DNA sequence variation within and around human minisatellites

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

by

Darren G. Monckton BSc. (Bath)
Department of Genetics
University of Leicester

October 1992
PREFACE

Although the work contained in this thesis covers the same general area, the work envelopes several distinct aspects within the overall theme. As such I have chosen to weight the bulk of the thesis into the central results sections. The introductory chapter (Chapter 1) is a brief introduction to the human genome, genomic variation, the place of minisatellites within it and current DNA typing systems. The results chapters (3-8) contain more specific introductions to their subject matter, actual results obtained and specific conclusions to be drawn from them. Each results chapter in effect forms a mini thesis, each of which may be read largely as a separate entity, whilst the discussion chapter (Chapter 9), attempts to draw general conclusions and relate the results obtained here with a wider view of minisatellite biology, genomic evolution and individual identification.

The methods used during the course of this work followed almost exclusively standard molecular biological procedures adequately described elsewhere. Thus, the fully referenced, materials and methods chapter (Chapter 2) does not contain an exhaustive list of the exact protocols used. Rather, the sequential use of a variety of previously described techniques to develop an overall new procedure, plus minor modifications of standard techniques and specific methodological details, are described, in context, within the results chapters.
CONTENTS

Preface
Contents
Abstract
Acknowledgements
Publications
Abbreviations

Chapter 1  INTRODUCTION
Genomic variation
Repetitive DNA
Dispersed repeats
Satellite DNA
Telomeric DNA
rDNA
Gene clusters
Micro-, mini- and midi-satellites
Minisatellites
Dinucleotide and other simple sequence repeats
Coding minisatellites
Triplet repeats and human disease
Midisatellites
Minisatellite internal variation
Minisatellite mutation processes
DNA typing systems and individual identification
Multilocus DNA fingerprints
SLP analysis
PCR amplifiable tandem repeats
HLA-DQα
MVR-PCR
This work

Chapter 2  MATERIALS AND METHODS
Materials
Chemicals, enzymes and other molecular biology reagents
Oligonucleotides
Human DNAs
Methods
DNA preparation
General methods for DNA handling
Enzymatic manipulations
Preparative gel electrophoresis
General PCR
DNA hybridization
DNA sequencing
Computing

Chapter 3  MINISATELLITE 'ISOALLELE' DISCRIMINATION IN PSEUDOHOMOZYGOTES BY SINGLE MOLECULE PCR AND VARIANT REPEAT MAPPING
Summary
Introduction
MS32
Internal mapping at MS32
Homozygote excess at minisatellite loci
This work
Results
Discussion
### Chapter 4  SINGLE STRANDED MOBILITY POLYMORPHISMS OF MINISATELLITE ISOALLELES

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Methods for detecting single base substitutions</td>
<td>1</td>
</tr>
<tr>
<td>Minisatellite isoallelism and internal variation</td>
<td>2</td>
</tr>
<tr>
<td>This work</td>
<td>2</td>
</tr>
<tr>
<td>Results</td>
<td>3</td>
</tr>
<tr>
<td>Single stranded mobility polymorphisms of PCR amplified MS32 alleles</td>
<td>3</td>
</tr>
<tr>
<td>SSMP of MS32 alleles on Southern blot analysis of total genomic DNA</td>
<td>3</td>
</tr>
<tr>
<td>PCR-SSMP at MS205 - Segregation analysis</td>
<td>4</td>
</tr>
<tr>
<td>SSMP analysis for MS1</td>
<td>5</td>
</tr>
<tr>
<td>SSMP analysis of unique sequence DNA</td>
<td>5</td>
</tr>
<tr>
<td>Discussion</td>
<td>6</td>
</tr>
</tbody>
</table>

### Chapter 5  BASE SUBSTITUTIONAL POLYMORPHISMS IN THE FLANKING DNA OF MS32 AND THEIR APPLICATIONS IN MVR-PCR

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>MVR-PCR at MS32</td>
<td>1</td>
</tr>
<tr>
<td>'Null' repeats during MVR-PCR</td>
<td>2</td>
</tr>
<tr>
<td>Allelic variation at MS32</td>
<td>2</td>
</tr>
<tr>
<td>Allelic analysis by MVR-PCR at MS32</td>
<td>3</td>
</tr>
<tr>
<td>This work</td>
<td>3</td>
</tr>
<tr>
<td>Results</td>
<td>3</td>
</tr>
<tr>
<td>Identification of three common polymorphisms in the flanking DNA</td>
<td>3</td>
</tr>
<tr>
<td>Primate sequence comparisons</td>
<td>4</td>
</tr>
<tr>
<td>Assays for the polymorphisms and heterozygosity analysis</td>
<td>4</td>
</tr>
<tr>
<td>Knockout MVR-PCR</td>
<td>5</td>
</tr>
<tr>
<td>Haplotype analysis of flanking DNA polymorphisms</td>
<td>5</td>
</tr>
<tr>
<td>Applications to mixed DNA samples</td>
<td>6</td>
</tr>
<tr>
<td>Discussion</td>
<td>7</td>
</tr>
</tbody>
</table>

### Chapter 6  ALLELIC STRUCTURE AND DIVERSITY AT MS32

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Allelic variation at MS32</td>
<td>1</td>
</tr>
<tr>
<td>MVR-PCR at MS32</td>
<td>1</td>
</tr>
<tr>
<td>This work</td>
<td>2</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>2</td>
</tr>
<tr>
<td>Pedigree analysis</td>
<td>2</td>
</tr>
<tr>
<td>Unrelated individuals</td>
<td>2</td>
</tr>
<tr>
<td>Allelic diversity</td>
<td>2</td>
</tr>
<tr>
<td>Diploid code diversity</td>
<td>3</td>
</tr>
<tr>
<td>MS32 MVR-PCR heterozygosity</td>
<td>3</td>
</tr>
<tr>
<td>Allele alignments- structural analysis</td>
<td>4</td>
</tr>
<tr>
<td>flanking haplotypes</td>
<td>5</td>
</tr>
<tr>
<td>population analysis</td>
<td>5</td>
</tr>
<tr>
<td>Conclusions</td>
<td>6</td>
</tr>
</tbody>
</table>
Hypervariable minisatellites have found increasing use in a wide range of applications including linkage analysis, relationship testing and forensic medicine. Conventional length analysis via restriction digestion, agarose gel electrophoresis and Southern blot hybridization produces at best, less than 100 resolvable alleles and involves inherently error prone length estimations. Forensic analyses based on such a system have recently been heavily criticised and the population assumptions used to calculate match frequencies questioned. However, allele length is not the only criterion by which minisatellite variability may be assessed. All well characterised hypervariable human minisatellites also show variation in the dispersion patterns of minisatellite variant repeats. At the most well characterised human minisatellite MS32, MVR variation has been assayed by both restriction analyses and a more direct MVR-PCR approach. Both procedures have shown that allele length analysis seriously underestimates the true variability at minisatellite loci. Many individuals incorrectly typed as homozygous by length analysis have been demonstrated to be heterozygous by internal MVR mapping. MVR allelic analysis has been used to directly show the existence of many hundreds of different alleles, each one absolutely defined, from which we can infer the existence of many thousands of alleles. With a theoretical capacity to define many millions of alleles MVR analysis adds a new dimension to the study of variation in human DNA. The high observed allelic variability is directly reflected in the very high individual specificity obtained when MVR-PCR is applied to total genomic DNA to obtain a diploid code of the two superimposed alleles. MVR diploid code analysis produces a highly portable digital output which provides unambiguous match criteria. Diploid code analysis may be used to obtain highly discriminatory information from mixed DNA samples and, via the use of allele specific flanking primers in knockout MVR-PCR, from admixtures far lower than currently approachable. MVR analysis of allelic variation has revealed the existence of a terminal mutation hotspot which has been confirmed in studies of de novo mutation events in pedigrees. This work has revealed the complex nature of minisatellite mutation, with, for the first time, strong evidence for a role for unequal interallelic exchange and preliminary evidence for a size increase bias.
ACKNOWLEDGEMENTS

