation events. This indicates that unequal exchange is not a common mutation mechanism involved and that a gene conversion process is the more likely dominant mode of mutation at the MS32 locus.

The three presumptive mutations recovered and MVR-mapped from the 110D locus, show no indication of mutational polarity. One (mutation 1) shows an 8 repeat unit duplication combined with a 4 repeat unit switch from 'EEEe' to 'YeeE' directly after the duplication event. There is no 'donor' site for this block of four repeat units in the rest of the allele and it may represent a gene conversion event similar to small patches of conversion seen with mutants at the endogenous locus (Jeffreys et al., 1994). Another mutant (mutation 2) is a deletion event which also shows switching of repeat unit type in one position, indicating that it is not likely to be a PCR artefact. The third recovered mutant (mutation 3) shows a large 22 repeat unit duplication at the 3' end of the allele. This type of duplication event is probably the result of a rare slippage event or unequal sister chromatid exchange at the 110D locus, although it remains possible that it is the result of a PCR artefact duplication.

Due to time constraints only one mutation was recovered from the 110C locus, though four gain mutants were recorded. However, the mutant recovered shows an interesting change to the 110C variant repeat unit array pattern. This mutant was recovered from analysis of the 3' most block of repeats in the 110C locus, the block downstream of the 5' flanking human DNA (figure 6.4 A). The 5' end of the mutant allele is almost identical to the progenitor allele with the exception of 2 repeat unit type switches, 'ee' to 'Ye' at positions 18 and 19 in the mutant array. However, at repeat unit number 30 there appears to have been a 15 repeat unit insertion (figure 6.4 C). The only possible donor site for this mutant is from the 5' block of repeat units. This indicates that there may have been some 'inter-array' interaction to produce this mutation event (figure 6.4 D). Additionally if the 15 repeat units...
donated from the 5' block are aligned with the mutant array the two repeat unit type switches at positions 18 and 19 could be aligned within one repeat unit, of a ‘yE’ motif in the 5’ repeat unit array. This indicates that these repeat unit switches may also have been ‘donated’ from the 5’ MS32 repeat unit block, possibly at the same time as the downstream 15 repeat unit donation (figure 6.4 D). The 15 repeat units which appear to have originated from the 5’ MS32 repeat array are followed by 8 repeat units of unknown origin. The ‘EyyEyEyE’ motif does not appear anywhere in the original construct or in the 110C variant repeat arrays and may potentially be the result of a complex conversion event.

B. Multi-copy loci

Mutation frequencies by pedigree analysis. In contrast to the single-copy loci for which no mutations were detected during pedigree analysis, the multi-copy loci mutation rates were high enough to be directly measured by pedigree analysis. The total mutation rates at the multi-copy loci are 9.8% (109) and 5.0% (110A) respectively. Of these total mutation events the presumed germline mutation rates at these loci (which may possibly include very early embryonic mutation events) are 4.4% for 109 and 3.1% for 110A. The germline mutation rate measured by pedigree analysis at the endogenous human locus is ~1% (Jeffreys et al., 1991b). As well as representing a higher germline mutation rate than at the endogenous locus the total multi-copy MS32 transgene mutation rates include examples of unexpected forms of MS32 instability.

Early embryonic mutation at the multi-copy loci. At both multi-copy loci there are examples of mutations occurring in the early embryo (figure 6.7, mutant 292 and figure 6.8, mutant 1594). The mutation events that appeared to be embryonic (identified by additional bands in the Southern blot profile) must have occurred early in the embryo, before the separation of the germline from the soma (day 5-6, reviewed by A. MacLaren, 1991), since germline transmission of both standard and mutant forms of the loci was observed when breeding from these mice (A.Collins et al., in press). Three distinct forms of mutation were observed to occur during this early window of embryonic mutation; firstly for both loci 110A and 109, an apparent loss or gain of MS32 repeat units from one or other construct blocks in the array, secondly, for locus 109 the loss of a specific Hind III site (figure 6.9) and thirdly the complete deletion of the 109 locus (figure 6.10).

Previously somatic mutation events at the MS32 locus have only been observed by Southern blot length analysis of some lymphoblastoid cell line DNAs (Armour et al., 1989b) and by single molecule amplification of size fractionated (for deletion events) blood DNA (Jeffreys et al., 1990). At microsatellite loci a high proportion (up to 40%) of presumptive germline mutations detected in lymphoblastoid cell line DNAs were found to be somatic in origin or the result of mutations during culturing of the cell lines (Weber et al., 1993; Banchs et al., 1994). For this reason the mutations observed in the Armour et al. (1989b) study may represent minisatellite mutation events that occurred in the cell line rather than bona fide somatic mutants. Likewise the events recovered from the Jeffreys et al (1990) study are atypical minisatellite deletion events that occur at a low frequency (0.07%) and show low levels of mosaicism. The only known example of mutational somatic mosaicism at a hypervariable human minisatellite locus is an event occurring early enough in development to be detectable as a third allele by Southern blot analysis of blood DNA probed with an MS31 locus probe (D. Neil, 1994). However this event was only one mutation in 3088 people tested giving a somatic mutation per locus per person of 0.008% (D.Neil, 1994).
level reported in this study accords well with calculations that estimated that the rate of somatic mutation at minisatellite loci would be very low (<10^{-5} per mitosis for MS31 and <4 x 10^{-6} per mitosis for MS32, Armour et al., 1989b). Therefore somatic mutation events can be assumed to be infrequent at minisatellite loci.

In contrast the rates of somatic mutation for different types of mutation events at the multi-copy transgenic MS32 loci are high. Correspondingly high somatic mutation rates have been observed with the endogenous mouse minisatellites Ms6-hm (Kelly et al., 1991) and Hm-2 (Gibbs et al., 1993). Ms6-hm has a high germline mutation rate of 2.5% and a high somatic mutation frequency resulting in 2.8% of mice showing somatic mosaicism by Southern blot analysis (Kelly et al., 1989, 1991). Germline mutations at Ms6-hm involve small length changes (<200 repeat units, 'GGGCA') with no significant bias for gains or losses of repeat units. However, there is both a deletion bias and a paternal origin of mutation bias in the somatic mutations at the locus, which in general involve changes of large blocks of repeat units (~500), suggesting that paternally and maternally derived Ms6-hm alleles may not be equally prone to mutation in the early embryo. There is no evidence of a parental sex bias in the embryonic or germline mutation events at the multi-copy loci 110A and 109 (A.Collick, personal communication). Hm-2 (repeat, GGCA) has a germline mutation rate of 3.6% and an extremely high somatic mutation frequency resulting in 20% of adult progeny being mosaic for a mutant Hm-2 allele (Gibbs et al., 1993). It is possible that the atypical somatic mutation behaviour of the multi-copy MS32 transgenes is due in part to influences of the mouse replication and recombination processes acting upon the integrated MS32 DNA. The fact that the 5' flanking DNA of these two mouse multi-copy minisatellites is different (109 is flanked by gamma satellite DNA, 110A is flanked by single-copy DNA, chapter 5) suggests that this is not a flanking sequence dependent process. The lack of observed somatic mutations by pedigree analysis of the single-copy loci may be a reflection of a generally lower mutation process at these loci.

Recently other tandemly repeated DNA sequences have been reported to be unstable during the early embryonic window of development. For example the CGG triplet repeat array in the 5' untranslated region of the FMR1 gene is subject to extreme germline and somatic instability leading to expansion of the CGG repeat array and resulting in the Fragile X mental retardation phenotype (FRAXA) (Kremer et al., 1991; Fu et al., 1991). In addition the Myotonic Dystrophy and FRAXE triplet repeat arrays show high levels of somatic instability (Reviewed by Kuhl and Caskey 1993 and Nelson 1993), detected as mosaicism in Southern blots of genomic DNA isolated from the blood and other tissues of single individuals. The sequences of the mouse minisatellites Ms6-hm and Hm-2 are a GGGCA pentamer repeat and a GGCA tetramer repeat respectively. This sequence shows similarity to the triplets CAG (in FRAXA and DM) and CGG (in FRAXE) which have been shown to be embryonically unstable. Therefore it is possible that the mechanisms proposed to result in instability at the triplet repeat loci (see chapters 8 and 9) also affect the mouse minisatellites. It is however difficult to see the relationship of these mechanisms to the large (12-20kb) multi-copy MS32 transgene insertions where the repeat unit length is 29bp.

Specific loss of a Hind III site at the 109 locus. The detection of a recurring mutation in different parts of the 109 mouse pedigree that did not appear to involve changes in the number of MS32 repeat units, indicated that an unusual type of mutation process was occurring. This mutant event results in the disappearance of a Hind III site between the second and third blocks of MS32 repeat units in the 109 locus array (figure 6.9). Currently the structural basis for this mutation is being analysed and the region of change in the MS32 flanking DNA is being
sequenced (A. Collick, personal communication). Interestingly many of the other mutations that appear to affect repeat unit copy number have changes localised to second and third repeat unit blocks around the flanking DNA junction where the missing site mutation can occur (Collick et al., in press), indicating that there may be structural properties of the multi-copy array at this location that promote both kinds of mutation process. This view is supported by the fact that the human minisatellite flanking DNA in the region between the second and third repeat unit blocks in the construct array has been structurally arranged. Restriction mapping shows that the Pst I site in this flanking DNA has been lost as have the sites for some of the flanking DNA primers (Collick et al., in press). Alternatively the localisation of a number of mutations in this region of the multi-copy locus may reflect a non-random mutation process at the locus by virtue of its multi-copy structure or possibly may result from the orientation of the 109 locus on the chromosome. Recent data has shown that the 9.8% mutation rate of the 109 locus is reduced in mice bred from animals carrying a 109 locus with the missing site mutation (A. Collick, personal communication). Perhaps this is an indication that the missing site form of the locus is more stable than the initial form and represents a stable version of the transgene 109 locus. More work on this interesting phenomenon will help to answer these questions.

Full deletion of the 109 locus. A third unusual mutation phenomenon seen with the multi-copy transgene locus 109 is the full deletion of the locus. There was some evidence for full deletion of the locus from my results (figure 6.10), although the extent of the deletions were not assessed over the full hemizygous x hemizygous pedigree. In the course of breeding further 109 type mice for mutation analysis, an additional 131 progeny of hemizygous x wildtype crosses were examined by A. Collick and P. Bois. A pattern of transmission distortion not obvious from the data presented in table 6.1 was observed, with only 68/205 progeny from hemizygous x hemizygous matings carrying a 109 locus. Half of these progeny would be expected to be transgene positive so this reflected a significant distortion from Mendelian segregation. Close analysis revealed that individual 109 line animals showed variability in their transmission of the locus, some with rates around 50% and others with substantially lower rates (e.g. 12.5%, Collick et al., in press). This type of effect was not seen for the single-copy locus 110D which was studied in an analogous manner. Furthermore this study revealed low levels of otherwise normal 109 transgene in some mice (Collick et al., in press) The dosage was quantified by using the mouse LINE derived 1.3kb Eco R I repeat (Chen and Schildkraut 1980) as an internal control for the amount of tail DNA digest loaded per lane, with signals from both MS32 and Eco R I repeat probes being measured by phosphoimaging. Mice with a low dosage of the transgene locus additionally were shown to transmit the locus at a reduced rate (Collick et al., in press). The 109 locus has 5' gamma satellite flanking mouse DNA (chapter 5). Previously instability has been observed for two selectable markers, thymidine kinase (Wahl et al., 1984) and CAD (Burner and Lo 1986) integrated into centromeric regions of mouse chromosomes, where mouse gamma satellite DNA is localised (Vissel and Choo 1989; Hastie 1989). Therefore it may be possible that the instability of the 109 locus leading to its full deletion in the early embryo is a result of the gamma satellite DNA behaviour. Perhaps non-satellite DNA actively promotes its own deletion by specifically destabilising the region or perhaps it is just lost through normal processes of satellite DNA turnover for example unequal exchange crossovers between different blocks of satellite (Vissel and Choo 1989).

Mutation rates and hemizygosity. The mutation rates of the 109 locus did not appear to increase in heterozygous mice or in the progeny from hemizygous x hemizygous matings. This would suggest that interallelic
events as observed widely with the endogenous MS32 locus mutants are not as important in the mutation processes of the multi-copy MS32 transgenes or that when the interallelic pathway is blocked, as in hemizygous x wildtype crosses that intraallelic processes can result in mutation. With the 110A locus all of the mutations were observed in F1 and F2 mice bred in Cambridge. No mutations were seen in mice subsequently bred in Leicester. This is a strange phenomenon and may result from a number of factors.

The founder mice were bred with non-transgenic (C57BL/6J.CBA/6J) F1 animals to produce the first generation of progeny including the mutant 110A mice 239, 292 and 295 (figure 6.7 and figure 6.11). Subsequent generations in the Cambridge colony were bred with the same genetic background of mice, but in Leicester breeding was with C57BL/6J mice. Perhaps there were small strain specific differences influencing mutation rate at the 110A locus, although not with the other transgenic loci. To test this P.Bois tested 30 positive 110A mice bred in Leicester with (C57BL/6J.CBA/6J) F1 and observed no mutation events. With a rate of 8/70 mutations seen in the initial 110A mice from Cambridge it was expected to see at least one or two mutations in these mice, if this was a purely CBA/6J related phenomenon. However the numbers of mice tested were too small to draw any significant conclusions as to the effect of the specific strain on mutation at the 110A locus.

Alternatively this effect may be a result of transgene processing. For example there may have been an initial 'mutation burst' at the beginning of the breeding program but subsequent processing of the transgene construct inserted into the mouse genome may have caused the transgene to become stabilised over time. Although mutation events at the mouse minisatellite Ms-hm have been shown not to cluster there have been examples of minisatellite mutation 'bursts' in the germline of cattle (Georges et al, 1990). The proposal for a mutation 'burst' experienced by an MS32 transgene in the mouse could be tested by repeating the microinjections and performing small pool PCR experiments on the loci obtained. This would give a direct mutation rate both germline (from mouse sperm) and somatic (from e.g. tail DNA) for individual MS32 transgene alleles in different parts of the pedigree. The mechanism by which such a mutation 'burst' would be controlled and the proposed stabilisation of the transgene insertion is unknown but may be the result of epigenetic factors such as DNA methylation of the transgene or mouse flanking DNA.

MVR-PCR analysis of 110A mutants identified during pedigree analysis. Mutation events at the multi-copy loci were assumed to involve the loss or gain of MS32 repeat units by virtue of length changes in mutant bands observed using Southern blot analysis. MVR-PCR on DNA recovered from four embryonic 110A mutations (figure 6.11 B, C) confirmed that repeat unit changes had occurred for these mutation events. MVR-PCR of a fifth 110A (germline) mutation (P30, figures 6.7 and 6.12 ) did not indicate any changes in the variant repeat unit array of the mutant block, indicating that a 5' mouse flanking DNA change may have occurred. Both MVR-PCR maps of the mutants 239 and 292 show an identical change in variant repeat unit structure, a seven repeat unit duplication. Both these animals are F1 progeny of the founder mouse 110 bred with a CBA/6J non-transgenic mouse. Southern blot analysis of siblings of 239 and 292 revealed a third mutation event, apparently the same seven repeat unit gain, though this mutant has not been analysed by MVR-PCR. It is possible that these events may represent some kind of 'favoured mutation event' at the 110A locus, perhaps analogous to the missing site mutation at the 109 locus or it may be that these are not germline mutation events but represent germline mosaicism due to a premeiotic event in the founder mouse 110. Since both mutation events are from hemizygous animals it is likely that the event is the result of either a replication slippage event or an unequal sister chromatid
exchange. These are mechanisms that have been observed at the endogenous MS32 locus, however they do not represent examples of the most prevalent interallelic events seen at the endogenous human locus.

In order to assess if such MS32 mutation properties as polarity have been transferred to the multi-copy mutation events, a large number of germline mutation events must be recovered and studied. In the current study there was not enough time available to follow up such lines of enquiry. There is the possibility that some of the germline mutation events identified by pedigree analysis may be the result of very early embryonic mutant events. These are indistinguishable from germline events by standard Southern blot analysis. However SP-PCR analysis of sperm DNA from multi-copy transgene positive mice would enable the calculation of the genuine germline mutation rate at these loci. SP-PCR using transgene construct flanking primer pairs E1/E2 or 32A/32D results in separate products from each individual repeat unit array/individual transgene construct in the multi-copy locus. Therefore any mutants recovered as individual hybridizing bands would probably represent MS32 repeat unit changes in a single repeat unit block and could therefore be analysed by MVR-PCR. However it would not be possible to ascertain from which part of the multi-copy construct array the mutant MS32 repeat unit block arose.

Unfortunately in this study it was not possible to obtain a SP-PCR pattern with a consistently ‘clean’ background in the lanes (figure 6.13) in order to establish the genuine germline (or somatic, from lung and brain tissue) mutation rates at the multi-copy loci. Although the number of additional bands, the majority of which were presumed to represent PCR artefact products, was reduced when the multi-copy structure was cleaved by restriction digestion into individual transgene construct monomers, it was never possible to be certain that the remaining bands were bona-fide mutation events. However this result suggested that these PCR artefacts were produced as a result of the multi-copy structure of the transgene. The reason for the production of so many PCR artefacts in these SP-PCRs is not clear, but they may result from the presence of inverted transgene construct blocks within the multi-copy array. It is not possible to indicate the orientation of the individual transgene construct blocks in the multi-copy array by restriction digestion, but it is possible that the presence of this inverted repeat results in multiple PCR artefacts perhaps by the DNA forming unusual conformations between complementary regions during the PCR.

**Conclusions**

Mutation analysis of the multi-copy MS32 transgenic loci has shown that transgenes of a human minisatellite can show instability in the mouse. From the current evidence the rates and processes of minisatellite mutation at these transgenic loci differ from those seen with the endogenous locus. With the endogenous locus SP-PCR can overcome the limited supply of mutants available from pedigree analysis to provide an unlimited supply of de novo mutation events. However with the MS32 single-copy transgenic loci the low mutation rates limit the numbers of putative mutation events available for study so that the conclusions drawn from this work so far are only very tentative. The level of artefacts seen with the multi-copy loci SP-PCR made assessment of mutation rates and recovery of individual mutation events intractable, but with high pedigree mutation rates it would have been expected to see this reflected in a high SP-PCR mutation rate. However the unusual forms of minisatellite mutation revealed at the multi-copy loci provide additional systems for studying the complex mutation processes that can lead to structural changes in the repetitive DNA sequences of higher eukaryotes.
Chapter 7

The effects of gamma radiation on mutation processes at the MS32 transgene locus 110D

Summary

Assessment of the effect of ionising radiation on germ cells and the resulting frequency of germline mutations is of great importance in risk estimation of radiation damage in human DNA. Direct measurement of germline mutation in humans is limited and mice are the animal model of choice as an indicator of the sensitivity of DNA to radiation. Radiation induction of germline mutation events has been studied in mice by assessing the progeny of irradiated animals for induced recessive and dominant mutations. For this process extremely large numbers of mice need to be analysed to elucidate the base rate of spontaneous mutation for comparison with the radiation induced mutation rate. Some mammalian minisatellite loci have been shown to be highly unstable, enabling the determination of spontaneous mutation rate by pedigree analysis. In addition, endogenous mouse minisatellites have been shown to be sensitive to mutation induction by low doses of gamma radiation. In this study the aim was to use the small pool-PCR (SP-PCR) technique to analyse many thousands of individual sperm representing potential progeny directly from irradiated male mice. The locus chosen for study was the MS32 transgene locus 110D. This is a single-copy transgene with a putative male germline mutation rate of 0.007%. Mice carrying this locus were irradiated over a dose range from low exposure (0.1 Gy) to high exposure (3.0 Gy), with gamma radiation from a \(^{60}\)Co source. Sperm DNA was prepared after six weeks to recover sperm that had been irradiated at the spermatogonial stage. This DNA was analysed by SP-PCR to assess if an elevated rate of germline mutation at this locus had been induced by the gamma radiation. Some putative transgene deletion events were observed but no gain mutations were seen.

Introduction

It is extremely difficult to detect changes in germline mutation rate in human populations resulting from environmental influences such as exposure to ionising radiation or to mutagenic chemicals (UNSCEAR 1986). Therefore the estimation of the genetic hazards of mutagens to man is dependent upon the extrapolation of data from experimental animal systems. The main purpose of mammalian mutagenesis experimentation has been to elucidate the factors important in mutation induction in the germ cells of mammals, as a model for man. Between the initial DNA change and the ultimate recovery of an induced mutation in the succeeding generation there are a number of intervening steps that will influence the number of mutations seen. They include DNA repair processes and the survival frequency of mutant bearing germ cells, zygotes and embryos.

Mouse germ cell data following spermatogonial exposure to ionising irradiation has been proposed as the most relevant experimental data upon which to extrapolate the genetic risk imposed on offspring from human individuals who have been exposed to single high level doses of irradiation, for example children of survivors of the Hiroshima and Nagasaki atomic bombings. A number of different methods have been developed to allow the recovery and quantitation of induced mutation events in mouse germ cells. The specific-locus method was developed by Russell...
(1951) and provided an efficient method for the screening of transmitted germ cell mutation events. The method involves breeding homozygous wild-type treated (or untreated control) mice to untreated tester stock mice which are homozygous recessive at seven marker loci: a, non-agouti; b, brown; CCh, chinchilla; d, dilute; p, pink-eyed dilution; s, piebald and se, short ear. The first six loci control coat pigmentation and the seventh locus controls external ear length. F1 offspring are expected to be heterozygous at the marker loci and therefore to express the wild-type phenotype. However if a mutation has occurred at one of the tester loci in the treated mice to produce the mutant form of the gene then the F1 progeny will be homozygous recessive and present a mutant phenotype. This method of screening for clear phenotypic mutations has the advantage that induced mutation events are simple and fast to score. This is important since very large numbers of mice (typically several hundred thousand; Russell and Kelly 1982; Ehling et al., 1985) must usually be screened to establish a control base spontaneous mutation rate. In addition large numbers of treated mice can also be screened for induced mutations. Using this test a number of factors have been shown to be affect ionising radiation induced mutation load; dose, dose rate, dose fractionation and germ cell stage (Ehling and Favour 1984; Favom” 1989). The spontaneous mutation frequencies observed by different groups using this test in Harwell, Neunherberg and Oak ridge were very similar (reviewed in Favom” 1989) indicating the consistency of this method.

There are however certain disadvantages and limitations of this test as a measure of induced mutations. Although the majority of phenotypic changes in the seven locus test are likely to result from recessive mutations at one or more of the seven loci in the treated mice, it is not possible to distinguish exactly what type of mutation event, recessive or dominant has occurred. Recessive mutation events can be identified in the specific locus test because a homozygous recessive tester strain is used, but any mutation type may also cause phenotypic changes resulting from changes in the complex genetic pathways leading to the physical characteristics used in the assay. In a randomly mating natural population, newly induced recessive mutations will only result in an observable mutant phenotype when they occur as a homozygote. Dominant mutations by definition express phenotypic effects as a heterozygote, and would be visible in the F1 generation after radiation exposure regardless of genotype or mating scheme in the parental generation. Data from the UNSCEAR report (1986) observed that in human populations dominant disorders have a significant incidence with a high spontaneous mutational component (defined as the fraction of observed patients with dominant disorders due to spontaneous mutation events) as compared to recessive or irregularly inherited genetic disorders. The spectrum of genetic changes that can result in a recessive phenotype are much wider than those that result in a dominant or ‘gain of function’ mutation (Kacser and Burns 1981). Therefore it can be hypothesized that the mutation load resulting in dominant genetic changes is different from that measured for recessive genetic changes, as assayed by the specific-locus test. For this reason it is not appropriate to base all characterisations of the mutational process in germ cells of mammals on results with recessive alleles, given that it appears that dominant alleles pose the major genetic risk (UNSCEAR 1986).

A number of procedures were therefore developed to assay for solely for dominant mutations in mammalian germ cells of the mouse. These included assays for visible phenotypic changes (reviewed in Searle 1974; Searle and Beechey 1986) and mutations altering mouse cataracts (Krotochilova and Ehling 1979). The induced mutation rates for these traits could be directly compared with the induced rates to recessive alleles using the specific-locus test, since experiments were designed to screen for both types of mutation in the same experimental animals. A combined approach to screen for mutations at various genetic endpoints in the same experimental animals was
developed by Ehling et al (1985) at Neuherberg. The genetic endpoints included recessive specific-locus mutants, dominant cataract mutations, protein charge mutations and enzyme activity mutations. In this study the loci assayed in the specific-locus test were shown to be the most mutable and there was a similar per locus mutation rate to enzyme activity alleles. In contrast the per locus mutation rates to dominant cataract alleles and protein charge alleles was one order of magnitude lower than the mutation rate to specific-locus allele. The per locus mutation rate to dominant alleles had previously been observed to be lower than that to recessive alleles, which adds further support to the premise that a more specific DNA lesion is required to effect a dominant change. In the Ehling et al (1985) study, radiation was shown to be relatively more efficient in inducing recessive mutations than dominant mutations. By comparison, the compound ethynitrosourea, which acts by base alkylation and results in a relatively high frequency of point mutations, induces a higher frequency of dominant mutations than recessive specific-locus mutations (Favour 1989).

Two extrapolation procedures are available to estimate the genetic risk in man based upon experimental data from the mouse. The first is called the doubling dose (DD). This is defined as the dose of radiation needed to double the spontaneous mutation rate per generation. It is obtained by dividing the average rate of spontaneous mutations at a given set of loci by the rate of induction of mutations at the same set of loci. The radiation DD has been estimated in the mouse by the recessive specific locus test to be between 0.32-0.37 Gy in stem cell spermatogonia (Russell and Kelly 1982), and from the observed mutation frequency to dominant cataract alleles to be 0.96-2.1 Gy (Kratochilova and Ehling 1979). To estimate the number of induced cases of recessive mutation in the human population the following calculation is used; induced cases of mutation = the spontaneous mutation rate x (dose/DD).

The second extrapolation used to estimate the risk of ionising radiation to human populations is called 'the direct approach'. It is based on a comparison between the estimation of induced mutation rate to dominant alleles in the mouse model, with an estimate of the total number of loci in the human population involved in dominant mutation related disease relative to the number of loci in the human population controlling the particular mutation indicator used in the mouse. To estimate the number of induced cases of dominant mutation in the human population the following calculation is used; induced cases of mutation = induced mutation rate per gamete per dose x (dominant loci/indicator dominant loci) x dose.

Because of the low spontaneous mutation frequency observed at most loci in human and mouse populations, extremely large numbers of mice (typically 10,000 to >100,000; Russell and Kelly 1982; Ehling et al., 1985; Favour 1989) with a defined genetic background are necessary for analysis in the evaluation of the genetic effects of radiation dose. However, some minisatellite loci in human populations and in the mouse have been observed to have a high spontaneous mutation rate to new length alleles (Jeffreys et al., 1988, 1991b; Kelly et al., 1989; Gibbs et al., 1993). The mutation rates at some minisatellite loci are high enough to be identified by the pedigree analysis of only a few hundred alleles (Jeffreys et al., 1988, 1990, 1991b; Armour et al., 1989b). It was proposed that minisatellite loci would provide a good system for assessing the effects of ionising radiation since the analysis of relatively few animals is required to establish the baseline spontaneous mutation rate, for comparison with any increase in minisatellite mutation rate, induced by radiation. In addition minisatellites are thought to be selectively
neutral, obviating the potential effects of selection against the propagation of some radiation induced mutation events during mouse development.

The first work to assess the effects of ionising radiation on minisatellite loci was performed in this laboratory by Y. Dubrova (1993). In this study male mice were irradiated with an acute exposure of 0.5 or 1.0 Gy of γ-radiation from a 60Co source. After six weeks these mice were mated with untreated females so that offspring would be derived from sperm with mutations induced at the spermatogonial stem-cell (A0) stage. Progeny from these crosses and from control untreated crosses, were analysed by conventional DNA fingerprinting methods using the multilocus probes 33.6 and 33.15 (Jeffreys et al., 1985b) to look for mutation events. The minisatellite mutation rates were observed to be elevated in mice treated with γ-radiation, to give a doubling dose estimated at 0.5 Gy. The frequency of mutation events did not appear to increase with radiation dose indicating that there may be some ‘mutation saturation’ effect at higher doses. This was the first evidence that minisatellite DNA may be sensitive to ionising radiation. Recently it has been proposed that some of the genetic effects of radiation such as the induction of tumours and other physical anomalies, occur at too high a frequency to be caused by radiation damages in single-copy genes alone (Sadamoto et al., 1994). Therefore these authors suggest that it is prudent to assess for radiation induced DNA damage in other types of DNA, for example tandemly repeated DNA sequences, which comprise large proportions of mammalian genomes (reviewed in Singer 1981, 1982; Hastie 1989).

This work. Mutation events at the hypervariable human minisatellite MS32 show a tendency to occur towards the 5' end of the locus and many of these events are interallelic with a high proportion of gene conversion type events (Jeffreys et al., 1991, 1994). The observation of a mutation rate polymorphism (the 'OHCG' polymorphic site, chapter 3, figure 3.6) controlled by a 'G' to 'C' transversion in the 5' flanking DNA suggests that the mechanism of many of the mutants at this locus are cis-modulated by elements interacting with the flanking DNA (Monckton et al., 1994). Analysis of transgenic mice carrying a human MS32 locus have indicated that some of these transgenic loci may provide a system by which the effects of γ-radiation on a human locus can be assessed directly in a mouse model system. Although the transgene locus 1H0D locus does not mutate at the high rate of the endogenous human locus, the mutation rate is high enough (0.007%) to measure a base spontaneous mutation rate using the SP-PCR technique (chapter 6). The recovery and analysis by MVR-PCR (Jeffreys et al., 1991b) of three putative mutation events from this locus has not indicated any polarity of mutation (chapter 6). This would suggest that the 5' MS32 flanking DNA of the 1H0D locus does not influence mutation processes at this locus. These results suggest that the 1H0D transgenic MS32 locus would make an ideal candidate for the study of the effects of gamma radiation upon basic mutation processes at a minisatellite locus since there are no cis-modulating effects to account for in the analysis of mutations and the low but measurable spontaneous mutation rate from SP-PCR, should facilitate the observation of the small increases in mutation rate by induced mutation.

It is proposed that any increases in this rate induced by γ-radiation could be accurately measured using the SP-PCR technique on germline and somatic DNA directly from exposed animals. This would obviate the need for breeding large numbers of mice since SP-PCR on sperm DNA can assay many thousands of potential progeny for γ-radiation induced mutation events. The advantage of the SP-PCR technique is that it additionally enables the recovery of individual minisatellite mutation events and the subsequent analysis of these events by MVR-PCR (Jeffreys et al., 1991b). Generally studies of radiation induced mutation concentrate on the rates of mutation and do
Table 7.1. Details of the 110D mice used in the radiation experiments.

<table>
<thead>
<tr>
<th>Mouse number</th>
<th>Date of birth</th>
<th>Assigned dose (Gy)</th>
<th>Actual dose (Gy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>990</td>
<td>1.1.93</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>557</td>
<td>27.12.92</td>
<td>0.1</td>
<td>0.092</td>
</tr>
<tr>
<td>991*</td>
<td>1.1.93</td>
<td>0.1</td>
<td>0.092</td>
</tr>
<tr>
<td>1096</td>
<td>28.7.93</td>
<td>0.1</td>
<td>0.092</td>
</tr>
<tr>
<td>558</td>
<td>27.12.93</td>
<td>0.2</td>
<td>0.213</td>
</tr>
<tr>
<td>992*</td>
<td>1.1.93</td>
<td>0.2</td>
<td>0.213</td>
</tr>
<tr>
<td>995</td>
<td>1.1.93</td>
<td>0.2</td>
<td>0.213</td>
</tr>
<tr>
<td>560</td>
<td>27.12.92</td>
<td>0.6</td>
<td>0.615</td>
</tr>
<tr>
<td>997*</td>
<td>1.1.93</td>
<td>0.6</td>
<td>0.615</td>
</tr>
<tr>
<td>920</td>
<td>1.1.93</td>
<td>0.6</td>
<td>0.615</td>
</tr>
<tr>
<td>561</td>
<td>27.12.92</td>
<td>1.6</td>
<td>1.62</td>
</tr>
<tr>
<td>998*</td>
<td>1.1.93</td>
<td>1.6</td>
<td>1.62</td>
</tr>
<tr>
<td>993</td>
<td>1.1.93</td>
<td>1.6</td>
<td>1.62</td>
</tr>
<tr>
<td>562</td>
<td>27.12.92</td>
<td>3.0</td>
<td>2.99</td>
</tr>
<tr>
<td>994*</td>
<td>1.1.93</td>
<td>3.0</td>
<td>2.99</td>
</tr>
<tr>
<td>1094</td>
<td>28.7.93</td>
<td>3.0</td>
<td>2.99</td>
</tr>
</tbody>
</table>

* indicates mice used in the initial SP-PCR experiments.

Date of irradiation = 20.10.93  Date of DNA recovery = 1.12.93

All mice are C57BL/6J genetic background.
Mice numbers 990-998, 557-562 and 1096 and 1097 represent three groups of siblings.

Since it is not possible to accurately choose particular exposures from a 60Co source the assigned dose values represent the levels of planned irradiation and the actual dose values are a measure of doses during the exposure.

Table 7.2. Putative radiation induced mutation events at the 110D locus.