Without doubt my greatest thanks go to Alec for his excellent supervision throughout the course of this project and his unwavering attention during the writing of this thesis. I consider it a great honour to have worked in his laboratory, especially at a time when such significant progress has been made and to have worked with the outstanding array of scientists collected in G19. I should like to thank everyone with whom I have worked in G19 both past and present, including Andy, Annette, David, Esther, Ian, John, Keiji, Mark, Maureen, Max, Moira, Nicola, Rita, Yuri and especially Ila, whose efficient practice has ensured the continued smooth operation of the lab. For lively discussions in the lab I should especially mention John, Nicola, Andy and Mark, whilst for extracurricular activities, including barbecues, parties, trips, slammer sessions and general drunkeness I should like to thank Annette, David, Carl, Cathryn, Ian (G19), Ian (tall), Karl, Katrina, Keiji, Lee, Lesley, Louise, Mark (again), Max, Mike, Nikki (the stick), Richard (big) and Sian. Annette deserves one final mention as a cooperative 'bay-mate' and expert (if a little slow) proof reader.

Many thanks must also be extended to the whole Genetics department which has been friendly and run both smoothly and efficiently throughout, in particular to Terry and Joan for orders and the like, to Sarah for liberal use of her printer and even to Margaret whose direct approach is required to run any large organization.

Finally I must extend great thanks to my family who have supported and encouraged me throughout my student days, although alas the time has now come to get a proper job, well maybe.
Some of this work has been published:


ABBREVIATIONS

AR       Androgen Receptor gene
CEPH     Centre d'Etude du Polymorphisme Humain
CFTR     Cystic Fibrosis Transmembrane Regulator
DGGE     Denaturing Gradient Gel Electrophoresis
DM       Dystrophia Myotonica, myotonic dystrophy
DSB      Double Strand Break
DSBR     Double Strand Break Repair
LINE     Long INterspersed repetitive Element
LTR      Long Terminal Repeat
MER      MEedium Reiteration sequence element
MT-PK    MyoTonin Protein Kinase
MVR      Minisatellite Variant Repeat
MVR-PCR  MVR mapping by PCR
NTS      Non Transcribed Spacer region
PCR      Polymerase Chain Reaction
PFGE     Pulsed Field Gel Electrophoresis
RFLP     Restriction Fragment Length Polymorphism
RTVL     ReTroViral-Like element
SBMA     Spinal and Bulbar Muscular Atrophy
SCE      Sister Chromatid Exchange
SINE     Short INterspersed repetitive Element
SLP      Single Locus Probe
SMD      Single Molecule Dilution
SSA      Single Stranded Annealing
SSCP     Single Stranded Conformational Polymorphism
SSMP     Single Stranded Mobility Polymorphism
STR      Simple Tandem Repeat
THE      Transposon-like Human Element
VNTR     Variable Number of Tandem Repeat
YAC      Yeast Artificial Chromosome
Chapter 1

INTRODUCTION

The concerted effort to map and sequence the twenty three chromosomes and three thousand million base pairs comprising the haploid human genome has progressed considerably since the inception of the 'human genome project' some ten years ago (Gilbert, 1982) and its large scale progression over recent years. World-wide collaboration has lead to significant progress in both genetic and physical mapping, with yeast artificial chromosome (YAC) and cosmid contigs now available for large portions of the genome. The move to large scale genomic sequencing awaits completion of the physical map and will be aided by further advances in sequencing technology. However, the mass sequencing of random cDNA clones is being pursued by some groups as a direct approach to obtaining coding gene sequences (Adams et al., 1992). Already though, problems associated with intellectual ownership and the patenting of anonymous DNA sequences have arisen (Anderson, 1991).

Even once the colossal task of actually sequencing the genome has been completed, in an anticipated twenty years time, the task of interpreting the abundance of new information will have barely begun. The eventual human genome sequence will be a composite of sequence obtained from many laboratories around the world and hence, a patchwork derived from many libraries and many individuals. Undoubtedly this primary sequence will in itself prove invaluable as a basic tool for genome analysis. However, in order to fully understand structure and function in the human genome we must also know the extent and limits of DNA sequence variation between individuals. Phenotypic variation is based largely on DNA sequence variation and with few exceptions the basis of phenotypic variation in humans is poorly understood. The most well characterised phenotypic consequences of sequence variation are either those relating to inherited genetic disease, or those with direct medical consequences such as blood antigen grouping. As alluded to, sequence variation plays a primary role in the basis of inherited genetic disease, as well as neoplasia and possibly the aging process. Obviously a thorough understanding of the genome, the underlying variation within it and de novo mutation will enable us to develop new strategies for determining the molecular basis for such genetically controlled disease processes. Once the genetic basis of disease pathogenesis is known, more directed approaches to biochemical analysis and effective therapy can be initiated. Individual genetic variation also forms the basis of the growing range of DNA typing systems of increasing application in the field of forensic biology and human population analysis. Thus a thorough understanding of genomic variation has applications in these areas as well.

Variation is ultimately based on mutation and accompanied by haplotypic mixing, as achieved through homologous recombination and the sexual process. The processes of mutation and the control of homologous recombination in man are poorly understood. A detailed characterisation of the diversity present in current human populations and the analysis of de novo mutation events in individuals will allow us to understand more fully such mutational processes, including the role and effects of recombination.

Genomic variation

Genomic diversity covers a wide spectrum of possible variation, in terms of size, frequency, phenotypic effect and molecular basis. Polymorphisms as small as single base changes are common in the genome (Jeffreys, 1979; Nickerson et al., 1992) and, although the majority are found in non coding DNA and are hence
presumably selectively neutral, single base substitutions can be the basis for genetic disease. One of the first human genetic diseases to be characterised at the molecular level, sickle cell anaemia, is the result of a single A to T transversion. Microdeletions and microadditions are less frequent, but can also be the basis of genetic disease, for example around 70% of the defective cystic fibrosis transmembrane regulator (CFTR) alleles producing cystic fibrosis contain a 3bp deletion (Δ508, Kerem et al., 1989). In fact, the CFTR Δ508 allele is so common in Caucasians that it qualifies as a polymorphism in its own right. Medium to large deletions and duplications can be found and are often associated with disease phenotypes, for example deletions and duplications within the globin genes are a common cause of thalassaemia (reviewed by Weatherall and Clegg, 1982) and similar such rearrangements within the large dystrophin gene are a common factor in many cases of Duchenne and Becker muscular dystrophy (for example see England et al., 1990), whilst more recently duplication of a large genomic fragment, incorporating the peripheral myelin protein, has been found to be associated with Charcot-Marie-Tooth disease (reviewed by Anon, 1992). A polymorphic medium sized DNA translocation (>15kb) with no known disease phenotype (Wong et al., 1991) and a polymorphism involving large variations (>180kb) in the length of the subtelomeric region of the short arm of chromosome 16 (Wilkie et al., 1991) have also been identified. Large-scale chromosomal changes including gross chromosomal deletions, duplications and aneuploidy do occur, but are almost exclusively associated with disease phenotypes, as well as being a common feature of tumour progression (Holliday, 1989).

Variation in the location of transposable elements, such as Alu and L1 elements, is rare, although some polymorphic loci do exist and some insertions associated with genetic disease have been observed (see Chapter 8).

The most variable loci in the human genome are the so called variable number tandem repeats (VNTRs), which as their name suggests show variation in the copy number of a tandemly repeated sequence. The term VNTR can be used to include a range of variable loci including mononucleotide repeats, dinucleotide repeats, or microsatellites, up through minisatellites, midisatellites, telomeric, rDNA and satellite repeats. Such loci often show extreme levels of variation in terms of repeat unit copy number with resulting high heterozygosities. Such high allele length variability is based directly upon rapid mutation to new length alleles by poorly understood insertion/deletion mechanisms.

Repetitive DNA

The presence of repetitive DNA is a ubiquitous feature of higher eukaryotic genomes, comprising an estimated 20-30%, or 600-900Mb, of the human genome (see Schmidd and Jelinek, 1982). This mass of DNA consists of a number of distinct DNA species, differing in the nature of their repetition, origin, structure and function. Around 10% is in the form of dispersed repeats, whilst the remainder is tandemly repetitive, although much of this may be in the form of dispersed tandem arrays.

Dispersed repeats. The major classes of dispersed repeats in the human genome are the Alu and L1 elements, each one dispersed throughout and comprising around 5% of the genome (Schmidd and Jelinek, 1982; Singer, 1982). Both Alu and L1 elements show features characteristic of transposable elements and are assumed to have have arisen from multiple RNA mediated transposition events from a limited number of source genes (Britten et al., 1988; Paulo di Nocera and Sakaki, 1990). Neither have any known function, although both are sometimes expressed, either as part of larger transcripts or more rarely as independent transcripts in their own right. Few insertional polymorphisms have been identified and only a few examples of de novo insertions causing genetic defects have been reported (see Chapter 8). In addition, internal sequence variation and/or
variation in the length of the Alu element polyA tract has been observed for the majority of Alu repeat loci analysed (Economou et al., 1990; Orita et al., 1990; Epstein et al., 1990). Dispersed repeats are also known to act as promoters of non-homologous recombination (between close, but non-homologous copies of the same dispersed repeat), often leading to deleterious deletions (Ariga et al., 1990; Berkvens et al., 1990) and duplications (Lehrman et al., 1987). Other, less abundant, dispersed repeats include Transposon-like Human Elements (THE, Paulson et al., 1985) and MEedium Reiteration sequences (MER, Jurka, 1990) and will be discussed in a more detailed appraisal of human dispersed repeats in Chapter 8.

Satellite DNA. Large arrays, many hundreds of kilobases in length, composed of tandem repeats of a complex sequence, form a sizeable proportion of the higher eukaryotic genome. By virtue of its often atypical G/C content, this DNA can frequently be separated from the bulk of the genome by density gradient centrifugation, giving rise to so called 'satellite bands' and hence the name satellite DNA. The most prominent form in primates is α-satellite, which is present at the centromere of every human chromosome and forms around 5% of the genome (reviewed by Willard, 1990). α-satellite is assumed to have a functional role in centromeric processes, such as homologue recognition, pairing and segregation during meiosis (Willard, 1990). The α-satellite blocks are composed of tandem arrays of a basic 171bp monomer repeated many thousands of times and spanning approximately 1-3Mb across each centromere (eg Ge et al., 1992; Haaf and Willard, 1992). Internal sequence variation between monomers is frequent and can be used to distinguish α-satellite subfamilies, which may be common to a group of chromosomes (eg Choo et al., 1988), or more often chromosome specific (eg Haaf and Willard, 1992). Variant monomers with alternative restriction enzyme sites have been used to identify higher order repeats within satellite blocks (reviewed by Willard and Wayne, 1987). Such higher order repeats are a common feature of satellite DNA and have been used in attempts to infer the processes operating to maintain homogeneity within satellite DNA (Willard and Wayne, 1987). Clustering of variants and high order repetition appear to suggest that the most frequent processes are intrachromosomal and largely short range (eg Willard and Wayne, 1987; Ge et al., 1992) and probably a result of unequal exchange (Smith 1976). However, the existence of α-satellite on all the human chromosomes and shared subfamilies demonstrates that exchange between non-homologous chromosomes has also played an evolutionary role.

The size and type of α-satellite blocks at homologous loci can often show extreme levels of individual variation (eg Ge et al., 1992; Haaf and Willard, 1992). However, conventional α-satellite assays require the use of high molecular weight DNA and pulsed field gel electrophoresis (PFGE); furthermore, the resulting banding patterns are complex and often require interpretation of intensity differences (Ge et al., 1992; Haaf and Willard, 1992). A few, more easily reproducible, but still relatively complex, PCR based assays for haplotyping variant repeats within centromeric α-satellite have been developed (Warburton et al., 1991), but still the application of α-satellite polymorphisms are largely restricted to linkage mapping and satellite evolution.

β-satellite has been less well studied, but makes up a considerable proportion of the short arm of the acrocentric chromosomes (Greig and Willard, 1992). The basic repeat unit is 68bp long and like other satellite sequences, forms high order tandem repeat arrays which can extend for up to hundreds of kilobases (Greig and Willard, 1992).

Both α-satellite (Choo et al., 1988) and β-satellite (Greig and Willard, 1992) have been mapped to and implicated, as possible breakpoints for the very common Robertsonian translocations which often involve the acrocentric chromosomes.
Telomeric DNA. Telomeres are the highly conserved functional ends of linear eukaryotic chromosomes (reviewed by Blackburn, 1991). All telomeres are comprised of a tandem array of a simple G-rich sequence, with the G-rich strand orientated 5' to 3' towards the terminus and protruding 12-16 bases beyond the complementary C-rich strand. The telomeric DNA is synthesized by a telomerase enzyme which acts as a ribonucleoprotein reverse transcriptase whose RNA template is an intrinsic part of the enzyme. The telomerase action allows complete synthesis of the linear chromosomal end which would otherwise be shortened by progressive rounds of replication. In humans the consensus repeat sequence is AGGGTT arranged in tandem arrays averaging 10kb in length. The exact length of the telomeric array is extremely heterogeneous within an individual, such that a Southern blot analysis using a telomeric probe produces a smear rather than a discrete band. Some evidence suggests that in humans at least, normal somatic cells lack telomerase activity with a resulting gradual decrease in telomere length through successive cell generations; the relevance of these observations to chromosomal instability, aging and senescence is presently unknown (Blackburn, 1991).

rDNA. The genes for the 18S and 28S rRNA in man are arranged as tandem arrays in the so called nucleolar organizer regions on the short arms of the five acrocentric chromosomes. Each cluster is comprised of around 40 repeats of a 44kb monomer made up of a 13kb transcribed portion and a larger nontranscribed spacer (NTS), which in itself contains a number of tandem repeats (see Arnheim et al., 1980). The transcribed region shows very high interspecies conservation, whilst the NTS is known to show extreme interspecies variation in terms of both sequence and length, even amongst closely related species such as the great apes (Arnheim et al., 1980). However, species specific variants have become fixed at dispersed loci within a genome, suggesting a relatively high rate of exchange between nonhomologous chromosomes operating over a relatively short evolutionary period (Arnheim et al., 1980). Individual variation in both the length of the NTS region (Arnheim et al., 1980) and in the proportions of specific base substitutions (Qu et al., 1991) have been observed, but the multicopy nature of the array makes Southern blot analysis usually dependent on intensity difference interpretations. Due to the large size of a single repeat (44kb) unit, higher order repeat analysis has so far eluded researchers. Even attempts to clone large rDNA arrays in YACs, for reasons as yet unclear, have proven largely unsuccessful (Labela and Schlessinger, 1989). rDNA evolution is assumed to involve unequal exchange between repeats on sister chromatids, homologous and non-homologous chromosomes, giving rise to the spread and fixation of variant repeats and variation in the length of the NTS (see Arnheim et al., 1980; Dover, 1982).