<table>
<thead>
<tr>
<th>Radiation dose (Gy)</th>
<th>Number of putative mutations &gt; 0.65kb</th>
<th>Putative mutation rate</th>
<th>Number of putative PCR artefacts seen &lt;0.65kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.007%*</td>
<td>5</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>0.03%</td>
<td>9</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>0.03%</td>
<td>1</td>
</tr>
<tr>
<td>0.6</td>
<td>3</td>
<td>0.1%</td>
<td>5</td>
</tr>
<tr>
<td>1.6</td>
<td>1</td>
<td>0.03%</td>
<td>unknown</td>
</tr>
<tr>
<td>3.0</td>
<td>0</td>
<td>?</td>
<td>unknown</td>
</tr>
</tbody>
</table>

* the unirradiated mutation rate is based upon the previously observed mutation rate at locus 110D (chapter 6).

Chapter 7, table 7.1
not analyse in detail individual events. It was hoped that this study would provide the first comprehensive analyses of a number of individual γ-radiation induced mutation events at the same allele.

Fifteen male mice (of C57BL/6J background) carrying a hemizygous copy of the 110D locus were identified. Trios of these mice were given acute doses of γ-radiation from a 60Co source of between 0.1 and 3.0 Gy. The doses were administered at the acute rate of 0.45 Gy/min. Three mice were exposed at each level to enable an average mutation frequency to be obtained. The mice were transported by courier to the MRC Radiobiology Unit in Harwell, Oxfordshire where irradiation from a 60Co source was performed by Mr. G. Howells and colleagues. The assigned doses were 0.1, 0.2, 0.6, 1.6, and 3.0 Gy. During irradiation the more accurate actual doses administered from the 60Co source were measured (table 7.1). After irradiation, the mice were returned to Leicester and kept under normal mouse house conditions before sacrifice. Six weeks later sperm and other somatic tissues were kept from each mouse. Sperm DNA was prepared from these mice and SP-PCR analyses were performed on one animal at each dose.

Results and discussion

Sperm DNA from exposed mouse numbers 991, 992, 997, 998, and 994 and from untreated mouse 990 were digested with Ava II. This enzyme cleaves in the mouse DNA flanking the transgene 110D locus. SP-PCR reactions containing DNA equivalent to 160 sperm per tube were performed using standard SP-PCR technique (as described in chapter 6) using sperm DNA prepared from these mice and from a human sperm sample. The human sample carried two alleles and was included as a positive sample to ensure that the correct number of PCR cycles to achieve amplification of individual mutant sperm molecules had been performed. The first attempt at amplification using the transgene flanking primers E1 and E2 resulted in a "background pattern" in each PCR lane on the Southern blot autoradiogram (data not shown). This pattern of additional bands was not as extensive as that seen for the multicity loci (figure 6.13, chapter 6) but the bands were of different intensities indicating that many of them were probably PCR artefacts rather than bona fide single molecule mutation events which would be expected to be of the same intensity. In addition the "background" bands in the PCR lanes were observed with the unirradiated mouse sperm and human sperm DNA samples. This type of pattern had not been observed before with these samples indicating that these initial SP-PCR reactions included many artefactual products.

These SP-PCRs were repeated using the same conditions but with a fresh digestion of the sperm DNA samples. The results are shown in figure 7.1. The "additional band background" in these lanes is very low. Four gain mutations were visible in the human sperm DNA sample, out of a total of 1440 sperm tested. This gave a mutation rate of 0.3% per gamete. This value did not differ significantly from unpublished data which showed a repeat unit gain mutant rate of 0.37% for the lower allele (A. Jeffreys, personal communication). Therefore in this experiment the SP-PCR technique was successful in the identification of mutations from single sperm.

No gain mutations were visible for the unirradiated control or any of the treated mice. There were a number of putative deletion mutations. These are split into categories in table 7.2. Putative deletion events larger than 0.65kb are visible on all three autoradiographs (figure 7.1, A, B and C). One putative deletion mutation event was observed at the 0.1 Gy and 0.2 Gy dose levels (two bands in the same PCR lane of a very low intensity in the 0.2
A, B and C. The SP-PCR was performed using transgene flanking primers 32A with 32D, for 25 cycles of 96° for 45s, 69° for 1 min, 70° for 3 min. Each SP-PCR tube had the equivalent DNA to represent 160 sperm and eighteen tubes were prepared from each sample of irradiated DNA, a total of 2800 sperm. Nine tubes of untreated sperm DNA and of human positive control sperm DNA were analysed at the same time. A 2μl aliquot of the 7μl PCR was electrophoresed through 1% agarose gel at 200V for 16 hours for gels 'A' and 'B' and for 20 hours for gel 'C'. These gels were Southern blotted and the membranes hybridised in a phosphate/SDS hybridisation mixture at 65° overnight with a 5kb MS32 repeat unit probe.
Gy dose rate sample were also observed, but assumed to represent artefacts of the PCR and not counted as mutation events. This gave a putative mutation rate of 0.03% per gamete. This was approximately four times higher than the spontaneous mutation rate observed at the 110D locus of 0.007%. The putative mutation rate is elevated to 0.1% for the 0.6 Gy dose level, approximately 14 times higher than the endogenous rate. Only one putative mutation was observed at a dose level of 1.6 Gy and no mutations were observed from sperm irradiated at a 3.0 Gy level. These results suggest that the 110D locus might be sensitive to induction of deletion mutation events by γ-radiation. The level of these putative mutations peaks at a dose of 0.6 Gy and appears to be reduced at higher doses. This may reflect a induced mutation 'saturation' effect between 0.5 and 1.0 Gy, as was previously reported by Y.Dubrova in the initial mouse minisatellite study (Dubrova et al., 1993). No deletion mutations are visible with either the untreated or 3.0 Gy dose sperm DNA. This may be because a dose of 3.0 Gy, which represents a high dose (Favom-1989) causes too much DNA damage in some of the immature spermatogonial stem cells for these cells to mature and be represented in the mature sperm used in the SP-PCR analyses.

The second class of putative deletion mutations are those smaller than 0.65kb (table 7.2). From figure 7.1 bands of this size are visible on the autoradiographs 'A' and 'B', but not on the last autoradiograph 'C'. Since these bands appear to be of different intensities and often are of the same size it was predicted that they were extremely likely to be the result of PCR artefact processes, for this reason they are not included in the calculation of putative induced mutation rates from γ-radiation.

Unfortunately however, it was not possible to identify the putative >0.65kb deletion events with confidence as bona fide mutation events, although it remains possible that some of these putative mutation events were real. There is evidence that mouse minisatellite loci are sensitive to mutation damage by ionising radiation (Dubrova et al., 1993) and therefore the MS32 transgene loci might be expected to be sensitive to radiation induced mutations as well. In addition deletion events have been reported as accounting for 90% of radiation induced mutation events (Sanaranarayanan 1993). However for the purposes of this study it was not possible to identify which putative mutations realistically represented γ-radiation induced mutation events. For this reason no mutation events were recovered and analysed by MVR-PCR and no analysis was performed using somatic tissues. The overall results were therefore inconclusive and it was not possible to say if radiation can induce new mutations in the MS32 transgene loci.

One approach that was not attempted due to time constraints was the use of a size selection for mutant sperm from the irradiated DNA panel. Since most mutants can be assumed to be deletions and the rate of mutation is probably quite low, a size fractionation of total sperm DNA as described for human sperm DNA in Jeffreys et al. (1990) would be appropriate. This technique would recover pools of DNA from a total DNA digest of various deletion 'catchment' sizes below the progenitor allele. Since the progenitor allele will not be present (in large amounts) to be amplified, huge numbers of sperm (many 1000s) can be analysed for potential mutation events in every PCR reaction. The limitation of this type of experiment would probably be the actual total number of mature sperm recovered from the mice for analyses, estimated from these mice to be approximately 120,000 sperm. This would give a more realistic estimate of induced mutation rate and the recovered mutation events would almost certainly be bona fide mutant events if they fell within the 'catchment' DNA size range.
Additional future studies might involve the study of the endogenous mouse minisatellite loci Ms6-hm and Hm-2 (Kelly et al., 1989; Gibbs et al., 1993). Recent evidence from one research group suggests that they have observed an elevated mutation frequency at the Ms6-hm locus (alternative name Pc-1 in this research group) in the progeny of mice treated with 60Co ionising radiation (3.0 Gy administered as an acute dose) identified by mutant bands on Southern blots probed with an Ms6-hm/Pc-1 repeat unit probe (Sadamoto et al., 1994). In this study the authors observe both gain and loss mutations at the Ms6-hm locus from irradiated male and female parental alleles. They also suggest that the level of induced mutation events at the Ms6-hm/Pc-1 locus is too high to be accounted for solely by DNA damage from a 3.0 Gy dose. They propose that this may reflect a previously unexpected mechanism of radiation induction of mutational processes at hypervariable tandem repeat loci in the mouse, that radiation may trigger genetic instability in mouse germ cells which in turn induces mutations at tandem repeat arrays throughout the mouse genome. This theory could be checked by PCR analyses of short tandem repeat elements in the mouse genome, followed by polyacrylamide gel electrophoresis, to ascertain if STR loci are unstable after exposure to radiation.

Interestingly Sadamoto et al. (1994) also noted that the elevation of mutation rate at the Ms6-hm locus was observed in sperm resulting from irradiation at the spermatid and spermatogonial stages, but not in sperm irradiated at the spermatogonial stages. In this experiment we looked at sperm which would have been irradiated at the spermatogonial stage. Dubrova et al. (1993) working in this laboratory reported that mutation frequencies in mouse DNA fingerprints were significantly increased by ionising radiation at the spermatogonial stage. In this study the minisatellite probes 33.6 and 33.15 detect many mouse minisatellite loci simultaneously (Jeffreys et al., 1985b, 1987b), and it is possible that distinct minisatellite regions may vary in their response to radiation exposure. However, it might prove fruitful to repeat the experiments in this chapter, but assaying for mutations at both the endogenous mouse loci Ms6-hm and the transgene locus 110D in sperm DNA irradiated at the spermatid and spermatogonial stages.

Some attempts were made to initiate a SP-PCR system at the mouse minisatellite locus Hm-2 (Gibbs et al., 1993) which is present in the C57BL/6J strain at a size accessible by PCR. However, so far this has not been successful. Problems stem in part from the extremely long arrays of tetrameric repeat units. The Hm-2 allele in these mice was shown to be 4.3kb, equivalent to >900 tetramer repeat units (data not shown). This size of Hm-2 allele can be amplified from 30ng of starting genomic DNA, but from SP-PCR levels of DNA the products appear to show a partial collapse to 1.3kb. However it would be interesting to pursue this type of experiment further, since a successful assay could be used to analyse the effects of many different toxicological substances on mouse germline and somatic mutation rates.

**Conclusions**

From the limited data in this study it was not possible to assess if y-radiation elevates the mutation rate at the Ms32 transgene locus 110D. However the panel of irradiated mouse DNAs produced for this project remain a valuable resource for further study.
Chapter 8

A study of tandem repeat sequence instability in HNPCC DNA

Summary

Hereditary non-polyposis colon cancer (HNPCC) is one of the most common cancer predisposition syndromes. Recently linkage analysis in two large kindreds demonstrated an association of HNPCC with markers on chromosome 2p. Most loci encoding tumour suppressor genes undergo somatic losses during tumourigenesis, but both chromosome 2p alleles were retained in HNPCC tumours. During a search for chromosome 2 losses it was noted that HNPCC tumours demonstrated a novel form of simple repeat sequence (STR) instability. One putative mechanism underlying this instability was suggested during the study of E.coli and yeast DNA containing defective mismatch repair genes which manifested as instability of dinucleotide repeats. Additionally an inability to repair slipped strand mismatches was observed in an HNPCC cell line. These studies suggested that defective mismatch repair proteins might be responsible for minisatellite alterations in the tumours of HNPCC patients. Further studies identified a human homologue of the yeast mismatch repair protein MSH2 (itself a homologue of the bacterial mutS), called hMSH2. This was shown to map to the region of chromosome 2p which showed linkage with HNPCC. Since then other mutations in different mismatch repair proteins have also been associated with HNPCC patients. In this study small pool PCR analysis was performed at five different minisatellite loci with repeat unit lengths between 9-54bp, on HNPCC normal and tumour DNA. The purpose of these experiments was to investigate if the tumour DNA instability observed at STR loci, extended to more complex tandem repeat arrays. No elevation of mutation rate was detected at the minisatellite loci in the tumour DNA by this method suggesting that the processes of strand misalignment and slippage which are involved in mutation processes at STR loci are not common at minisatellite loci. However some apparently clonal mutation events were identified during PCR analysis of two tetramer loci, indicating that the instability observed at other STRs extended to tetrameric loci.

Introduction

It has long been established that some cancer types, including common ones such as colorectal and breast cancers tend to run in families. This observation suggested that in addition to the effects of diet and other environmental factors on the development of these diseases, there was a tendency for susceptibility to these diseases to be inherited. One form of susceptibility to colorectal cancer (CRC) is known as familial adenomatous polyposis (FAP)(Mulvihil 1983). Patients with this disorder develop numerous benign colon polyps, a few of which subsequently develop into malignant carcinomas. FAP accounts for only 1% of CRC cases in the Western world and can be initiated by a defect in the APC tumour suppressor gene (Kinzler et al., 1991). A second form of CRC that shows a familial aggregation is hereditary non-polyposis colon cancer (HNPCC) which is estimated to account for between 4-13% of all CRC in industrial nations (Peltomaki et al., 1993 and references therein) and to affect as many as 1 in 200 individuals (Lynch et al., 1993). However, using physical examination it is difficult to distinguish between bona fide familial cases of this disease and sporadic cases, since unlike the FAP forms of CRC,
there are no distinguishing diffuse polyps or other physical features. Therefore HNPCC kindreds are usually defined by the so called "Dutch criteria" as kindreds where at least three relatives over two generations, with at least one less than 50 years of age, have non-polyposis CRC (Vasen et al., 1991). In order to provide better counselling and health monitoring for individuals suspected as being at risk from an inherited susceptibility of non-polyposis CRC, it was necessary to look for a method of genetically identifying those at risk rather than relying on speculative evidence for familial aggregation of the disease.

Recently groups led by Vogelstein and de la Chapelle carried out linkage analysis of HNPCC kindreds using more than 300 short tandem repeat (STR) markers and found that HNPCC segregated with a marker located on chromosome 2p (Peltomäki et al., 1993). In these studies two large kindreds were used with a mean onset of CRC of 41.1 years and 44.4 years. In addition to CRC, individuals in these kindreds were also susceptible to other forms of cancer including breast cancer. 345 markers were assayed and a significant LOD score of 6.39 was obtained for one marker (D2S123) suggesting close linkage. In these studies the precise location was not obtained, but estimated to be 2p15-16.

Some previously characterised cancer predispositions are caused by suppressor gene mutations (reviewed in Knudson 1993). In Aaltonen et al (1993) the researchers reasoned that if the pathogenesis of familial CRC differed from that of the more common sporadic forms, then the genes previously implicated in the latter may not be involved. Therefore the next step in the analysis was to examine tumour DNA from CRC affected individuals in the HNPCC families for changes in genes already known to undergo mutation in some sporadic cases of CRC, such as the tumour suppressor genes KRAS (located on chromosome 12), P53 (chromosome 17) and APC (chromosome 5). Mutations in these genes were observed in a large proportion of the HNPCC tumours analysed; for KRAS (61%), P53 (64%), and APC (57%). This was an interesting result since previously mutations in the APC candidate tumour suppressor gene had been established as the cause behind some cases of FAP, a syndrome which predisposes to CRC (Mullighan 1983). One possibility that had been proposed was that mutations in the HNPCC gene could substitute for mutations in the APC gene and that tumours would progress along either an "APC" or "HNPCC" pathway. This possibility was therefore ruled out by the observation of a number of mutations in the APC gene in the HNPCC patients.

Some genes responsible for cancer predisposition result in allele loss in tumours (Fearon and Vogelstein 1990). However analysis for allele loss of D2S123 in HNPCC tumours did not reveal any loss of heterozygosity at this locus. This was an unexpected result, but perhaps even more surprising was the observation of an unusual pattern of changes in the microsatellite profile of the D2S123 locus in tumour DNA when compared to normal DNA from the same individual (Aaltonen et al., 1993). The profile from tumour DNA showed additional bands (both gains and losses of repeat units) to the pattern from normal DNA, interpreted as an increase in instability at the locus in the tumour. This heterogeneity of microsatellite markers was not limited to chromosome 2p, but was observed genome-wide. These observations were supported by findings from other groups who showed that in a substantial proportion of both familial and sporadic CRC tumours (13%), there was a remarkable instability of mono-, di- and trinucleotide loci (Thibodeau et al., 1993; Ionov et al., 1993; Lindblom et al., 1993). On the basis of this data it was estimated that the total number of mutations at microsatellite loci alone could be as many as 100,000 per cell (Aaltonen et al 1993). Thus it was suggested that predisposition to HNPCC was associated with a susceptibility to
DNA replication errors. Tumours resulting from such a process were termed RER+ (replication error repair) tumours and the candidate gene for HNPCC was predicted to encode a factor involved in the DNA replication or repair pathways.

The best defined mismatch repair pathway is the *Escherichia coli* MutHLS pathway which promotes long patches (~2kb) of excision repair, and is dependent on the mutH, mutL, mutS and mutU(uvrD) gene products (reviewed Modrich 1991). This pathway is known to be the most active repair pathway in *E.coli* and to increase the fidelity of DNA replication as well as to act upon mispaired bases in recombination intermediates (Fishel et al., 1986). MutS protein binds to mismatched nucleotides in DNA. MutH protein interacts with GATC sites in DNA which are hemi-methylated on the adenine and is responsible for excision of mismatches on the unmethylated novel DNA strand. Specific incision of the unmethylated strand results in increased fidelity of replication since excision repair is targeted to the newly replicated and therefore unmethylated DNA. MutL facilitates the interaction between mutS which is bound to the mismatch and mutH bound to the hemi-methylated Dam site, resulting in the activation of mutH (Grilley et al., 1989). MutU(uvrD) is the helicase that appears to act in conjunction with one of the single-stranded DNA exonucleases to excise the unmethylated strand containing the mispaired bases. The "gap" left is filled by the action of DNA polymerase III helicase, single-stranded binding protein, and DNA ligase. Unstable repeat unit tracts have been observed in the *E.coli* strain mutS, which are deficient in mismatch repair function (Levison and Gutman 1987).

Although the prokaryote pathway is the best understood, genetic analysis suggests that *Saccharomyces cerevisiae* has a similar mismatch repair system to the bacterial MutHLS pathway (Reenan and Kolodner 1992). Study of the stability of simple sequence repeats in the yeast *S.cerevisiae* demonstrated that mutations in any one of three genes involved in DNA mismatch repair (PMS1, MLH1 or MSH2) resulted in a massive (100- to 700-fold) increase in instability of these simple sequence repeat unit tracts (Strand et al., 1993). MSH2 is a yeast homologue of the bacterial mutS protein, and PMS1 and MLH1 are mutL homologues. Each of these three genes plays an indispensable role in DNA replication fidelity. In the same study pol2 and pol3 mutant yeast strains, which have mutations that eliminate the proof-reading functions of the two yeast DNA polymerases (POLII and POLIII) were found not to cause much effect on repeat unit tract instability. Therefore it is proposed that DNA polymerase in vivo has a very high rate of slippage on templates containing simple repeats, but that the majority of these errors are corrected by cellular mismatch repair systems. These results strongly suggested that the HNPCC gene may encode a protein involved in DNA mismatch correction.

Such a mutation may represent a functional human homologue of PMS1, MLH1, MSH2 or another component of the yeast mismatch repair system, for example a DNA helicase or a single-stranded binding protein. Fishel et al (1993) identified a human homologue of the bacterial mutS and *S.cerevisiae* MSH proteins called hMSH2. To isolate the hMSH2 gene they utilised a degenerate PCR approach that had previously been successful in the identification of the mutS homologues of *S.cerevisiae*, due to the conserved nature of the yeast sequence compared to the bacterial sequence. Primers were used in a PCR containing human cDNA prepared from poly(A)+ RNA as a template and a 360bp fragment was recovered. This was the expected fragment size predicted from conserved sequences in the bacterial and yeast genomes. Sequence analysis showed that this DNA had an 81% sequence identity with the *S.cerevisiae* MSH2 protein and it was used as a probe to screen a human cDNA library to find the
full length clone. Comparison of the hMSH2 amino acid sequence with other mutS homologues showed that it was most highly related (mean 41% sequence identity) to the S.cerevisiae protein.

The hMSH2 protein was expressed in E.coli under the control of the lac promoter and was observed to cause a dominant "mutator" phenotype resulting in a ten-fold increase in mutation rate (Fishel et al., 1993). This is an interesting effect since it appears that the hMSH2 protein is in some way blocking the bacterial mismatch repair pathway. It may result from the hMSH2 mutS homologue having the ability to bind to mismatched nucleotides in the bacterial DNA, but lacking the ability to interact with other bacterial mutHLS proteins such as mutH, to allow excision of the mismatched nucleotides.

In addition, the HNPCC cell line H6 was shown to be defective in mismatch repair in vitro and in vivo (Parsons et al., 1993). Transfection of heteroduplex plasmids, engineered with a run of dinucleotide repeat units, was performed into H6 cells. Analysis of these repeat sequences showed that they acquired mutations with a two-fold higher frequency in H6 cells than in a control cell line, confirming the "mutator" hypothesis (Aaltonen et al., 1993; Fishel et al., 1993) as a contributory factor in HNPCC. Experiments with heteroduplexes containing single base mismatches confirmed that H6 cells were deficient in the repair of slipped-strand mismatches since neither base-base mismatches nor extrahelical nucleotides were repaired in H6 cell extracts. This suggests that in addition to the known role of mutS mutants in the process of destabilising microsatellite repeats in E.Coli, Yeast and Man (Levison and Gutman 1987; Strand et al., 1993; Aaltonen et al., 1993) that RER+ cells may also prove to be unstable to transition and transversion mutagenesis (Parsons et al., 1993).

In a separate study Leach et al (1993) mapped the HNPCC locus to the human chromosome 2p22-21. Physical and linkage mapping were used to delineate a 0.8Mb region of chromosome 2p harbouring the HNPCC gene. cDNA clones of the hMSH2 mutS homologue were identified and shown to reside within this 0.8Mb region of chromosome 2p. Alterations in this hMSH2 gene were identified by PCR based sequencing, both in the germline and tumour tissue, of patients with RER+ tumours. These results provided the final evidence that mutations of hMSH2 are most likely responsible for HNPCC and the RER+ phenotype of some CRC tumours.

More recently other mismatch repair gene mutants have been shown to be associated with HNPCC patient tumours, for example hMLH1 (a homologue of mutL, assigned by linkage data to a locus on chromosome 3p21-23, Bronner et al., 1994) and hPMS1 and hPMS2 (homologues of the mutL gene, Nicolaides et al., 1994). This is not a surprising observation given that the mismatch repair complex is composed of many different proteins, and that malfunction of any one may give rise to an HNPCC phenotype. There are now four mutant mismatch repair genes associated with HNPCC tumours, which may go some way to explaining the high incidence of this disease.

How might a malfunction of mismatch repair cause cancer? The findings of various groups studying instability at HNPCC tumours raises a number of questions. Why do mutations in hMSH2 appear to result in a predisposition to certain types of tumours, for example CRC and breast cancer, rather than tumours of every cell type? hMSH2 is expressed in lymphocytes and many other cells and therefore its presumptive function would appear to be a 'housekeeping' one leading to the puzzling question of why mutations are only observed in the tumour cells of these patients and not in normal cells. DNA samples from lymphocytes of affected HNPCC
individuals have been observed to be heterozygous for observed mutations of hMSH2 (Lesch et al., 1993). This suggests that the presence of a single functional copy of hMSH2 might be sufficient for normal development and therefore that the transformation from normal to tumour tissue may require that the remaining allele becomes lost or inactivated. However, it is known that there is no significant allele loss in HNPCC tumour DNA (Aaltonen et al., 1993). It is possible that the acquisition of secondary mutations in the hMSH2 gene may be enough to initiate tumourigenesis. Although there is no direct experimental evidence that cells which are heterozygous at the hMSH2 locus give rise to mutations at a higher than average frequency, it has been observed that in extracts of "KK" cells, which are derived from an HNPCC individual, the capacity to control mismatch repair is lower than that of normal cells from the same lineage (Pusons et al., 1993). It is possible that the reduced dosage of the hMSH2 gene in heterozygotes could lead to an increase in mutations not being repaired genome-wide. In this case the second hMSH2 gene would be more likely to accumulate a second non-repaired mutation.

A second possible scenario is that a mutation in the second hMSH2 allele is not required. In this case the heterozygous dose of functional hMSH2 would be sufficient to correct mismatches until a second unrelated event brought about neoplasia. During the subsequent period of uncontrolled growth the reduced dosage of functional hMSH2 may not be sufficient to cope with an increased level of mismatches, resulting in an accumulation of post-replicative mutations. Previously it has been observed that E.coli mutator cells, defective in proof-reading function generated so many heteroduplexes that heteroduplex repair became exhausted yielding an extreme mutator effect (Schapper and Radman 1989). It is possible that in humans the rapid proliferation associated with neoplasia could saturate the post-replication repair capacity offered by a single functional copy of the hMSH2 gene. During this period, the rate of tumourigenesis would be increased due to a rapid accumulation of mutations in other genes involved in growth control, such as APC. This is consistent with data from Aaltonen et al. (1993) who observed a high percentage of alterations in the tumour suppressor genes APC, P53 and KRAS.

The second scenario is perhaps more consistent with the observation that mutations appear to only be observed in the tumour itself. With the first scenario all heterozygous cells whether normal or tumour should give rise to mutations at a higher rate. In the second scenario a mutator phenotype will only be initiated in tumour tissue. The presence of one mutant hMSH2 allele would therefore give a predisposition to the disease, but the transition to tumourigenesis would be triggered by an unrelated event, perhaps the growth of a benign polyp. In the second scenario it may be possible that the tissue specificity of the tumour is dictated by which additional gene is mutated, for example APC, KRAS and P53, all previously associated with CRC (Aaltonen et al., 1993; Bodmer et al., 1994).

This work. The primary aim of this study was to ascertain if the tandem repeat instability seen in HNPCC tumours at STRs is shared by more complex and longer repeat sequence arrays, such as minisatellite loci. There is evidence from previous studies of gastrointestinal tumours for VNTR instability, though in these earlier studies the tumours were obviously not correlated with RER status (Armour et al., 1989b). This suggested that in a subset of these tumours, which may have corresponded to sporadic CRC cases linked to hMSH2 or another mismatch repair mutant, minisatellite loci were unstable. This indicated that it was worthwhile studying some established HNPCC tumour DNAs for evidence of VNTR instability thus allowing for assessment of the extent of strand misalignment and slippage in generating minisatellite instability.

Chapter 8, page 5
It was hoped that if an elevated mutation rate was observed then it might be possible to directly quantify the extent to which instability may vary between individuals. Additionally we proposed that there may be an elevated mutation rate in non-tumour tissue in patients who carry the HNPCC gene.

The proposal was to analyse any minisatellite mutation events in RER+ DNA samples by the use of small pool PCR (SP-PCR) on a panel of normal and tumour DNAs from patients who showed linkage to the D2S123 marker (Peltomaki et al., 1993) and had been established as showing STR instability (Aaltonen et al., 1993). It was hoped that the single molecule mutation techniques developed in this laboratory would be able to measure even low induced frequencies of mutation, even if there was no substantial mutational mosaicism, as was the case in microsatellite analysis. Structural analyses of mutant alleles would enable us to address questions of processes of mutation. Pairs of non-tumour colon cell line DNA (referred to as "normal" DNA) and CRC tumour cell line DNA from three different individuals were kindly provided by B. Vogelstein, from the Dept. of Medical Oncology at John Hopkins Medical School, Baltimore, MD, USA.

SP-PCR reactions were performed using suitable alleles identified in the HNPCC patient DNA samples for five minisatellite loci; MS32, MS205, MSG21, MS1 and pXg3. In addition analysis of some tetrameric STR loci was performed to ascertain if the HNPCC instability extended to these loci.

**Results**

**Bulk DNA PCR analysis for evidence of instability and for allele sizing.** In the initial PCR analyses 50ng samples of the normal/tumour DNAs were amplified using flanking primers extending across the whole locus for 20 cycles, according to the parameters in table 8.1. These products were electrophoresed through 1% agarose, then Southern blotted and hybridised in a phosphate/SDS hybridisation mixture at 65° overnight with a minisatellite probe (data not shown). This was done to identify any gross instability in the tumour DNA, as represented by mutational mosaicism and to enable the identification of any minisatellite alleles of a suitable size for SP-PCR analysis. The most efficient allele size for SP-PCR is <5kb. During this analysis it became apparent that one or more of the alleles at one of the normal/tumour DNA paths was consistently different for each minisatellite locus amplified. Since this "mutation" between normal and tumour DNA was not observed at the other two pairs of normal/tumour DNAs, it was assumed to be the result of a sample mix-up in the laboratory where the DNAs were made.

**SP-PCR of the HNPCC samples.** SP-PCR was performed at five minisatellite loci (as described in figure 8.1) for ~1680 individual somatic cell equivalents from both tumour and normal DNA. It was immediately apparent that there was no detectable elevation in minisatellite mutation rate for the tumour DNA as compared to the normal DNA in the two patients CO86 and CO88. Two small somatic mutations (one loss and one gain of repeat units) were identified in the normal DNA from patient CO88 at the MS32 locus. These were not recovered for analysis and no other mutations at the other loci were seen. Under these SP-PCR conditions the technique does not amplify particularly well across the MS205 locus (this work and J. Armour, personal communication), and the autoradiograph profile of the SP-PCR at this locus (figure 8.1 D) reflects this.
Table 8.1. Details of the tandem repeat unit loci used in HNCC DNA analysis.

<table>
<thead>
<tr>
<th>Locus name and type</th>
<th>Repeat unit size (bp)</th>
<th>Primers and annealing temp</th>
<th>DNA probe size (kb)</th>
<th>Mutations</th>
<th>Non-tumour DNA</th>
<th>Tumour DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>i/ minisatellite</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS32</td>
<td>29</td>
<td>MS32 A + B, 67(^\circ)</td>
<td>5.0, repeat units</td>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>MS1</td>
<td>9</td>
<td>MS1 A + B, 67(^\circ)</td>
<td>4.6, repeat units</td>
<td>?</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>p(_)g3</td>
<td>33/37</td>
<td>p(_)g3 C + D, 67(^\circ)</td>
<td>6.3, repeat units</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>MS205</td>
<td>45/54</td>
<td>MS205 A + B, 67(^\circ)</td>
<td>5.1, repeat units</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>MS621</td>
<td>32/55</td>
<td>MS621 A + C, 67(^\circ)</td>
<td>5.0, repeat units</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>ii/ STR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wg0c7</td>
<td>3</td>
<td>wg0c7 A + B, 65(^\circ)</td>
<td>(labelled primer)</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>wg1c4</td>
<td>4/5</td>
<td>wg1c4 A + B, 60(^\circ)</td>
<td>(labelled primer)</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>mk2c8</td>
<td>4</td>
<td>wg2c8 A + B, 62(^\circ)</td>
<td>(labelled primer)</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>

(Primer sequences are detailed in table 2.1)
General SP-PCR. The SP-PCRs were performed on aliquots of normal and tumour DNA from HNPCC individuals CO86 or CO88, depending on which individual had the best size of allele(s) for the analysis. DNA samples were digested with Mbo I and the concentration accurately assessed using comparative gel electrophoresis and fluorimetry (see chapter 2 for details). Sets of six duplicate SP-PCR reactions were performed using aliquots of DNA corresponding to 40, 80 and 160 individual cells. SP-PCR was as described in chapter 6. PCR cycle conditions for all five loci were 96° for 45s, 67° for 1 min, 70° for 3 mins for 25 cycles in a 7μl reaction. 2 μl of product was electrophoresed through 1% agarose, Southern blotted and hybridised in a phosphate/SDS mixture overnight at 65° with 10ng of the appropriate minisatellite probe (as detailed in table 8.1).

A. Amplification was with the primers MS32 A and MS32 B. 40, 80 and 160 refer to the number of input molecules. CO88T refers to input DNA of individual CO88 tumour DNA and CO88 N of a normal DNA input. The size of the amplifiable allele in individual CO88 was 3.5kb. Two mutant events are identified in the non-tumour DNA.

B. Amplification was with MS1A and MS1B. The size of the amplifiable MS1 allele in individual CO86 was 3kb.

C. Amplification was with pA,g3 A and pA,g3C. The size of the amplifiable pA,g3 allele in individual CO88 was 6.5kb.

D. Amplification was with MS205A and MS205B. Both alleles in individual CO88 were of amplifiable sizes; 2.4 and 3.6kb.

E. Amplification was with MS621A and MS621C. Both alleles in individual CO86 were of amplifiable sizes; 2.15 and 2.25kb.
Chapter 8, figure 8.1
STR analysis of the HNPCC samples. PCR amplification with one primer end labelled with $^{33}$P was performed for one triplet repeat locus and two tetramer loci (as described in Armour et al., 1994) on DNA from the individuals CO86 and CO88 plus two positive control CEPH individuals (CEPHA and CEPIHb). Aliquots of the samples were analysed by polyacrylamide gel electrophoresis and were electrophoresed alongside a M13 phage sequencing ladder. Although the different nucleotide content between the M13 sequence and the STR loci means that accurate sizing by comparison with the known M13 sequence is not possible, the presence of the sequence enables the determination of the number of nucleotides lost or gained in mutant bands.