Gene clusters. Many genes are arranged in linear clusters, often containing not only identical but also related genes. These clusters are assumed to have arisen from a single source gene, via duplication and rounds of unequal exchange to produce arrays of genes which are able to diverge and form the basis of multigene families (for review see Maeda and Smithies, 1986). These clusters differ from the tandemly arrayed rDNA genes in that the intergenic sequences are usually greatly diverged and of varying lengths. Examples of such gene clusters in man include the histone genes (Heintz et al., 1981), the immunoglobulin genes (see Baltimore, 1981), the haptoglobin genes (Maeda and Smithies, 1986) and most notably the globin genes. The human globin genes are arranged in two clusters, the α-cluster of four genes and three pseudogenes, covering over 28kb on chromosome 16 and the 50kb β-cluster containing five genes and a pseudogene on chromosome 11 (see Weatherall and Clegg, 1982; Orkin and Kazazian, 1984). Variation due to deletions and duplications in these regions are frequent and many form the basis of common thalassaemias (for reviews see Weatherall and Clegg, 1982; Orkin and Kazazian, 1984). Many of these rearrangements are assumed to arise through unequal interchromosomal exchange, for example between the two α-globin genes to give either the deletion, α3.7, or
duplication, $\alpha\alpha$anti-$3.7$ and between the $\gamma$- and $\beta$-globin genes to give the Hb Lepore gene (Weatherall and Clegg, 1982; Orkin and Kazazian, 1984).

Micro-, mini- and midi-satellites

Apart from the large blocks of satellite DNA, higher eukaryotic genomes are also blessed with the occurrence of other smaller dispersed blocks of simpler sequence repeats. The categorisation of these repeat blocks into micro-, mini- or midi-satellite, or indeed some other categorisation such as simple tandem repeat (STR), is somewhat arbitrary, with no clear boundaries between them. In fact it is not always clear to which particular aspect of repeat biology, ie monomer repeat length or total repeat array length, the qualifier applies and in reality attempts to define precise classifications based on our current understandings are probably meaningless. Such sequences are probably best viewed as comprising a continuum of structures from short mononucleotide repeats through to some very large arrays such as the midisatellites, merging into the extensive satellite arrays. The molecular processes operating on such sequences are likely to be shared by various members of each group throughout the continuum, with varying relativities and dependence on factors other than purely repeat sequence and/or monomer/array length. Nonetheless, for reasons of simplicity some divisions in the discussion of these sequences will be made, although as discussed, these may not reflect true biological distinctions.

Minisatellites. The first observation of a highly polymorphic RFLP having more than the usual two allelic states associated with a standard restriction site dimorphism was made by Wyman and White in 1980. Over the next few years a series of such loci were fortuitously isolated from essentially random clones of genomic regions, usually containing a gene of interest (eg Higgs et al., 1981; Bell et al., 1982; Capon et al., 1983) and in some cases the basis of variability shown to be due to variation in copy number of tandem repeat arrays (Bell et al., 1982; Capon et al., 1983; Wyman et al., 1986). The major breakthrough in minisatellite analysis came with the demonstration by Jeffreys et al., (1985a) that multiple polymorphic tandem repeat loci (for the first time termed 'minisatellites'), of similar sequence, could be detected on genomic Southern blot analysis using a tandem repeat probe hybridized at low stringency. The 'patterns' revealed by this type of analysis were shown to be highly individual specific and soon became known as 'DNA fingerprints' (Jeffreys et al., 1985b). The potential applications of such multilocus DNA fingerprints were obvious and quickly realised; for example, individual identification (Jeffreys et al., 1985b), kinship testing (Jeffreys et al., 1985c), forensic medicine (Gill et al., 1985), monitoring bone marrow transplants (Thein et al., 1986), following tumour progression (Thein et al., 1987) and for identifying cell lines (Thacker et al., 1988).

The cross hybridizing properties of minisatellite probes also allowed the directed cloning of more minisatellite loci (Jeffreys et al., 1985a; Wong et al., 1986) and over the last few years many such hypervariable human minisatellite loci have been cloned (Wong et al., 1987; Nakamura et al., 1987a, 1988; Armour et al., 1990). During high stringency Southern blot hybridization these hypervariable minisatellite clones detect only their cognate single locus (single locus probes, SLP) and have proven to be very informative genetic markers (Nakamura et al., 1987a), with sometimes very high allele length heterozygosities which can exceed 90% (Wong et al., 1987). The application of minisatellite loci to linkage mapping is though limited by their non-random dispersion through the genome; in situ hybridization (Royle et al., 1987) and linkage mapping (Nakamura et al., 1988; Armour et al., 1990) place the majority of these loci in subtelomeric regions and/or at the distal end of linkage maps. These markers can also be used for individual identification and forensic analysis (Wong et al., 1987; see below), detecting allele loss in tumours (Vogelstein et al., 1989) and the detection of chromosomal abnormalities such as uniparental disomy (Malcolm et al., 1991).
Standard single locus minisatellite analysis is performed using genomic digests and Southern blot hybridization. PCR analysis can be successfully applied to minisatellite loci, although cycling conditions require careful control to prevent artifactual 'collapse' of products (Jeffreys et al., 1988a). (PCR collapse is defined as the production of aberrant length products due to the misaligned priming action of incomplete single stranded products (Jeffreys et al., 1988a)). PCR amplification of tandem repeat loci is also limited by the size of alleles that can be efficiently amplified, with a maximal amplifiable size of about 10kb (Jeffreys et al., 1988a), although many of the most variable loci frequently have alleles much larger than this (10-30kb, Wong et al., 1987). A few loci of limited allele size range have been identified and provide systems for rapid PCR analysis (Boerwinkle et al., 1989; Budowle et al., 1991).

**Dinucleotide and other simple sequence repeats.** The existence of large numbers of short dinucleotide repeats dispersed throughout the genomes of man (estimated at 50,000-100,000 copies of CA repeats) and other higher eukaryotes has been known for some years (Miesfeld et al., 1981; Hamada and Kakunaga, 1982; Sun et al., 1984; Tautz and Renz, 1984). Whilst repeat length variation, as determined by DNA sequencing, between different clones had been observed (Slighorn et al., 1980; Shen and Rutter, 1984), no rapid method for assessing individual variation at these loci was available. The development of PCR (Saiki et al., 1988) though, has enabled allelic status at these loci to be determined simply and rapidly and has been used to show that many such loci are indeed polymorphic, with heterozygosities up to 90% (Weber and May, 1989; Litt and Luty, 1989). Such potential marker loci are widespread in the genome, with no aberrant clustering obviously apparent and are relatively easy to clone, sequence and develop into useful marker systems (eg Hudson et al., 1992). Their adoption by the linkage mapping community has been rapid and widespread (eg Matsutani et al., 1992; Small et al., 1992), with microsatellite based maps constructed or under construction for at least 12 human chromosomes (see Wang and Weber, 1992).

Other short simple sequence repeats (trimeric and tetrameric) have also been developed as marker systems and these too, appear relatively informative, common and well dispersed throughout the genome (Edwards et al., 1991, 1992).

Reports of large arrays of simple sequence motifs seem surprisingly absent from the human genetic literature, although recent data suggest that this apparent deficiency may not reflect the true biological situation. A previous search of the Genbank sequence database revealed over 100 human CA repeats of at least 6 or more repeats, although none of the sequences identified contained more than 30 repeats (Weber, 1990). Since this search, a large imperfect CA repeat has been found in a subtelomeric region of chromosome 16p, with alleles ranging from 30 to ~450 repeats (Wilkie and Higgs, 1992). This array is however surprisingly uninformative (Wilkie and Higgs, 1992), especially in light of the observation that informativeness generally increases with allele length for previously identified CA repeats (Weber, 1990). Even more recently the identification of large unstable triplet repeats associated with human heritable disease has been reported (see below). Armour et al., (personal communication) have isolated two very large imperfect simple sequence repeat arrays. One of these has been partially characterised and appears to consist largely of blocks of CCW repeats and other triplet repeats (Armour et al., personal communication). Alleles lie in the range from 2 to at least 7kb, with an estimated heterozygosity in excess of 70%. This locus, along with the 16p locus (Wilkie and Higgs, 1992) and the large triplet repeats (Fu et al., 1991; Brook et al., 1992), is however extremely difficult to propagate in a bacterial host, as well as being somewhat refractory to PCR amplification (Armour et al., personal communication). It now seems probable that large simple sequence repeats may indeed be more common in the genome than
previously realised, with their apparent absence reflecting the difficulties associated with their cloning and analysis.

**Coding minisatellites.** Although the majority of hypervariable minisatellites cloned to date are non-coding (Wong *et al*., 1987; Nakamura *et al*., 1987a, 1988; Armour *et al*., 1990), a few variable coding minisatellites have been identified in humans. For example, the mucin protein (MUC1) locus on chromosome 1 shows extreme length variability (heterozygosity >80%) at the DNA level, which is also reflected in the electrophoretic mobility of the resultant protein product (Swallow *et al*., 1987). Detailed analysis has demonstrated the polymorphism is due to variation in the number of 60bp (20 amino acids) coding repeats, with the smallest allele containing 21 repeats and the largest 125, producing a protein of RMM about 225 kDa (Gendler *et al*., 1990). Other, less variable, coding minisatellites include other mucins (Gum *et al*., 1989), proline rich proteins (Azen *et al*., 1984) and the involucrin gene (Simon *et al*., 1991). The large family of collagen genes (reviewed by Vuorio and Crombrugghe, 1990) also contain extensive regions of tandem repetition, although variation in copy number of tandem repeats has not been reported. The collagen genes are essential structural genes present in all higher eukaryotes, showing extreme levels of conservation, even in terms of gene structure, between distantly related species (Vuorio and Crombrugghe, 1990). These genes appear to be under very high selective pressure, with most variation almost always associated with a severe phenotype (see Vuorio and Crombrugghe, 1990). At least two *de novo* deletions of a single repeat unit have though been identified (Wallis *et al*., 1989; Hawkins *et al*., 1991), both associated with severe dominant phenotypes, suggesting that collagen repeats are prone to similar mutational mechanisms operating at other tandemly repeated loci, but are maintained in a non-polymorphic state by intense selective pressure. The effect of high selective pressure operating at coding minisatellites has not been fully investigated, but can be assumed, in general, to limit population variability.

Recently a new class of variation, associated with tandem repeats in genic DNA and genetic disease, has come to light and is discussed below.

**Triplet repeats and human disease.** Recently a novel basis for inherited genetic disease in man has been identified for three diseases, so far, X-linked spinal and bulbar muscular atrophy (SBMA, sometimes known as Kennedy’s disease), fragile X mental retardation and myotonic dystrophy (Dystrophia Myotonica, DM)(reviewed by Caskey *et al*., 1992). All three have been found to present a phenotype associated with variation in the copy number of a triplet repeat located in the disease gene. In SBMA a polymorphic CAG triplet repeat (polyglutamine) has been found in exon 1 of the androgen receptor gene (AR). Normal individuals have a copy number of around 11 to 31 repeats (La Spada *et al*., 1991; Edwards *et al*., 1991, 1992), whilst affected individuals have approximately double sized arrays containing from 40 to 52 repeats (La Spada *et al*., 1991). The molecular pathology of the AR repeat has not been fully elucidated, but may be related to tissue specific transcriptional regulation via the polyglutamine tract (La Spada *et al*., 1991). Some *de novo* mutation events have been observed at this locus (Caskey *et al*., 1992), but they are small and relatively rare. This data, along with the relatively narrow allele size range for disease association, suggest that a single duplication to the larger disease allele may be ancient (Caskey *et al*., 1992). In fragile X the defect has been localised to an unstable triplet repeat ((CGG)$_n$) in the 5' untranslated region of the FMR-1 gene (see Fu *et al*., 1991). This repeat is again polymorphic in the normal population (6 to 54 repeats), with disease associated alleles showing an increased size range (200+, Fu *et al*., 1991). Progression of the fragile X phenotype appears to be associated with disease specific methylation of the FMR-1 gene and surrounding DNA, and may be the result of shut down of FMR-1 itself, or another closely linked gene (Caskey *et al*., 1992). For DM the defect has also been identified
as expansion of a CTG repeat (the same sequence as in SBMA) located in the 3' untranslated region of the
dystrophia myotonica protein kinase (MT-PK) gene (see Brook et al., 1992). Again, this sequence is polymorphic in the
normal population, 5 to 30 repeats and expanded in affected individuals, greater than 50 and up to 2000 repeats
(Brook et al., 1992). The expanded DM alleles show very strong linkage disequilibrium with flanking markers,
highly suggestive of a single ancient ancestral mutation (Harley et al., 1992; Aslanidis et al., 1992). How
repeat expansion in the 3' untranslated region of the MT-PK gene results in a dominant phenotype is unclear at
present, although no evidence for a change in methylation status has been observed (Caskey et al., 1992). Both
fragile X and DM display the relatively rare phenomenon known as anticipation, whereby disease severity
increases in successive generations. The molecular basis of anticipation has now been identified as an increase in
size of the triplet repeat associated with progressive instability in successive generations (Fu et al., 1991; Brook
et al., 1992). Both have allele designations in ranges of normal, premutation (very mildly or non-affected, but at
increased risk of having affected offspring) and full mutation (affected individuals). Alleles in the premutation
range have an increased rate of mutation, being relatively unstable in the germline. However, once alleles
progress into the full mutation range instability increases dramatically with germline mutation rates of 1 and
extensive somatic mosaicism (Fu et al., 1991; Brook et al., 1992). The molecular basis of triplet repeat
expansion and instability is presently not known, with no evidence for unequal homologous exchange having
been obtained. Interestingly, fragile X expansion is limited to the female germline, whereas in DM expansion
can occur in either germline (Fu et al., 1991; Brook et al., 1992). These findings present an exciting new
challenge to medical molecular genetics, combining tandem repeat biology with a disease phenotype.

Midisatellites. The first 'midisatellite' was identified by Nakamura et al., (1987b) near the
telomere of the short arm of chromosome 1 and consists of a very large tandem array (50-200kb) of 40bp repeats.
A further three midisatellites have since been identified (Page et al., 1987; Gray and Jeffreys, personal
communication) and all have been shown to be highly polymorphic (alleles 10-200 kb) using PFGE, or through
Southern blot analysis of internally cutting restriction enzymes to produce highly complex patterns of variable
cosegregating haplotypes.