The tetramer repeat locus wglc4 (heterozygosity = 90%) on chromosome 8 (Armour et al., 1994) has a basic repeat unit of AGGA/G. In addition to the majority of polyprune tetramer repeat units there are a few (2-3, in the alleles sequenced, J. Armour personal communication) pentamer AAGAA repeat units. Amplification of this locus gave apparently clonal mutation events in the tumour DNA of individuals CO86 and CO88 (figure 8.2 A). If it is assumed that the mutant band lies closest in size to its progenitor allele, then for individual CO86 the smallest allele has a gain of 5 nt (or a gain of one pentamer repeat unit). In individual CO88 tumour DNA the larger allele shows a 4nt or one repeat unit loss, and the smaller allele shows an 8 nt or two repeat unit gain.

The tetramer repeat locus mlc2c8 (heterozygosity = 80%) on chromosome 9 (M. Koutsourakis, unpublished data) has a number of variant repeat units; (CCTT), (CTTT), and (CCCT). Amplification of this locus gave apparently clonal mutation events in the tumour DNA of individuals CO86 and CO88 (figure 8.2 B). In individual CO86 the larger allele has been deleted by a length of 8 nt, or two repeat units. The smaller allele remains unchanged. In individual CO88 both alleles have changed in the tumour DNA. Assuming that the mutant band lies closest to its progenitor allele then the largest allele has a gain of 4 nt or one repeat unit and the smaller allele a loss of 4nt or one repeat unit.

Discussion

STR loci. Analysis of tumour DNA from two HNPCC patients for two tetramer repeat loci indicated that mutation events had occurred in the tumours of these individuals (figure 8.2) and that the tumour DNA was largely clonal for these mutation events. The mutation events generally involved the loss or gain of whole 4 nt repeat units. In the case of locus wglc4, analysis of patient CO86 revealed a 5nt change in the mutant tumour profile. This is likely to correspond to a gain of one of the pentameric repeat units observed in this locus (J. Armour, personal communication).

No trace of the progenitor alleles was present in the mutant DNA profiles of the tetramer loci. This may reflect the fact that a single mutant cell contributed to the tumour, or that a clonal region of the tumour was used to produce these tumour cell lines. Data on the production of these cell lines in the USA was not available, but it is possible that a non-clonal pattern would be observed if these experiments were performed on DNA directly from tumour tissue. A recent study from Australia (Young et al., 1994) also demonstrated instability at tetranucleotide loci in HNPCC tumour DNA. The profiles obtained in these PCR reactions indicated the presence of the progenitor alleles in addition to mutant alleles. The authors suggest that the progenitor bands may result from stromal contamination, but it also remains possible that the tumour tissue is non-clonal. Support for this idea comes from Shibata et al. (1994) who suggested that genomic instability in simple repeated sequences occurs early in tumorigenesis and that it
Figure 8.2. PCR analysis of two tetramer loci in HNPCC DNA.

A. **Locus wg1c4.** Amplification of 60ng of CO86 and CO88 normal and tumour DNA samples, was for 20 cycles of 96° 45s, 60° 1min and 70° 1min in a 10µl reaction, using the primers wg1c4A (end labelled with 33P) and wg1c4B. A 2µl aliquot was electrophoresed through a 6% polyacrylamide gel as described in (Armour et al., 1994). CEPHa and CEPHb are CEPH lymphoblastoid DNAs used as positive controls to check that a 'clean' PCR amplification is obtained. 50ng of CEPH DNA was amplified under the same conditions as the HNPCC DNAs. CO86N and CO88N refer to normal DNA from these individuals and CO86T and CO88T refer to tumour DNA. The normal DNA represents the progenitor bands. TCGA represents a section of M13 control sequence, prepared by single stranded sequencing according to the Sequenase protocol. The sequence is labelled with 35S.

B. **Locus wg2c8.** Amplification of the CO86 and CO88 normal and tumour DNAs and the CEPH DNAs was for 20 cycles of 96° 45s, 62° 1min and 70° 1min in a 10µl reaction using the primers wg2c8A (labelled with 33P) and wg2c8B. A 2µl aliquot was electrophoresed through a 6% polyacrylamide gel as described in (Armour et al., 1994). Lane labels are as in A.
Shibata et al. (1994) suggested that genomic instability in simple repeated sequences occurs early in tumorigenesis and that it persists after neoplastic transformation. In this study microsatellite loci were shown to accumulate consecutive somatic slippage events of a single or a few repeat units during the culturing of colon tumour cells. Analysis of single cell clones from cultured tumour cells showed different microsatellite profiles indicating that genomic instability persists during tumourigenesis.

This information indicated that it might be fruitful to analyse the CO86 and CO88 tumour cell lines for evidence of persistent instability at the tetramer loci. Since we did not have cultured cells it was not possible to obtain individual single cell clones for analysis, as described by Shibata et al. (1994). However, at the minisatellite locus MS32 the SP-PCR technique has been able to identify single cell mutation events (Jeffreys et al., 1994). Therefore it was decided to attempt to extend the use of the SP-PCR technique to the analysis of tetramer loci. This was done by performing SP-PCR reactions using one end-labelled primer and separating the products by polyacrylamide gel electrophoresis. However, the number of PCR cycles required to amplify from such small amounts of DNA to give a product of the correct relative intensity to the progenitor band to represent a single mutant molecule detectable by autoradiography, was between 27-30 cycles (data not shown). The amount of non-specific products formed in these PCRs when compared to those from 20 cycle PCRs (performed in the original experiments, figure 6.2) was greatly increased, and it was not possible to identify clearly any mutant bands. For this reason this line of experimentation was not pursued, although it is possible that with the use of a nested primer strategy (non-labelled amplification with outer primers followed by nested amplification with end labelled primers) it might be possible to amplify the progenitor and any mutant bands at these tetramer loci, without non-specific products.

Instability at STRs is thought to result either from recombination between misaligned repeated sequence tracts or from DNA polymerase slippage along the replicating strand during replication. In the first mechanism two tandem repeat arrays (possibly located on sister strands of a partially replicated molecule, on sister chromatids or on separate chromosomes) pair in a misaligned configuration. Crossing over between the misaligned arrays results in a deletion from one array and a duplication in the other. In the DNA polymerase slippage model of repeated sequence instability, the primer (leading) strand transiently dissociates from the template and reanneals out of register. If the misalignment involves the folding back of the template strand then a repeat unit deletion is seen, or if the primer strand is folded back then an increase in repeat units is the result.

The hMSH2 mutation status of RER+ tumours means that the mismatch repair system in these individuals is defective. Therefore mutations which naturally accumulate during DNA replication from either of the above mechanisms of recombination or strand slippage, are not properly corrected by the appropriate repair systems. Thus when the HNPCC gene is defective additional mutations are observed since they are not repaired. The fact that the mutations in dinucleotide and trinucleotide loci have been shown to accumulate after tumourigenesis (Shibata et al., 1994) lends weight to the idea of a failure to correct mutation events rather than an induction of mutation events by the HNPCC gene.

From this study of tetramer loci, clonal mutation events were observed in the tumour DNA. This indicates that strand slippage or recombination processes between misaligned tetramer repeats occur naturally at these loci and that these are not repaired in RER+ individuals. However, from this data and from other groups data showing
instability at dinucleotide and trinucleotide loci, it is not correct to say (as has been proposed by some groups) that the major form of mutation in non-HNPCC individuals at these loci is by the above two mechanisms. In non-HNPCC individuals there may well be other as yet undefined pathways to new alleles which are just not elevated in tumour DNA from HNPCC individuals since they are not modulated by a mismatch repair pathway.

**Minisatellite loci.** Somatic mutation events are rare in minisatellite loci (Armour et al., 1989b; Jeffreys et al., 1990, 1991, 1994). Additionally in the male germline, replication slippage and unequal recombination have not been established as the major mutation mechanisms operating at the minisatellite locus MS32 (Jeffreys et al., 1991, 1994). Therefore the lack of minisatellite mutations in the tumour tissue of the HNPCC individuals CO86 and CO88 may be simply explained by the fact that although these individuals have defective mismatch repair systems, slipped mismatches at minisatellite loci are not common to remain unrepaired in HNPCC individuals. In STR loci single strand breaks occurring within the repeated region during replication have been proposed as being the initiation events for strand displacement and replication slippage (Richards and Sutherland 1994). In contrast studies in this laboratory have indicated that minisatellite mutation is influenced by cis-acting elements outside the repeat unit array possibly initiated by double-stranded breaks in the DNA, making a predominant role for replication slippage seem unlikely.

The minisatellite MSI has a 9bp repeat unit. Allowing for the technical difficulties of identifying changes in small numbers of repeat units at this locus in normal or tumour DNA due to its size in individual CO86 (3kb) the SP-PCR results did not indicate a significant increase in unrepaired mismatch slippages in tumour DNA at this locus. Analysis of the wglc4 locus indicated that a mutation had occurred in the pentamer repeat unit part of this locus. There are only 4-5bp in repeat unit length between the MSI minisatellite and the tetramer(pentamer) STR wglc4 locus, and it would be interesting to know what parameters define the importance of mismatch repair processes in controlling mutation at these different classes of loci. The wglc4 locus has an AAAA<sup>G</sup> repeat unit and the pentamer repeat units at this locus are AAGAA. By comparison the MSI locus shares the common minisatellite G/C-rich "core" sequence (chapter 1). It would be interesting to test the theory that A/T-rich tandem repeat units may be more prone to mutation by strand slippage than G/C-rich repeat sequence, and therefore that mismatch repair systems play a more important role in these type of loci. One A/T-rich minisatellites is the Apo(B) minisatellite, at which MVR-PCR has been established (Desmaris et al., 1993) and it may be that this locus and other longer A/T-rich tandem repeat units might show some instability in HNPCC tumour DNA.

**Conclusions**

From these experiments it is clear that tandem repeat sequences of 4-5bp are destabilised in the tumour DNA of RER<sup>+</sup> individuals. This is likely to result from a failure of the hMSH2 protein to initiate binding to mismatches occurring during the replication of these loci. These mutations appear clonal in the cell line tumour DNA provided. In contrast an elevated level of mutation was not observed at five minisatellite loci with repeat unit sizes of 9-54bp), during analysis of a large number of somatic tumour cells from HNPCC tumour DNA. This indicates that strand slippage and recombination between misaligned repeat units are not common processes operating upon minisatellite loci.
Chapter 9

Discussion

Much of the work in this thesis has been discussed in detail in the relevant chapters. Therefore this final discussion will concentrate on how the results from mutation analyses of the multi-copy and single-copy MS32 transgenes relate to recent developments from studies of human minisatellite mutation performed in this laboratory in parallel with the MS32 transgensics project. In addition an evaluation of minisatellite mutation processes in comparison to turnover processes at simple tandem repeat loci will be discussed.

Human MS32 locus hypervariability

Allele diversity as revealed by MVR-PCR at the human MS32 locus. Many human minisatellite loci are highly polymorphic. This was first demonstrated from Southern blot allele length analysis studies which identified polymorphism as variations in allelic repeat unit copy number (Wong et al., 1986, 1987; Nakamura et al., 1987). MVR-PCR now provides the best method for determining allelic variability through the analysis of variant repeat unit interspersion (Jeffreys et al., 1991b). Previous studies of the MS32 locus in Caucasian populations using MVR-PCR revealed an extreme level of allelic diversity of MS32 MVR-PCR codes (Jeffreys et al., 1991b). In the present study Asian pedigrees were analysed using MVR-PCR, demonstrating that the level of MS32 allelic variability in Asians was as high as that seen in Caucasian populations and subsequently in other population groups such as Japanese, and Malaysians (K. Tamaki and C.L.Koh, unpublished data) and that no population specific alleles were identified, both important considerations in the potential forensic use of the MS32 locus.

In contrast to the high levels of MS32 locus variability seen in Caucasian, Asian, Japanese and Malaysian populations, diversity between alleles of an African origin has been shown to be markedly reduced (Monckton et al., 1994). Reduction in variability arose from an excess of some alleles detected more than once in the sample, including a common short (38 repeat units) allele, and from a second set of alleles which were fixed for α-type repeat units for the first 50-70 repeats. The reasons underlying this unusual locus behaviour will be discussed later.

Forensic applications of minisatellites. Currently minisatellite loci are widely used in the forensic analysis of DNA samples, both in the production of multi-locus 'DNA fingerprint' profiles and as single-locus probes (reviewed in Jeffreys and Pena 1993). The hypervariable nature of these loci makes them invaluable as individual specific markers (Jeffreys et al., 1985b). The most common applications are the determination of disputed family relationships in paternity and immigration cases (Jeffreys et al., 1985c, 1991b) and in forensic casework for individual identification, by comparison of DNA profiles from criminal suspects to the profiles of DNA from crime scene samples (e.g. Gill et al., 1985; Gill and Werrett 1987). After the introduction of 'DNA fingerprinting' into the courts there was some controversy as about differences in the DNA typing procedures of different laboratories (e.g. Thompson and Ford 1991) though these fears have been largely allayed by the standardisation of practical procedures between laboratories (Gill et al., 1992; NRC DNA Technology in Forensic Science, 1992).
There were also scientific arguments about the population assumptions used in the calculation of expected band matching frequencies and subsequent statistical assessment of positive individual identifications (e.g. Lewontin and Hart 1991; Clark and Kidd 1991). The statistical evaluation of a DNA profile requires knowledge of allele frequencies and the assessment of the Hardy Weinberg equilibrium assumptions for human populations. The major objection to the statistical analyses used were that significant population sub-structuring, of allele length frequencies, may exist. For less variable loci, with a smaller number of alleles it is possible that some allele frequencies may be more common in different population groups and that in inbred population groups a homozygote excess may occur. However, for the hypervariable loci commonly used in forensic analyses, similar allele length distributions are found in different population groups (reviewed in Jeffreys et al., 1991). In an attempt to bring to a close arguments about the technology, the National Research Council (NRC) of the United States National Academy of Sciences commissioned a report. This concluded that by performing modest surveys of 100 individuals from representative population groups within a country, the likely allele variations can be established (NRC DNA Technology in Forensic Science, 1992). In addition this committee established the use of a 'ceiling frequency' approach to estimate band match significance. This uses the maximum allele frequency in a population, or 5%, whichever is the greater, as a basis for estimating genotype frequencies and therefore band match probabilities. This results in a very conservative estimate weighted in favour of the defence in criminal proceedings, although the reduction in informativeness can be compensated for by the use of more tests, for example establishing a six locus match rather than a four locus match. Although some level of controversy still exists (e.g. Morton et al., 1993) it is hoped that the new conservative estimates will lead to a complete acceptance of the accuracy of DNA typing results.

There are technical disadvantages to conventional multi-locus probe (MLP) analyses. They require at least 500ng of good quality DNA, which may not be available from degraded forensic samples, and the poor resolving power of agarose gel electrophoresis in comparison to the number of minisatellite loci that hybridise in a multi-locus profile, means that it can be difficult to interpret these profiles. This difficulty is compounded by the need to make comparisons between 'identical' profiles run on separate blots in different laboratories. The highly complex nature of the data from multi-locus profiles means that it is not practical to store this information in computer databases for use in forensic comparisons or in research for the study of allele diversity and population structures. The use of single-locus probes, which require less input DNA and are easily resolved, goes a long way to alleviate these problems (reviewed Jeffreys and Pena 1993).

**MVR-PCR as a forensic tool.** The underlying allelic diversity at the MS32 locus potentially provides an ideal system for personal identification which can also be used in parentage analysis. As stated in chapter 3, consideration of the current world population size and the mutation rate at MS32 indicates that the actual numbers of different MS32 alleles could be >10^8 and that MVR-PCR at this locus could in theory distinguish between 7 x 10^23 different alleles solely by comparison of a-, t- and o- type repeat unit codes over the first 50 repeat units from the 5' end of the allele. The simplicity of diploid and single allele codes means that they can easily be stored in computer databases and the digital nature of the code means that inter-blot comparisons can be easily made. In addition the technique is PCR based and can be applied to small amounts of DNA such as those found in forensic samples. Currently MVR-PCR is well established at another hypervariable minisatellite locus, MS31A (D7S21) and multi-plex MVR-PCR of both MS31A and MS32 in the same PCR reaction has been successfully performed.
It is estimated that one additional locus as variable as MS32 and MS31A would be sufficient to meet the rigorous level of power needed for widespread use of MVR-PCR in forensics (D. Neil 1994) although additional analysis, perhaps using STR loci typing, might still be required to distinguish between closely related individuals such as siblings who have a \( \frac{1}{4} \) chance of sharing the same diploid code. One additional locus is \( p\lambda g3 \) (D7S22) at which MVR-PCR is in development (T. Guram, personal communication). In addition MVR-PCR has been developed at two short minisatellite loci MS205 (D16S309) and MS621 (D5S110). These loci have lower variability in their MVR-PCR profiles and it is not possible to produce diploid codes from total genomic DNA (Armour et al., 1993 and J. Armour personal communication). For these reasons these loci are of great value in forensics, but may play a role in population studies. Of great interest forensically is the discovery of a Y-chromosome specific minisatellite (M. Jobling, personal communication). Since this locus has no homologue it would be possible to generate single allele MVR-PCR codes from genomic DNA and importantly it provides a male specific marker, which would be of immense value in the analysis of rape samples where there is a mixture of the victim’s and the assailant’s DNA. Analysis of mutation at this locus will prove interesting since no interallelic interactions (except for sister chromatid exchanges) can occur.