Minisatellite internal variation. Every hypervariable human minisatellite characterised in
detail has been found to vary not only in terms of allele length, but also in repeat unit sequence (eg Bell et al.,
1982; Capon et al., 1983; Owerbach and Aagaard, 1984; Wong et al., 1986, 1987). Thus for each locus two or
more minisatellite variant repeat (MVR) types can be defined. These variant repeats have been found to exist not
just as isolated copies, but also as multiple copies dispersed throughout the array (Owerbach and Aagaard, 1984;
Wong et al., 1986, 1987). PCR amplification of the human minisatellite MS32 (D1S8) allowed the first survey
of MVR dispersion patterns (MVR maps) in large numbers of alleles to be made (Jeffreys et al., 1990). For this
locus two MVRs, differing by the presence/absence of a HaeIII restriction site, could be mapped within
amplified alleles by restriction mapping to give a binary code for the internal dispersion of the two MVRs
(Jeffreys et al., 1990; see Chapter 3 for details). This approach revealed extreme levels of variation in MVR
maps, displaying for the first time the true extent of allelic variation at minisatellite loci (Jeffreys et al., 1990).
However, this initial approach was limited in its application, being only of real use in studies of minisatellite
variation, mutation and evolution. The next major breakthrough came with the simplification of the technique
using MVR specific PCR primers (MVR-PCR) to assay the same sites of internal variation at MS32 (Jeffreys
et al., 1991; see Chapter 5 for details). MVR-PCR can be applied not only to separated alleles, but also to total
genomic DNA to reveal a diploid code derived from the superimposition of both single allele codes. MVR-PCR
has potential applications in studies of allelic variation (see Chapter 6), minisatellite mutation (see Chapter 7), population analysis (see Chapter 6), kinship testing and individual identification (see Chapter 5). MVR-PCR analysis is currently being applied to a number of other minisatellites where similar levels of internal variation are being revealed (Armour et al., unpublished data; Neil et al., unpublished data; see Chapter 7).

**Minisatellite mutation processes.** The hypervariability observed at minisatellite loci is a direct consequence of a high germline mutation rate. For some loci this rate is so high that *de novo* germline mutation events can be observed directly in pedigrees (Jeffreys et al., 1988b). Not surprisingly a direct correlation between mutation rate and variability has been found, with mutation rates as high as 5% per gamete (Jeffreys et al., 1988b). The mutational processes operative at minisatellite loci are presently poorly understood, although since the initial demonstration of large numbers of dispersed minisatellites in the genome it has been speculated that they may play a role in homologous recombination (Jeffreys et al., 1985a). The evidence linking minisatellites and recombination is reviewed in detail in Chapter 7, along with discussion of recent evidence derived from internal structural analysis involving recombination and other mechanisms.

**DNA typing systems and individual identification.**

Every individual, with the sole exception of identical twins, has their own unique genetic make-up, which provides in principle the basis for the unique identification of each individual by the use of DNA typing systems. The applications of such systems are obvious in terms of criminal forensics, for matching suspect samples to scene of crime samples and in civil cases for establishing family relationships. A variety of DNA typing systems have been developed over recent years, but all are based on assaying highly polymorphic loci.

**Multilocus DNA fingerprints.** The classic multilocus DNA fingerprint is highly individual specific and under ideal conditions can be used to match samples with near certainty from a single test (Jeffreys et al., 1985b, 1991a). Unfortunately, multilocus fingerprints do have several associated weakness that limit their application in forensic medicine where samples are often small, of poor quality and quite frequently mixed. These limitations include: the need for reasonably large amounts of DNA (at least 500 ng); the occurrence of ‘band shift’ (the aberrant migration of a DNA sample due to contamination with co-purifying impurities, see Thompson and Ford, 1991); and their susceptibility to the overall limitations of agarose gel electrophoresis (ie relatively poor resolving power, gel distortions, requirement of DNA size markers, etc., all resulting in an inability to accurately determine band size). In addition, the patterns obtained with a multilocus approach are very complex, limiting the ability to database profiles and match samples run on different gels, on different days in different laboratories. Furthermore, the identification and interpretation of mixed DNA samples is difficult with such complex patterns.

**SLP analysis.** Although the limitations of sample impurity and agarose gel electrophoresis are increased for SLP analysis compared to multilocus DNA fingerprinting, it does offer some advantages. These include a greater technical ease of production, the ability to type very small amounts of DNA (as little as 10 ng, Wong et al., 1987), an increased ability to interpret mixed DNA samples (Wong et al., 1987) and a greater ease in defining the simpler resultant profile for computer databasing and inter-laboratory comparisons. Although less informative per test, sequential experiments with a series of hypervariable loci can produce a highly informative profile (Wong et al., 1987). For each locus the expected match frequency is calculated from an allele frequency database under the assumption that the population is at Hardy-Weinberg equilibrium. The match frequencies at each loci are then multiplied together, under an assumption of linkage equilibrium, to obtain the overall match probability. The statistical power of the system is primarily limited by the resolving power of agarose gel
electrophoresis. Minisatellite length alleles vary in a quasicontinuous fashion, where the potential difference in allele lengths is below that resolvable by current electrophoretic technology (at best 100 distinguishable alleles, Wong et al., 1987). The result of this inability to precisely define allelic length is the need to "bin" alleles into groups covering a small size range, or to define an allele with qualifying error margins, resulting in a decreased statistical significance of a match (Budowle et al., 1991). In addition, the inability to resolve closely spaced alleles is responsible for at least part of the apparent excess of homozygotes observed at some minisatellite loci (see Chapter 3 for discussion of homozygote excess and isoallelism). Over the last few years the forensic applications of minisatellite analysis have been brought into question, with many of the limitations discussed above being highlighted, as well as examples of bad laboratory practice and over-enthusiastic data interpretation (see Lander, 1989). More recently concerns about the population assumptions, specifically the potential effects of population substructuring, made in calculating expected match frequencies have been expressed (Cohen, 1990; Lewontin and Hartl, 1991). These concerns have evoked a lively discussion in the scientific literature (see for example Chakraborty and Kidd, 1991; Brookfield, 1992) and brought chaos to the court-rooms of the USA (see Lander, 1991). In response, the National Academy of Sciences set up an investigating committee who concluded that DNA typing technology was scientifically sound and that the population substructural arguments could be allayed, in the absence of further investigation, by use of a conservative "ceiling" approach to estimating match significance (DNA Technology in Forensic Science, National Academy Press, Washington, D. C. 1992). This approach uses the maximum allele frequency identified in any subpopulation (termed the ceiling frequency), or 5%, whichever is the greater, as the basis for estimating genotype frequencies and hence match probabilities. The loss in informativeness resulting from this conservative estimate can be recovered by the use of an increased number of tests.

PCR amplifiable tandem repeats. The very short microsatellite dinucleotide repeats are highly amenable to forensic analysis of degraded DNA samples, although limited by there relatively low informativeness, necessitating the use of a large battery of loci, as well as the PCR stuttering effect frequently observed during polyacrylamide electrophoresis. The application of these systems to individual identification from highly degraded DNA samples has been adequately demonstrated in two recent high profile cases (Hagelberg et al., 1991, 1992). The potential of other PCR amplifiable simple sequence repeats and minisatellites for individual identification has been highlighted, but here too some concerns exist over the ability to precisely define allelic lengths, primarily as a result of the existence of variant length and/or sequence repeats producing non-integral alleles and necessitating binning (Boerwinkle et al., 1989; Budowle et al., 1991; Edwards et al., 1992). However, the population genetic concerns raised over standard Southern blot analysis of hypervariable minisatellites apply even more so to the less variable PCR based allele length systems, which have lower mutation rates and are more prone to population bottlenecks and drift effects. Furthermore, the general application of PCR to forensic biology has not been fully tested in the courts with relatively few cases having proceeded that far. Concerns over potential contamination with such a sensitive system are likely to be substantial, with an increased onus on typing laboratories to comply with strict quality control procedures.