**Proposed mechanisms of minisatellite mutation**

**Recovery of mutation events.** In parallel with the creation of MS32 transgenic mice for analysis, a great deal of effort was put into the analysis of the endogenous MS32 locus in humans by MVR-PCR. Large scale analysis of blood DNA samples from different population groups and analysis of lymphoblastoid DNAs from CEPH families have enabled the creation of an MS32 diploid code and single allele database containing over 1000 codes with examples from many different population groups (A.J. Jeffreys personal communication). Similar, smaller, analyses were also performed for the MS31A locus (Neil et al., 1993) and the MS205 locus (Armour et al., 1993). De novo mutation events were identified by pedigree analysis (Jeffreys et al., 1988, 1991b; Royle et al., 1990) although in general the recovery of minisatellite mutations by pedigree analysis was time consuming and did not give individual allele mutation rates. To identify a greater number of mutation events at the MS32 locus, for analysis, the new technique of small pool PCR (SP-PCR) was developed (Jeffreys et al., 1994). This enabled the recovery of an unlimited supply of de novo mutation events from the same individual for analysis by MVR-PCR. This technique is currently being applied to the MS205 locus (J. Armour and C. May, personal communication) and there are plans to extend it for analysis of the MS31A locus (D. Neil personal communication). The majority of the mutations studied were germline events, but the SP-PCR technique additionally enabled the recovery of some somatic mutation events for comparison of the processes of mutation between possibly meiotic germline processes and mitotic somatic mutation events (Jeffreys et al., 1994).

**Paternal bias of mutation.** A number of the minisatellite loci commonly used in paternity testing, show a bias of paternal mutation events, identified by pedigree analysis, including MS31A (D. Neil 1994), MS205 (Jeffreys et al., 1994, J. Armour personal communication) MS43A and \( p\lambda g3 \) (J. Henske, unpublished data). This phenomenon is also observed at the human minisatellite locus CEB1, where an extreme 15% male mutation rate is observed in contrast to a 0.3% female mutation rate (Vergnaud et al., 1991), and at the mouse minisatellite locus Ms6-hm (Kelly et al., 1989). It is possible that some of these mutation events may arise because of the greater number of germline mitoses in the male, however this would suggest a predominantly mitotic mutation process, which is not...
predicted from observations at the MS32 locus (see below and Jeffreys et al., 1991b, 1994). Elevated male germline mutation rates may result from the genomic position of minisatellite loci. This has been inferred by the observed expansion of sub-telomeric regions in male linkage maps, possibly connected with elevated levels of male-specific recombination in these regions (Nakamura et al. 1988, 1989). Currently the observed mutations at MS32 do not show any sex specific bias, however MS32 is not used in paternity testing, since the commonly used enzymes in such analyses, (Hinf I and Hae III) destroy the MS32 locus. Therefore relatively few mutation events have been studied and it is possible that a sex-specific bias has not been observed due to the small numbers of pedigree mutations analysed, only seven to date (Jeffreys et al., 1991b, 1994).

Characterisation of germline mutation events at three minisatellite loci. MVR-PCR was used to identify the type and possible source of de novo mutation events at the minisatellite loci MS32, MS31A and MS205 (Jeffreys et al., 1991b, 1994). Comparison of the structures of mutant alleles in children (identified during pedigree analysis) with their parental alleles and of mutant alleles detected by SP-PCR at MS32 with the progenitor alleles enabled the molecular dissection of the mutation events to give clues about the mechanisms creating variability at minisatellite loci.

Polarity of variation. Previously allele alignment analysis of both the MS32 and MS205 loci and to a lesser extent the MS31A locus, had indicated that there was a polarity of variation at one end of the alleles, designated as the 5' end; in groups of alleles with closely related MVR-PCR maps the differences between alleles had a tendency to occur at this variable end of the allele (Monckton 1992). When mutation events first identified from pedigree analysis were recovered and analysed by MVR-PCR they confirmed the polarity of mutation at the 5' end, as predicted from the MVR-PCR map variation patterns (Jeffreys et al., 1991b, 1994; Armour et al., 1993). In addition SP-PCR analysis at the MS32 locus showed that 90% of the mutant alleles which had gained repeat units, showed extreme polarity (Jeffreys et al., 1994). Interestingly deletion mutations identified both by SP-PCR and by pedigree analysis, also seem to show polarity in their breakpoints.

Mutant size gain bias. The majority of the mutation events identified both by pedigree and SP-PCR analysis were the result of an increase in the number of minisatellite repeat units. Of seven paternal mutation events identified by pedigree analysis at MS31A, all were gain mutations ranging from 5-22 repeats added at the 5' end of the allele. Only one female mutation event was observed, a deletion of 52 repeat units (Jeffreys et al., 1994). Pedigree analysis of MS205 revealed three small repeat gain mutations showing strong polarity and two deletion events, one with 23 repeat units lost (Jeffreys et al., 1994). The increases at MS32 identified in pedigrees and by SP-PCR tended to be small, and commonly involved the addition of a few repeats near the hypervariable end of the alleles. SP-PCR analysis of the MS32 locus showed that 93% of the sperm mutation events were changes of up to 20 repeat units and that 74% of these were gain mutations. This bias of size gains of minisatellite repeat units invalidates previous computer modelling of mutation processes involved in minisatellite evolution (Gray and Jeffreys 1991) which were based on the assumption that minisatellite repeat units are lost and gained with the same frequency. Instead of previous predictions for minisatellite evolution it is now apparent that a size gain bias will greatly accelerate the evolutionary expansion of tandem repeated arrays. At MS32 it is estimated that the observed gain bias would result in the size of alleles growing at a mean rate of 1 repeat unit per 43 generations (~1000 years)(Jeffreys et al., 1994). Currently it is not known if and how these tandem repeated arrays are prevented from
expanding indefinitely. It is possible that there is selection against any abnormal chromosome structure or gene function that may be caused by overlarge expansions of minisatellites, or it may be the result of a size related instability for deletion events. A few large deletions at long tandem repeat arrays could potentially counteract the effects of more frequent small gains of repeat units. It is possible that such deletions could be initiated by random double strand breaks (DSBs) which might induce a collapse of the repeat unit array. Such random DSBs would be likely to increase in frequency with the length of the repeated DNA and are therefore more likely to occur in long tandem arrays.

**Evidence for interallelic events at minisatellite loci.** In almost half of the male germline mutation events analysed at MS32, MS31A and MS205 there is evidence for the interallelic transfer of minisatellite repeat units during the creation of the mutant allele (Jeffreys et al., 1994). The process is non-conservative and both alleles are able to gain or lose repeat units. As yet no flanking marker exchanges have been associated with MS32 mutation events, confirming earlier predictions that interallelic unequal exchange of repeat unit blocks is not a major mechanism of mutation at these loci. The type of repeat unit transfer observed bears the hallmarks of typical gene conversion events. Anomalous regions of repeat units of unclear origin were also present in both pedigree derived paternal mutations and SP-PCR recovered mutation events at MS32. Of the four maternal mutations identified by pedigree analysis at MS32, none showed evidence of interallelic exchange, but rather extra repeat units appeared to result from intrallelic duplication events. At MS31A most of the exchanges involved the transfer of a segment of repeat units from the donor allele into an internal site within the recipient allele, close to the 5' end of the repeat unit array. None of these mutations involved the exchange of flanking DNA, markers suggesting again a gene conversion-like process. However, recent evidence (D.Neil personal communication) has identified one MS31A mutation event where a large deletion is associated with both interallelic repeat unit transfer and the exchange of the appropriate flanking markers to suggest that it may be the result of an unequal exchange event. Only one of the mutants identified by pedigree analysis at MS205 showed evidence of an interallelic transfer, however it was not possible to analyse for exchange of flanking markers at this individual due to the homozygosity of the father (Jeffreys et al., 1994).

Putative mutation mechanism for minisatellite loci. Similar studies at the CEB1 (D2S90) locus (Buard and Vergnaud 1994), have also shown a 25% frequency of gain mutations with polarised interallelic transfer, without the exchange of flanking markers. Observations of interallelic repeat unit transfer imply that the recipient allele is in some way activated for mutation before transfer. The correlation of de novo mutation events by interallelic transfer with a polarity of mutation at the 5' end of the repeat unit array, suggests that this activation process may be modulated by elements active upon or near the 5' region of the array. One simple and speculative model for minisatellite mutation, (sometimes referred to as the gap expansion model, GEM), was proposed by Jeffreys et al (1994) and involves mutational activation by the introduction of a double stranded break (DSB) near the beginning of a tandem repeat array. The proposal is that the positioning of this DSB is directed by an initiator 5' of the array (figure 9.1) and that gains of repeat units can result if the free ends of such a break diffuse apart. The resulting gap could be bridged by strand invasion from either the sister chromatid or the homologous allele, followed by the repair of the single-stranded gap.
Figure 9.1. A speculative model for minisatellite mutation.

The recipient allele is shaded red, and the donor allele is shaded black. Flanking DNA is diagonally striped, repeat units are in full colour.

A. The 'initiator' factor binds to the 5' flanking DNA of the minisatellite, and introduces a DSB in the repeat units close to the start of the repeat array.
B. The DSB expands to form a gap and exonuclease activity creates single-stranded 5' overhangs. The donor allele pairs with the recipient allele.
C. The gap is bridged by strand invasion.
D. The resulting single-stranded gap is filled by repair synthesis (the blue repeats) and the conversion complex is resolved, perhaps by the extrusion of the donor allele strand from the recipient allele.
E. Short domains of mismatch repair over the heteroduplex created around the filled gap result in the scrambling of donor and recipient repeat units near the donor segment.

This figure was taken from Jeffreys et al (1994) and adapted by D. Neil (1994).
Deletion events would be expected to result from an alternative mutation pathway, but potentially also involving the action of DSBs. If the free ends of a DSB diffused laterally together prior to misalignment with the homologous chromosome or sister chromatid, then a deletion or repeat unit 'collapse' could occur. Additionally in vitro experiments have been used to show that DSBs in mitotic cells are preferentially repaired via a non-conservative single stranded annealing (SSA) mechanism (Lin et al., 1984). During this process 5' exonucleases digest the free 5' strand adjacent to a DSB resulting in a 3' overhang. Next regions of intramolecular homology can align and any non-homologous DNA is excised. Duplex DNA is regenerated by gap filling DNA synthesis. The repeat units of minisatellite loci would provide regions of internal homology and the potential for high levels of SSA mediated repeat unit collapse resulting from random interstitial DSBs is presumably high. Interestingly few of the deletion germline mutations analysed at MS32 showed evidence for interallelic exchange, but often showed the simple loss of runs of contiguous repeats (Jeffreys et al., 1994).

By definition somatic mutation events must be mitotic in origin and it is interesting to look at the MVR maps of these mutants in comparison with those from germline mutation events. SP-PCR was performed upon the MS32 locus directly in blood cell DNA (Jeffreys et al., 1994). In contrast to the mean germline mutation rate of 0.8% per sperm for MS32, the somatic mutation rate was shown to be greatly reduced at this locus (0.06% per cell). This value was consistent with earlier studies which indicated that somatic mutation events are rarer, occurring <$1x10^{-5}$ per mitosis (Armour et al., 1989). Therefore if mitotic events do occur in the male germline they can only make up a small proportion of the mutation events identified from sperm DNA. This suggests that the majority of male germline mutation events arise during gametogenesis, possibly during meiosis (Jeffreys et al., 1988a). As predicted from a meiotic process of mutation almost all of the MS32 sperm mutations recovered were different, with the exception of a few examples of mutational mosaicism where the mutants were non-polar. Either these arose before meiosis, or represent "preferred" mutation events (Jeffreys et al., 1994; Monckton et al., 1994). Structural analysis of the blood mutants of MS32 recovered by SP-PCR showed no evidence for a size gain bias. No anomalous repeats, of the type thought to originate from complex gene conversion-type processes, were observed and most of the events appeared to result from simple intrallelic loss of repeat units (Jeffreys et al., 1994). Limited data means that it is difficult to draw conclusions as to the level of mutational polarity in the somatic process. However, it can be concluded from current data that at least some of the germline mutation events the MS32 locus arise from a different mutation process than the somatic mutation events, possibly meiosis specific.

The conversion model shown in figure 9.1, cannot of course explain the mechanism behind many of the more complex events at minisatellite loci, for example the presence of anomalous repeat units for whom no intra- or interallelic donor is apparent. Many of the mutation events appear to involve intrallelic duplications, often in combination with interallelic exchange. This suggests that a complex multi-step process may be occurring, perhaps due to extra breaks being introduced into the conversion complex (Jeffreys et al., 1994) and a role for replication slippage in introducing some of the duplication events cannot be ruled out. Knowledge of the mutation processes operating on the endogenous MS32 locus provide valuable information for making comparisons with the mutation status of the MS32 transgenic loci. This enables assessment of whether the transgenic MS32 loci provide a good model for in depth analysis of the factors modulating the human minisatellite mutation processes. For example do they display the mutational polarity, interallelic repeat transfer and sperm specificity of mutation observed at the endogenous locus.

Chapter 9, page 6
**MS32 transgenic mice**

**Creation of transgenic animals.** In this study we have demonstrated that it is possible to create mice transgenic for a hypervariable human minisatellite locus and to achieve germline transmission of this transgene insertion. The process of transgene construct integration was revealed to be far from simple, resulting in eight integrant structures, all differing from the original construct and to involve a significant level of MS32 transgene construct insertion into mouse gamma satellite DNA, the major mouse satellite sequence (see chapter 5). Structural analysis of the eight separate insertions revealed that three of the integrations had involved the insertion of more than one copy of the original construct. These were referred to as multi-copy loci, in contrast to the five insertions comprising a single copy of the original construct, often rearranged or truncated at the 5' or 3' terminus.

**Transgene mutation processes.** Subsequent mutation analysis of six of these MS32 transgenic loci revealed that the mutation behaviour of the multi-copy and single-copy transgenes differed. Mutation rates at the single-copy loci 110D and 110C were measured by SP-PCR to be 0.007% and 0.04% per sperm respectively. These values were greatly reduced from those seen at the endogenous human locus of 0.8% (from SP-PCR) and 1.0% (from pedigree analysis). In addition, mutational polarity was not obvious in the small number of mutation events that were recovered from the 110D locus for analysis by MVR-PCR. However, some events involved the introduction of anomalous repeat units perhaps by complex gene conversion-type events (see figure 6.3, chapter 6). One very interesting mutation was obtained for analysis by MVR-PCR from the 110C locus. This showed a clear gene conversion event with the upstream block of repeats acting as a donor into the downstream repeat unit block (see figure 6.4, chapter 6). All of the single-copy mutation events analysed by MVR-PCR showed evidence for more than one type of mutation process, for example repeat unit deletion or duplication, in conjunction with the gain of anomalous repeat units or repeat unit type switches.

The multi-copy transgenes showed a different pattern of mutation. The loci 109 and 110A had mutation rates high enough to be measured by pedigree analysis. The germline mutation rates at these loci were 4.4% and 3.1% respectively. These values were much higher than the rate observed at the endogenous human MS32 locus, which may have been a reflection of the fact that the 109 locus was composed of six construct repeats and the 110A locus had eleven construct repeats. However, although the mutation events occurred in different construct blocks throughout the 110A array, all 109 mutation events were limited to the region between the second and third repeat unit blocks. In addition both multi-copy loci showed high levels of somatic mosaicism for mutant alleles, 109 (4.4% of progeny studied) and 110A (1.9%). One germline 110A mutation was recovered and analysed by MVR-PCR, revealing a seven repeat unit duplication at the non-variable end of the original construct. However, the complex structure of the two multi-copy loci and the fact that SP-PCR was unsuccessful at these loci, meant that mutation analyses were restricted to changes in gross transgene locus structure rather than MVR-PCR analysis of mutant repeat unit arrays. This severely limited any comparisons of repeat unit array mutation processes at the multi-copy loci with those observed at the endogenous MS32 locus.

Recent evidence from restriction mapping data and structural analysis by PCR, of the progenitor 109 locus and some 109 mutants, by A. Collick and J. Drew, has suggested that an inverted transgene construct structure exists at this locus, between the second and third construct blocks in the array (Figure 9.2). This has the effect of creating a
Figure 9.2. Schematic of the transgene 109 multi-copy locus showing the position of the inverted transgene constructs.