HLA-DQα. Although quite common in the genome, base substitutions only usually have two allelic states and a subsequent maximal heterozygosity of 50%. Such low informativeness has meant that base substitutional polymorphisms are not generally of practical use in individual identification. The exception to this rule is the human leukocyte antigen (HLA) system where relatively large numbers of base substitutions exist over a small distance, forming relatively informative haplotypes. The major histocompatibility (MHC), or HLA,
system comprises a set of informative genetic markers that can be assayed at both the protein and DNA level (Saiki et al., 1986). One of the most informative of these loci is the HLA-DQα locus on chromosome 6 comprising at least 10 different alleles which may be assayed by PCR amplification of a short (<250bp) genomic fragment and hybridization with allele specific oligonucleotides (Saiki et al., 1986). This method is fast, gives unambiguous allelic definition and can be applied to very small amounts of DNA and/or degraded DNA. The informativeness though is insufficient for individual identification and the system is mainly used for exclusion (Tamaki et al., 1991). Similar assays are available for other HLA loci, although their tight genetic linkage again limits variability through linkage disequilibrium. Concerns have also been raised about the use of coding DNA typing systems, in terms of the phenotypic information that may potentially be derived from it; HLA haplotypes have been linked with a number of diseases and population specific selective pressures have previously been identified (eg Hill et al., 1991). DNA typing has already been restricted to anonymous non-coding loci in German courts, although paradoxically serological HLA typing is still permissible as evidence.

**MVR-PCR.** MVR-PCR produces an unambiguous digital diploid code from small amounts of DNA which is highly individual specific (Jeffreys et al., 1991; see Chapter 5). The digital codes are simple to read, immune to many of the problems associated with band size measurements on agarose gels and highly amenable to computer analysis and totally portable between laboratories. Match criteria are unambiguous and the use of diploid code phenotypes negates the use of Hardy-Weinberg allele frequency assumptions in calculating match frequencies (Jeffreys et al., 1991). MVR-PCR is applicable to mixed DNA samples (see Chapter 5), degraded DNA and kinship testing, overcoming many of the limitations associated with other DNA typing systems (Jeffreys et al., 1991). MVR-PCR is being commercially developed by Cellmark Diagnostics and is currently being evaluated by over 25 forensic labs world-wide; it is anticipated that within a few years MVR-PCR will become a standard DNA typing tool.

**This work**

The analysis of sequence variation is fundamental to the functional analysis of the human genome, understanding genetic disease processes, genomic evolution and individual identification. Minisatellites form some of the most variable loci in the genome, highly informative both as linkage markers and as individual identifiers. This work concerns the analysis of sequence variation within and flanking a number of human minisatellite loci. The demonstration that internal variation at minisatellites is extreme and highly informative has enabled a new direction in minisatellite analysis. The occurrence of isoallelism (ie alleles of the same length, but of differing internal structures) is investigated by restriction analysis at MS32 in Chapter 3 and by single stranded mobility polymorphism, in Chapter 4, for several minisatellites. Chapter 5 describes the identification and characterisation of a number of base substitutional polymorphisms in the flanking DNA of MS32 and demonstrates their application to allele specific (knockout) MVR-PCR. The processes underlying variation at minisatellite loci are presently poorly understood; the use of internal variation to study allelic structure, Chapter 6 and de novo mutation events, Chapter 7, is described and discussed in respect to such processes, with specific reference to the role of interallelic recombination and its potential mechanistic basis. The analysis of mutation in man has been largely restricted to the analysis of population variation and a limited number of de novo mutation events detected by pedigree analysis. The advent of PCR has heralded a new approach to mutation analysis through the specific recovery of rare mutation events from bulk genomic DNA. The development of such techniques will have tremendous application in the assessment of genetic load in man and the effect of environmental factors.

*Chapter 1 Page 11*
Chapter 2

MATERIALS AND METHODS

The methods used during the course of this work followed almost exclusively standard molecular biological procedures adequately described elsewhere. Thus, the fully referenced, materials and methods outlined below do not contain an exhaustive list of exact protocols. Instead, only brief descriptions are given of the general methods used. Also, the sequential use of a variety of previously described techniques to develop an overall new procedure, plus minor modifications of standard techniques, exact experimental conditions used and specific techniques are described, in context, within the results chapters.

Materials

Chemicals, enzymes and other molecular biology reagents. All reagents used were standard, purchased from established suppliers of molecular biological reagents (Fisons, FMC Bioproducts, Gibco BRL, NEB, Sigma and Pharmacia).

Oligonucleotides. Oligonucleotides for hybridization probes and PCR amplification were synthesized by J. Keyte and D. Langton in the Department of Biochemistry, University of Leicester.

Human DNAs. Lymphoblastoid cell line DNAs from 40 large Caucasian families were supplied by H. Cann and J. Dausset of the Centre d'Etude du Polymorphisme Humain (CEPH), Paris, France. The source of blood DNAs were: unrelated Japanese individuals, donated by Y. Katsumata, Nagoya University, Japan; British mother-father-single child trios, unrelated Pakistani individuals and large Bangladeshi families from P. Debenham, Cellmark Diagnostics, Oxford, UK; several Fanconi's anaemia families from C. Mathew, Guy's Hospital, London, UK; and individuals from Papua New Guinea and Melanesia by K. Kidd, Yale, USA. Departmental blood DNAs were prepared (see below) from venous blood samples taken by J. Armour.

Methods

DNA preparation. Venous blood and sperm DNA was prepared under PCR clean conditions as previously described by Jeffreys et al. (1990).

General Methods for DNA Handling. General methods for handling DNA, gel electrophoresis, Southern blotting etc were performed as adequately described by Sambrook et al. (1989).

Enzymatic manipulations. DNA restriction enzymes and other DNA modifying enzymes were used according to the manufacturers instructions with the supplied buffer systems.

Preparative gel electrophoresis. For the preparation of size fractionated DNA, the DNA was electrophoresed on agarose gels as normal and the required size fraction was excised from the gel under UV illumination by comparison with standard DNA size markers. Dialysis membrane was cut into sheets just wider (~2mm) than the gel fragment and deeper than the depth of the gel and boiled in 10mM Tris-HCl (pH 7.5), 1mM EDTA for 10 minutes. A slot was cut in a fresh gel (the same concentration etc, as the initial gel) and the gel fragment inserted into this slot with the dialysis membrane inserted vertically in front of its leading edge (ie the edge towards which the DNA will migrate under electrophoresis). The DNA was electrophoresed at 10-15 volts/cm until all of the DNA was loaded onto the membrane (approximately 10 minutes, depending on the size fraction required), either as judged by eye using a hand held UV illuminator, or by comparison with the
migration of a bromophenol blue marker dye. With the voltage still applied, the membrane was rapidly removed to a 1.5ml microfuge tube using tweezers. The tube was then spun for 3 minutes at 15,000 rpm with the corner of the membrane trapped in the lid to gravitate the buffer containing the DNA to the bottom of the tube. The DNA thus recovered was purified by standard ethanol precipitation (Sambrook et al., 1989).