The 5' MS32 repeat array flanking DNA is shown as a green box and the 3' flanking DNA as a red box, whilst the MS32 repeat unit array is represented by gray shaded boxes. The restriction enzyme sites shown on the progenitor transgene 109 locus and 109 mutants are Ssp I (S) and Hind III (H). The order number of the six transgene construct blocks from the 5' end of the multi-copy array is indicated above each block.

Recent evidence from restriction and PCR analysis of the structure of the transgene 109 progenitor and mutant alleles by A. Collick and J. Drew (data not shown), indicated that final four copies of the original transgene construct in the array are inverted with respect to the first two copies. This is shown in this schematic by the orientation of the black arrows beneath the individual transgene constructs in the multi-copy array of the progenitor allele. The arrow heads represent the 3' MS32 flanking sequence. The inverted nature of this sequence means that there is a construct 'tail to tail' array of the MS32 3' flanking sequence between the second and third constructs in the array.

Schematics of the three mutation events described in chapter 6 (figures 6.8-6.10) are also shown indicating the flanking DNA status between the second and third transgene construct blocks in the array. The 745 mutation has only the 3' flanking DNA from the inverted third construct in the array, the 709 mutation has no 3' flanking MS32 DNA at this position and the inverted repeat of the 3' flanking DNA in the type II mutation has been changed by a small deletion event (A. Collick personal communication, data not shown).
Chapter 9, figure 9.2

**original construct**

5' flanking DNA (green)  MS32 repeats  3' flanking DNA (red)

1. 2.5
2. 2.2
3. 3.7
4. 2.5
5. 3.9
6. 2.5

**Progenitor 109**

inverted repeat of MS32 3' flanking DNA

**745 mutation**

**709 mutation**

**Type II/missing site mutation**
palindromic sequence of MS32 3' flanking DNA and MS32 repeat units at this position. It is possible that this type of sequence may form tertiary structures, perhaps affecting mutation processes at the 109 locus. This is inferred by the suggestion, from two observations, that it is this region of the multi-copy array that appears to be important in the 109 mutation process. First, all 109 mutation events characterised so far (19 out of 20) have one 'breakpoint' in this region, secondly, recent evidence from A.Colllick and J. Drew has shown that once changes are made to this region as in the 109 mutants 745, 709 and 'type II' mutants (figure 9.2) these alleles no longer mutate. 9.8%, of progeny from parents carrying the 'progenitor' 109 allele are mutant, however mutation rates for progeny from these mutant loci are; 745 (0/63 alleles tested), 709 (0/113) and 'type II' (0/155). Therefore it could be hypothesised that once this palindromic sequence is disrupted, (709 and 745 have lost all or part of the MS32 flanking DNA, and a small deletion of 3' flanking DNA has occurred in the 'type II' mutant), that tertiary structures cannot form and instigate mutation processes. This suggests that the inverted construct repeat structure may play a role in some mutation events at the 109 locus.

A good model of endogenous MS32 mutation processes? Minisatellite mutation is not restricted to the germline, since it has been previously shown to occur in clonal tumor cell populations (Armour et al., 1989) and by single molecule analysis of normal somatic DNA (Jeffreys et al., 1990), but it is not a common process at minisatellite loci and the only known example of a bona fide somatic mosaicism for a mutant allele, identified by Southern blot analysis is at the MS31A locus (D. Neil personal communication). For this reason the observation of extremely high somatic mutation rates at the multi-copy loci suggests that they are not mutating in an analogous fashion to the endogenous human MS32 locus.

The structure of the single-copy MS32 transgene loci was more similar to that of the endogenous locus, which initially suggested that they might provide a better model of human minisatellite mutation. There are some processes that occur at the transgene loci, for example duplication events and regions of anomalous repeat unit gain that also appear to occur at the endogenous human MS32 locus (Jeffreys et al., 1994). However, the rate of mutation at the single-copy loci is far below that observed at the endogenous locus indicating that if mutation processes are shared they are not initiated at the same rate. In addition where transgene mutations were recovered from the single-copy loci 110D and 110C, and analysed by MVR-PCR, they did not show the property of polarity toward the 5' end of the MS32 repeat unit array, as observed with the majority of human MS32 mutation events. Since the 110C and 110D transgenes were only analysed as hemizygous loci, it is obvious that no interallelic exchange from a 'donor' allele, as indicated in the model of minisatellite mutation (figure 9.1) could occur. Strand invasion of DSB gaps by sister chromatids was also predicted as a possible 'donor' mechanism for some mutation events, and the repeat unit duplications observed at some of the recovered 110D mutants may have arisen by this way, if not by strand slippage during replication.

These data suggest that the transgenic loci analysed to date do not provide a good model of the mutation processes at the endogenous human MS32 locus and it is difficult to extrapolate the information about mutation processes at these loci to the processes occurring at the human locus. Nevertheless the multi-copy loci mutate at a very high rate, providing a wealth of interesting mutation behaviours for study and these loci have provided a fascinating insight into the properties of integration and subsequent mutation processes of tandemly repeated DNA transgenes.
Hotspots of meiotic recombination in yeast. Minisatellite loci have previously been suggested as being 'hotspots' of meiotic recombination (Jarman and Wells 1989). Other meiotic recombination hotspots have been characterised in yeast. An initiation site (IS) for meiotic gene conversion is located in the promotor region of the ARG4 locus on Chromosome 8 in *Saccharomyces cerevisiae* and the initiation site for the introduction of DSBs has been localised to a 14bp poly-A sequence (Scultes and Szostak 1991; Sun et al., 1991). Nested deletion analysis of the DNA sequences flanking the 'IS' indicated that the ARG4 upstream region is composed of three regions involved in the control of DSB induction, and that mutating these upstream regions can lead to a reduction in the number of DSBs induced (De Massey and Nicolas 1993). This indicates that both the position and frequency of DSBs in *S.cerevisiae* are controlled *in cis*. The hotspot mutation *ade6-M26* in *Schizosaccharomyces pombe*, is a G to T transversion close to the 5' end of the open reading frame (Szankasi et al., 1988) and shows a 10-fold stimulation of conversion. However, the trans-placement of parts or all of the *ade6* gene (containing M26) to the other arm of chromosome 3 leads to a loss of the enhanced meiotic recombination effects (Ponticelli and Smith 1992). This has parallels with the observation of a loss of sperm-specific interallelic mutation events at the MS32 locus when it is 'removed' from its endogenous human chromosome and inserted into a new genome, and suggests that other factors in addition to *cis*-active flanking DNA, such as chromatin structure, may play a role in MS32 mutation processes.

Evidence for a *cis* - acting mutation initiator at the MS32 locus.

Influence of MS32 flanking DNA? High variability is created at minisatellites by a high germline mutation rate to produce new length alleles. Structural analysis of MS32 allele mutations revealed that mutations were largely confined to the 5' end of the array. Furthermore SP-PCR analysis of MS32 alleles ranging in size from 23 to >160 repeat units showed a mean mutation rate per gamete of 0.8%, irrespective of the total repeat unit array length (Jeffreys et al., 1994). These processes of mutational polarity and mutation rate constancy, independent of repeat array length, suggested that minisatellite mutation at MS32 is regulated by elements outside the array, possibly acting in the flanking DNA adjacent to the hypervariable end of the repeat array.

One of the initial aims of the transgenic MS32 project was to assess the effect of flanking DNA sequence, either human MS32 or mouse, on mutation processes. For this reason ~200bp of 5' and 3' MS32 human flanking DNA sequence were included in the original transgene construct. It was hoped that this would provide enough flanking sequence information to convey the property of mutational polarity to the transgenic loci. The results of these studies have been inconclusive, due to the small number of mutation events available for analysis by MVR-PCR. No polarity has been observed in any of the transgene mutation events analysed to date, although it remains possible that new transgene constructs incorporating different amounts of MS32 flanking DNA may be able to reproduce the mutational polarity observed at the endogenous locus. If such a scenario was achieved then a systematic mutagenesis of the flanking DNA required to confer polarity might enable identification of the crucial sequence motifs required for this process.

An alternative method of identifying possible important flanking DNA regions directly in man, was performed in parallel to the transgenic mouse project. This involved the search for naturally occurring polymorphisms in the flanking DNA. Some of these have been useful for flanking haplotype studies, to look at minisatellite evolution in
different population groups and also to analyse for the exchange of flanking markers in mutation events (Monckton 1992; Nell and Jeffreys 1993; Armour et al., 1993; Jeffreys et al., 1994).

The flanking O1C variant. During this analysis a variant position associated with alleles from a particular population group, alleles of an African origin, was identified (Monckton et al., 1994). A group of these alleles were atypical since they were found many times in this population group, resulting in an overall reduced allelic diversity in African alleles. Sequencing of the 5' MS32 flanking DNA in these alleles revealed a G to C transversion 48bp upstream of the MS32 repeat unit array and 16bp upstream of a diverged MS32 repeat, preceding the array. This was termed the O1C site and further investigation showed that the O1C polymorphism was not observed in any Japanese alleles and very rarely in Caucasians (frequency ~0.004) (Monckton et al., 1994 and see figure 3.6, chapter 3).

Reduced mutation rate at O1C alleles. As predicted from the low level of O1C allele diversity, the mutation rate at O1C alleles was shown to be reduced by SP-PCR analysis of a number of O1C alleles from Afro-Caribbean and Zimbabwean sperm DNA and from two Caucasian O1C examples (Monckton et al., 1994). The mutation rate for O1G alleles, for 3-20 repeat unit gain mutations was 0.13-0.88% per sperm in comparison with the O1C allele rate of 0.00-0.02% per sperm, with the exception of one O1C linked allele which showed a mutation rate of 0.12%. Although higher than the rate of the other O1C alleles this was nevertheless a low mutation rate compared with the average for O1G alleles. There was a preferential suppression of repeat unit gain mutation events. In contrast to O1G alleles where the majority of mutations are the result of small gains of repeat units, small deletions were more common than gains for the O1C alleles. The mutation suppression at the O1C allele did not act in trans at O1G alleles as shown for an O1C/O1G heterozygote individual, where the O1G allele showed a normal mutation rate. Importantly it was observed that even though it was not mutating itself, the O1C allele in a O1C/O1G heterozygote individual, could act as a donor in interallelic mutation events at the O1G allele. This emphasises the non-conservative nature of the repeat unit transfer process. The set of O1C alleles identified, do not all share the same MVR codes or flanking haplotypes, but all share a reduced mutation rate, indicating that it is the O1C polymorphism in the 5' MS32 flanking DNA that is associated with a reduced mutation rate.

The discovery of a group of stable MS32 alleles suggests that interallelic polar mutation processes are not intrinsic to minisatellite tandem repeat arrays and is consistent with the involvement of a cis-acting mutation initiator as proposed in figure 9.1. Although it seems remarkable that a single base transversion can be enough to result in mutation suppression, it is possible that this transversion inactivates some component of a mutation initiator element. In this scenario at least some part of the mutation initiator must associate with the flanking DNA very close to the mutational hotspot in the repeat unit array. In this way the O1G site could actually form part of the initiator element possibly providing a binding site for a protein that activates mutation. It is also possible that the transversion event might allow another protein to bind to this site in preference to the mutation initiator protein, thus blocking the effects of the mutation initiator.

The sequence surrounding the O1C polymorphic position is interesting since it forms an 18bp same strand palindrome;

TTGGTTGGA//AGGGTGTTT
A transversion event to the O1C variant disrupts this palindrome, it remains to be proved whether this has the effect of destroying a binding site for a mutation initiator factor.

MVR-PCR analysis of the few mutation events recovered from O1C alleles showed that mutational polarity was less prevalent in these alleles, and that the mutations occurring may involve other processes such as replication slippage and deletion. The reduced rate of mutations would suggest that these were not usually important mechanisms of mutation at minisatellite loci, but are simply residual mechanisms, when the major mutation process is 'blocked'. The types of mutation seen at low rates (0.007% and 0.04%) at the single-copy MS32 transgenic loci involved small duplications and deletions. Therefore it is possible that the low mutation rates in these transgenes were as a result of a failure of the 'normal' processes of mutation, initiated by cis-acting factors at the endogenous OIG locus. Since all the mouse transgenes were demonstrated to have OIG variants, it is possible that either the mutation initiator element is not conserved in the mouse genome, or that additional regions of the human MS32 flanking DNA, than those included in the transgene construct, are required to initiate mutation.

Minisatellite binding proteins. A number of proteins that bind to G/C-rich tandem repeat sequences have been identified (Collick and Jeffreys 1990; Wahls et al., 1991; Collick et al., 1991; Yazamaki et al., 1992). The protein Msbp-1 was isolated by using synthetic binding substrates with homology to the minisatellite G/C-rich 'core' sequence. It has been shown to specifically bind to many single-stranded G-rich repeats (Collick et al., 1991). Minisatellite binding protein activity may be a reason for the conservation of the minisatellite 'core' sequence. It is possible that single-stranded DNA binding proteins might bind to a single-stranded 'core' sequence after the introduction of a DSB into a minisatellite repeat array. It has been speculated that such a protein might act to stabilise the G-rich strand to enable strand invasion and the formation of a gene conversion heteroduplex (Neil 1994), but as yet firm evidence for a functional role of either these proteins or the minisatellite 'core' sequence, remains to be established. The putative minisatellite mutation initiator element (figure 9.1) appears to associate with the 5' MS32 flanking DNA and as such would not be described as a minisatellite repeat unit binding protein. It remains to be seen from protein/DNA association studies at the MS32 locus if this putative protein exists and binds at the OIG site, and indeed if it modulates the interactions of other proteins, including minisatellite binding proteins.

A comparison of the mutation processes at minisatellite loci with those observed at STR loci

In addition to the turnover of minisatellite tandem repeats another area currently of widespread interest is the mechanisms of mutation at STR loci, in particular the mechanisms of mutation that result in the expansions of triplet repeat arrays associated with human genetic disease. It is of great interest to discover if any of the processes operating at minisatellite loci also play a role in generating polymorphism at shorter, simpler, repeat units. By observing the properties of mutation events at STR loci, we can draw clues as to the most likely mechanisms underlying instability at these loci in the same way that early observations of polymorphic minisatellite loci implied possible mechanisms of mutation. Most of the data currently available relates to the association of STR instability with human genetic disease. Although the processes behind this instability may not reflect the processes at all STR loci, there is a growing body of evidence that STR instability is due to specific processes that are similar to those operating at other tandem repeat loci.
loci, particularly less variable ones, they can help in unravelling the possible mechanisms of generating polymorphism at these loci.

**Bias toward size increases at triplet repeat disease loci?** The triplet repeat arrays associated with human disease, share a propensity towards an increasing array size, as observed at minisatellite loci, although the level of repeat unit expansions observed at these loci far exceeds the rate at which minisatellite loci are predicted to expand in size (1 repeat unit every 43 generations, Jeffreys *et al.*, 1994). However, it is not yet apparent if this bias to repeat gain mutations is seen at less polymorphic and non-disease associated STR loci. In addition there is currently no data on the instability of disease associated triplet repeats in the normal size range. It would be possible to assay for triplet repeat instability in the sperm and somatic DNA from such individuals using SP-PCR to measure the rate at which normal alleles expand. Some of the expansions observed in triplet repeat arrays, for example those adjacent to the at the FRAXA, FRAXE and MD can be massive, for example an increase in array size from 200-2000 repeat units in fragile X and Myotonic Dystrophy. Neither of these repeat arrays are within the coding region of the genes they affect and this might explain why their expansions are huge in comparison with coding triplet repeats, such as the CAG repeat associated with Huntington's disease where selection against expressed long tandem arrays may exist.

For many human disease associated triplet repeats there is evidence for an intermediate expanded allele state, often referred to as the 'premutation allele' (reviewed by Mandel 1993). For example, the CAG repeat at the Myotonic Dystrophy locus in normal individuals has an array length of 20-35 repeat units. Some individuals carry the 'premutation' allele with an array of 50-100 repeats and when this is expanded to 100-2000 repeat units the disease phenotype is manifested. This term 'dynamic mutation' was coined to describe this type of multi-step mutation process at triplet repeat loci (Richards and Sutherland 1992) and can explain the phenomenon of anticipation (disease symptoms becoming progressively worse over generations) observed for some of these diseases.

**A polarity of variation?** At the spinocerebellar ataxia type I locus, 98% of normal alleles were shown to have interruption of their CAG repeat by CAT repeats, whilst disease alleles consisted entirely of CAG repeats (Chung *et al.*, 1993). The suggestion is that variant repeat units may somehow act as 'anchors' to minimise replication slippage, since pure CAG repeats would have a higher probability of slipped strand mispairing during replication with the subsequent expansion (or contraction) of the repeat unit array. Recent haplotype analyses, using microsatellite markers flanking the FRAXA repeat array, suggest that fragile X alleles may have arisen from a limited number of ancestral chromosomes (Oudet *et al.*, 1993; Zhong *et al.*, 1993; MacPherson *et al.*, 1994; Kunst and Warren 1994). There is an ATG variant of the major CGG triplet repeat at FRAXA (Fu *et al.*, 1991). Uninterrupted (CGG)n arrays are more commonly expanded than (CGG)n interspersed with ATG repeats (Snow *et al.*, 1994) and sequence analysis has shown that pure (CGG)n runs appear predominantly at one end of FRAXA alleles (Snow *et al.*, 1994; Kunst and Warren 1994), and that within pedigrees, expansions of repeat units occur at the same end of the allele. This suggests a possible polarity in the expansion of the FRAXA repeat array, perhaps mediated simply by the stabilising effects of ATG repeats at the non-expanding end of the allele on controlling slippage processes, or perhaps modulated by some cis-acting flanking sequences involved in modulating more complex conversion events, in a similar manner to the controls operating at the MS32 locus.
Sex bias of mutation? There is evidence for sex specific effects on the level of expansion at some disease associated triplet repeat loci. For example, large expansions of the triplet locus associated with Huntington's disease (HD), spinal-bulbar muscular atrophy (SBMA) and spinocerebellar ataxia type I (SCA1), are seen primarily after male germline transmission (Duyao et al., 1993; Banfi et al., 1994 and reviewed Mandel 1994) and in Fragile X (FRAXA), large expansions to the repeat arrays resulting in the disease phenotype, are exclusively limited to female germline transmission. There is evidence for a paternal bias of mutation at minisatellite loci such as MS31A, MS205 and CEB1 (Jeffreys et al., 1994; Vergnand et al., 1991). The greater number of mitotic divisions in the male germline inferred that mitotic processes of mutation may be responsible for some of the mutation events at these loci. Recent evidence from MS32 suggested that meiotic mutation processes were probably more prevalent at minisatellite loci (Jeffreys et al., 1994 and chapter 6) although it remains possible that the expansions at the HD, SCA1 and SBMA triplet repeat loci are caused primarily by mitotic mutation processes. The expansions at these triplets are smaller than those observed at the non-coding triplet repeats associated with Fragile X (FRAXA) and Myotonic Dystrophy (DM). This suggests that there may be different mutation processes operating on different triplet repeat loci. For example, discontinuous gene-conversion-type events have been reported to result in a decrease of triplet repeat units at the DM repeat and from a FRAXA repeat (O'Hoy et al., 1993; Van den Ouweland et al., 1994). This is a process associated with minisatellite mutation (Jeffreys et al., 1991) and it is possible that the same process might act to increase triplet repeat unit size.

Somatic instability at triplet repeat loci? The triplet repeat expansions identified in many disorders exhibit remarkable variation both within the cells of a single individual (particularly in FRAXA and DM) and also between members of a sibship (Fu et al., 1991; Brook et al., 1992). These have been loosely interpreted as the result of a mixture of mitotic and meiotic mutation processes. Mitotic instability is easily observed as allelic variants in the somatic tissues of an individual, although it is more difficult to conclusively identify meiotic instability. This is assumed to be represented by instability during the transmission of a premutation allele from a parent to an affected sibling. However since it is generally DNA from leukocytes that is analysed it is impossible to say if the expanded premutation in the child is a result of very early embryonic instability. SP-PCR of sperm and blood DNA from individuals carrying a premutation will enable assessment of whether instability is occurring in the male germline. At the FRAXA triplet repeat locus somatic mosaicism has not been observed in individuals carrying the premutation allele, but is commonly observed in the tissues of individuals carrying a fully expanded allele (Wohrle et al., 1992). This phenomenon was assumed to represent a continuous mitotic instability, over a threshold number of repeat units. However, more recent analysis of mitotic stability of Fragile X mutations in differentiated cell lines, showed that the pattern of different repeat lengths were transmitted to progeny cells in a stable fashion (Wohrle et al., 1993) and analysis of repeat array lengths in different tissues of fragile X patients showed that the expanded allele was not present in every tissue (Reyniers et al., 1993).This suggested that the genotypic mosaicism in these patients was generated post-conceptionally, and that parents would transmit the premutation FRAXA allele and that the expansion would occur in a very early time window of development (Wohrle et al., 1993). Expansion of the premutation to the full mutation is maternal specific. This may reflect the fact that in the early embryo the sperm contributes only DNA material, whilst the oocyte cellular transcription and replication machinery is maintained. It can therefore be hypothesized that maternally inherited premutation alleles might be expanded by 'initiators' of expansion existing in the oocyte and maintained in the early embryo.
Human minisatellites are generally assumed to be relatively somatically stable (Armour et al., 1989; Jeffreys et al., 1990, 1994) and do not show the widespread somatic instability observed at some triplet repeat loci. However, the mouse minisatellites Ms6-hm and Hm-2, which have the repeat unit sequences GGGCA and GGCA respectively, have been demonstrated as being highly unstable in an early embryonic window of development (Kelly et al., 1989, 1990; Gibbs et al., 1993). Although the repeat unit arrays of the mouse minisatellites are, at several kb, much larger than the triplet repeat loci, there may be some similarities in the processes of somatic mutation, perhaps the similarity in sequence between the mouse minisatellites and the typical CAG expanded repeats in humans, plays a role in their somatic instability.

Trans-acting factors. In contrast to evidence for cis-acting modulators of minisatellite mutation, there is evidence at STR loci for the influence of trans-acting factors in generating instability at these loci. Instability at STR loci has recently been associated with multiple forms of cancer (reviewed Richards and Sutherland 1994). As discussed in chapter 8, recent developments in the study of hereditary non-polyposis colon cancer, have indicated that the STR instability in colon tumour DNA from patients with HNPCC, results from the effects in trans of an aberrant mismatch repair protein, hMSH2. This protein is normally part of the human homologue of the bacterial mutHLS DNA repair pathway responsible for the repair of heteroduplex DNA, but in HNPCC individuals it is unable to perform this function. The malfunction of other trans-acting factors may also cause the STR instability manifested by other cancers and perhaps other heritable diseases.

Putative STR mutation mechanisms

Replication slippage. The features of STR polymorphism are consistent with mutation by a replication slippage process. For example, it has been noted that polymorphism in dinucleotide repeats, is proportional to perfect repeat copy number (Weber, 1990), and that instability increases with increased numbers of perfect repeats at the FRAXA and SBMA triplet repeat loci (Kunst and Warren 1994; Snow et al., 1994). Normal slippage processes would be expected to be more frequent in long uninterrupted repeat unit arrays, where there are no 'unique' sequence motifs to 'anchor' the replicating strands. Repeat units are lost or gained in the slippage process by a model where the primer strand transiently dissociates from the DNA template and reanneals in a misaligned configuration. The dissociation may be caused by a single stranded nick in the DNA. If the primer strand misaligns upstream of where it dissociated then a gain of repeat units results and vice versa. Triplet repeats at several disease loci become unstable above a similar size threshold of 40-50 repeats, perhaps above this size single-stranded nicks are more likely to occur in the array during replication initiating the loss or gain of a few repeats by slippage. If two or more single-stranded breaks occur in the same Okasaki fragment of the lagging strand during replication, then expansions involving much greater numbers of repeat units may be able to occur since the fragments of the discontinuously replicating DNA strand are not anchored at either end by non-repeat sequence (Richards and Sutherland 1994).

Unequal recombination. Even though there is a growing assumption that human STR loci mutate predominantly by slippage processes (reviewed in Kunkel 1993; Lustig and Petes 1993; Weils and Sinden 1993; Strand et al., 1993; Richards and Sutherland 1994) unequal exchange cannot be ruled out as a mechanism for generating polymorphism at these loci. The overall repeat unit array lengths at alleles could be altered by unequal sister chromatid exchange or by unequal meiotic exchange. In this type of mechanism two tandem repeat arrays, which
may be located on sister strands of a partially replicated molecule) pair in a misaligned configuration. Crossing over between these arrays results in the deletion of repeat units from one array and a duplication in the sister strand (reviewed in Petes and Hill, 1988). Unequal gene conversion is a related mechanism which can result in an alteration of repeat units by a non-conservative event. At STR repeats this type of event can be difficult to distinguish from replication slippage events, unlike at minisatellite loci where the presence of variant repeat unit types make conversion-type events easier to identify, but there is as yet no evidence to exclude this type of process from taking part in generating instability at shorter less variant repeat units.

**Minisatellites and disease**

Recent evidence has associated particular alleles at some minisatellite loci with an increased frequency of disease. For example the HRAS1 minisatellite is a highly polymorphic minisatellite locus located 1 kb downstream of the proto-oncogene H-ras-1 (Krontiris et al., 1993). It consists of four common progenitor alleles and a much larger number of rare alleles. The rare alleles are derived by mutation from the common four alleles, and a number of the rare alleles have been shown to be associated with many forms of cancer (Krontiris et al., 1993). It is proposed that new mutations of the HRAS1 minisatellite disrupt the controlled expression of nearby genes including the H-ras-1 proto-oncogene, resulting in an elevated rate of cancer. This was inferred by the fact that the HRAS1 minisatellite has been shown to bind members of the rel/NF-kB family of transcriptional regulatory factors *in vitro* (Trepicchio et al., 1992; Krontiris and Green 1994). The suggestion is that a 'saturation' effect is in operation, with the minisatellite allele preferentially binding the transcription factor normally bound by the H-ras-1 proto-oncogene, though it seems remarkable that a simple change in allele length could be solely responsible for such an effect. Another minisatellite locus implicated in human genetic disease is the minisatellite locus 365 bp upstream of the insulin gene. Certain sizes of minisatellite alleles at this locus have been associated with insulin dependent diabetes mellitus (Lucassen et al., 1993). It was thought that since the minisatellite was directly adjacent to the insulin regulatory sequences that different array lengths might affect the binding of the insulin transcription regulator, perhaps by affecting the chromatin structure (Hammond et al., 1992).

There may be a number of minisatellite alleles that are associated with expressed DNA regions and where new alleles can result in disease. Selection pressures against polymorphism at these loci would mean that they have not been identified due to a bias in the study of highly polymorphic minisatellite loci to date. It will be interesting to see if following the discovery of human disease caused by expansions in triplet repeat loci, investigators will analyse more closely the regions of tandemly repeated sequence near to their gene of interest.

**Future work**

**Human minisatellite mutation.** To date, analysis of the best characterised human minisatellite MS32 has revealed a wealth of information as to the mutation processes underlying variability at this locus. MVR-PCR has proved invaluable in determining the extent of allelic diversity and in conjunction with SP-PCR has revealed the properties of mutational polarity, bias of gain mutations and length independence of mutation. The discovery of a flanking DNA polymorphism associated with a reduction in mutation rate was the final piece of evidence for the involvement of a cis-acting element in modulating mutation processes. As yet, no paternal bias of mutation has been observed at MS32, in contrast to the situation at other human minisatellites such as MS31A and MS205. This
might reflect differences in the relative importance of mitotic and meiotic mutation events at these different minisatellite loci. Therefore a logical next step in the analysis of minisatellite mutation is to extend the analyses performed at MS32 using SP-PCR, on other minisatellite loci. This will allow comparative studies between human minisatellite loci. These experiments are currently in the pipeline (D. Neil, personal communication). A recent exciting observation has been the identification of a group of MS31A alleles of Japanese origin, which share some polymorphic characteristics with the groups of non-mutating OIC-linked alleles in Africans; they are small, frequently occurring and show little variability (D. Neil, personal communication). If SP-PCR analysis of these loci reveals that they are mutating at a lower rate then the 'normal' MS31A locus, then it will be appropriate to see if there are any flanking DNA polymorphisms associated specifically with these alleles. This data might result in the identification of a second minisatellite locus where mutation can be shown to be directly influenced by cis-acting factors acting in the flanking DNA.

Environmental influences on minisatellite mutation. Previous studies of DNA fingerprints of the progeny of irradiated mice (Dubrova et al., 1993) have shown an increase in mouse minisatellite allele length changes with an increase in mutation rate. Detailed analysis of individual mutation events was not possible, but these results do demonstrate a possible role for DSBs, (thought to be the mechanism of radiation damage in DNA) in mutation processes at minisatellite loci. Currently blood samples from 150 families accidentally exposed to high levels of radiation during the Chernobyl nuclear reactor disaster in the Ukraine are being analysed (Y. Dubrova, personal communication) using conventional DNA fingerprinting techniques. However it would be exciting to additionally analyse these somatic samples by SP-PCR. Although the SP-PCR experiments performed with irradiated mouse tissues in Chapter 7, were inconclusive, a larger study of the endogenous MS32 locus in these human irradiated samples may prove extremely fruitful. If consistent data can be obtained, this will provide the first direct measurement of individual induced mutation in vivo. Such mutation events could be recovered and analysed in detail by MVR-PCR. Additionally in the laboratory the induction of somatic length change mutations in cell line DNA by various established DNA mutagens (such as EMS) is being investigated using the SP-PCR technique (C. May, personal communication).

New models for minisatellite mutation. Although analysis of MS32 in vivo has revealed the influence of cis-acting factor in minisatellite mutation, this was only possible due to the serendipitous discovery of the OIC variant in the 5' MS32 flanking DNA and its association with non-mutating alleles. Further analysis of the regions of the MS32 flanking DNA involved in this process would be greatly enhanced by the development of an animal model of human MS32 minisatellite instability. To this end the development of "mark II" MS32 transgenic mice is underway. The DNA construct used to produce these animals will be designed to incorporate more of the MS32 5' and 3' flanking DNA. Currently the recovery of additional cosmid clones containing MS32 has extended the amount of cloned locus flanking DNA to 17.5kb of 5' flanking sequence and 40kb of 3' (A. Jeffreys, personal communication). By using recently developed PCR techniques for long range PCR amplification of DNA fragments up to 35 kb in length (Bar-nes 1994), much larger constructs for transgenesis can be produced.

Although mutation analysis of the preliminary transgenic loci, described in this thesis, was disappointing in that transgene mutation processes did not reflect those seen at the endogenous locus, this work has shown that transminisatellitic animals can be produced. The microinjection method of introducing genetic material into cells
has been successfully demonstrated as a way of producing minisatellite transgenes, and would be appropriate for producing the "mark II" transgenes. However, experience dictates that rather than concentrating on the structural analyses of both multi-copy and single-copy insertions, the most appropriate course for the analysis of any "mark II" transgene loci created, would be to initially identify only mutating single-copy loci for further analysis. Although multi-copy transgenes have been demonstrated in this thesis to have many interesting mutational properties, they are unlikely to ever result in a good model of human minisatellite mutation. Single-copy loci could be identified by conventional Southern blot analysis to exclude multi-copy insertions and SP-PCR analysis would enable assays for evidence of mutations. If the SP-PCR experiments were performed using flanking primers in the original transgene construct, this would provide another level of selection for single-copy transgenes that had maintained the basic 5' and 3' flanking sequences around the repeat unit array. More detailed structural analysis of appropriate transgenes could then be performed. This emphasis of analysis, would allow the screening of many transgene integrations to identify the most suitable insertions for further study.

If a transgenic model of MS32 mutation could be successfully produced, then systematic mutagenesis of the flanking DNA could be performed to enable identification of flanking DNA regions important in mutation. The methods could also be extended to create transgenic models of mutation at other minisatellites where the serendipitous discovery of the OIC variant may not be repeated. Accurate models of human minisatellite behaviour in mice would also enable the quantitative analysis of the effects of mutagenic chemicals and radiation upon minisatellite mutation rate in the germline and soma.

One of the important questions about minisatellite mutation is whether mutation is a largely meiotic or mitotic process. A fully mutating model of MS32 mutation in the mouse would enable the exciting possibility of SP-PCR analysis of mouse sperm fractions. Spermatogonial cells go through a mitotic division to differentiate into primary spermatocytes (Langman 1981). After a round of replication, these primary spermatocytes undergo a first meiotic division to give secondary spermatocytes and a second meiotic division to give spermatids which mature into sperm. If the meiotic mutation process is important in MS32 mutation as predicted (Jeffreys et al., 1994) then meiotic mutation events at an MS32 transgene locus would not be observed in spermatogonia and primary spermatocyte cells, which have not undergone a meiotic division, but ought to be represented in secondary spermatocyte and spermatid DNA.

There would also be a great interest in the animals transgenic for different types of tandemly repeated sequence. Mice transgenic for some triplet repeat disease loci, for example HD and MD, are being developed (M. MacDonald, personal communication; D. Monckton, personal communication). These may provide models for the assessment of therapies for these diseases. In addition it would be interesting to observe the behaviour of a STR transgene constructs composed differing lengths of a run of C/G rich repeats, followed by a run of A/T rich repeats to ascertain if mutation events occur preferentially in one type of sequence and/or whether the length of the repeat array has an influence on the level of instability as predicted from observations at disease associated triplet repeat loci.

Whilst these transgenes have provided a variety of interesting instability phenomena for analysis, a fuller understanding of the mechanisms underlying minisatellite mutation awaits further investigation.
Chapter 10

References


References


References


characterization of a hypervariable region downstream of the human α-globin gene cluster. EMBO J. 5: 1857-1863.


References


References


References


References


References


References