**General PCR (Saiki et al., 1988).** Since the polymerase chain reaction will produce a product from very small amounts of template, precautions were taken to ensure that reagents and materials used for PCR were kept free of contaminating DNA. Thus PCR reactions were prepared using PCR dedicated reagents and clean pipettes, with pipette tips and tubes taken directly from the manufacturers packaging. Single molecule dilutions and PCR reactions were prepared in a laminar flow hood to prevent aerial contamination. Glassware and gel equipment for use in single molecule analysis were cleaned by soaking in 1M HCl. All PCR reactions were performed in conjunction with appropriate zero DNA controls, which consistently gave no products.

PCR amplifications were generally performed, unless otherwise stated, in a 7.5μl reaction mix comprising 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, 1μM of each primer and 0.375 units of *Taq* polymerase (Amersham or Cetus) using 100ng of input total genomic DNA. Reaction mixtures were overlaid with a drop of paraffin oil in a 0.5ml microfuge tube and amplifications carried out in a Perkin-Elmer Cetus DNA thermal cycler (Perkin-Elmer Cetus, Connecticut, USA). Cycling conditions were denaturation for 1 minute at 96°C, 1 minute annealing at A°C and E minutes extension at 70°C for X cycles. Annealing temperatures, extension times, cycle numbers and primer sequences are given in the relevant text.

**DNA hybridization.** For filter hybridizations using double stranded template, 10ng probe were labelled by the random hexamer priming method (Feinberg and Vogelstein, 1984) incorporating α-³²P-dCTP. Tandem repeat probes were hybridized in a modified (Wong et al., 1987) phosphate/SDS solution (Church and Gilbert, 1984) and unique sequence probes were hybridized in a nonfat dried milk solution (Johnson et al., 1984) supplemented with 6% polyethylene glycol. Hybridizations were carried out at 65°C overnight in a shaking water bath or a Hybaid rotating bottle hybridization oven. Filters were washed at high stringency (0.1x SSC, 0.1% SDS) for 1 hour at 65°C. Oligonucleotide probes (10pmoles) were end labelled with γ-³²P-ATP and hybridized overnight at 50°C in a Denhardt's/SSPE solution as described by Sambrook et al. (1989). Filters were washed for 5 minutes in 5x SSC at 50°C. Autoradiography was performed as described (Sambrook et al., 1989) with exposure times of from 1 hour to 1 week, depending on signal strength and band intensity required.

**DNA sequencing.** Single stranded template DNA was generated by asymmetric PCR (Gyllensten and Erlich, 1988). Unique sequence DNA was sequenced in the presence of the detergent NP-40 by the di-deoxy chain termination method as previously described (Bachman et al., 1990) using T7 polymerase (Pharmacia). Minisatellite repeat DNA was cycle sequenced (Innis et al., 1989) for 10 cycles with a 67°C annealing temperature and a 2 minute extension using *Taq* polymerase (Amersham) and endlabelled oligonucleotide primers. Polyacrylamide gel electrophoresis was performed as described by Sambrook et al. (1989).

**Computing.** DNA sequences were analysed using a VAX 8650 Mainframe computer operating on VMS 5.4-2, using the Genetics Computer Group Sequence Analysis Software Package version 6.2 programs developed at the University of Wisconsin (Devereux et al., 1984). Digital MVR data was analysed with software written by A.J. Jeffreys (Jeffreys et al., 1991, and unpublished) in VAX BASIC V3.4 and MicroSoft QuickBasic, running on a VAX 8650 Mainframe and an Apple Macintosh personal computer, respectively.
Chapter 3

MINISATELLITE 'ISOALLELE' DISCRIMINATION IN PSEUDOHOMOZYGOTES BY SINGLE MOLECULE PCR AND VARIANT REPEAT MAPPING

Summary
The hypervariable human minisatellite MS32 has a heterozygosity of 97.5% based on detectable differences in allele length using standard Southern blot analysis. It has previously been shown that the basic repeat unit is in itself variable, and that this may be used to map the internal structure of an allele. This method has already been used to establish that alleles of the same length may have differing internal structures between non-related individuals. I now extend this approach to demonstrate that two apparently homozygous individuals are in fact heterozygotes. For each individual the two comigratory alleles were separated, without cloning, using single molecule dilution (SMD) of genomic DNA and recovery with PCR. Mapping of the variant repeat units revealed highly diverged internal structures and, for one individual, a size difference of one repeat unit (29bp). SMD and PCR recovery provide an efficient system for separating comigratory alleles without prerequisite for knowledge of sequence differences. These results demonstrate the poor resolving power of agarose gel electrophoresis and the associated limitations of allele length analysis.

Introduction
MS32. The human minisatellite MS32 was cloned from a λ library with Sau3A1 size fractionated (5 to 15kb) inserts and detects alleles in the size range of approximately 2 to 30kb, with an allele length heterozygosity of ~97.5% (Wong et al., 1987). In situ hybridization and linkage mapping have localised MS32 to an interstitial position on the long arm of chromosome 1 (1q42-43) and it has been assigned the locus name D1S8 (Royle et al., 1987). MS32 is comprised of a tandem array of a 29bp G-rich sequence expanded from within a retroviral long terminal repeat (LTR)-like repeat element (Figure 3.1A; Wong et al., 1987; Armour et al., 1989b). As with other sequenced minisatellite loci (see Chapter 1, p8) MS32 shows variation in the internal sequence of the repeat array, incorporating two commonly polymorphic base substitutions (Figure 3.1B; Wong et al., 1987). Small (<5kb) MS32 alleles may be PCR amplified up to levels detectable on ethidium bromide stained agarose gels and alleles up to approximately 10kb may be amplified and detected by Southern blot hybridization (Jeffreys et al., 1988a; see Figure 3.1A for primers). Dilution of bulk genomic DNA has been used to show that single molecules of MS32 alleles may be faithfully amplified and that this approach can be used to recover individual single MS32 alleles from heterozygous individuals (Jeffreys et al., 1988a, 1990).

Internal mapping at MS32. In order to examine detailed allelic structure at MS32, a method has been developed to assay the dispersion patterns of two minisatellite variant repeats (MVRs). Fortuitously, one of the major sites of variation within MS32 repeats, an A-G transition, creates or destroys a HaeIII restriction
Figure 3.1. The human minisatellite MS32: Structure, repeat unit and internal mapping.

(A) Structure of an example MS32 allele. Showing repeat units and derived internal map (open boxes, t-type repeats and shaded boxes, a-type repeats), the retroviral LTR (hatched box), restriction enzyme sites (S=Sau3A1, F=HinfI and H=HaeIII) and PCR primers (black boxes). Primer sequences are: A, 5' - TCACCGGTGAATTCCACAGACACT - 3'; B, 5' - AAGCTCTCCATTTCCAGTTTCTGG - 3'; C, 5' - CTTCCTCTGTTCTCCTCAGCCCTAG - 3'; D, 5' - CGACTCGCAGATGGAGCAATGGCC - 3'. Derivatives C1 and D1 incorporate a 5' extension TCACCGGTGAATTC containing an efficiently cleaved EcoRI restriction site (underlined), which was used to generate PCR products with a unique EcoRI site suitable for end labelling and mapping.

(B) MS32 Repeat Unit. Sequence of the MS32 repeat unit, showing the two polymorphic base substitutions, the variant HaeIII site and the constant HinfI site.

(C) Diagrammatic representation of an internal mapping autoradiograph of the example MS32 allele. MVR mapping is achieved by amplification of alleles with primers C1 and D (left) or C and D1 (right), end labelling at the EcoRI site and partial digestion with HaeIII (H) and HinfI (F), followed by agarose gel electrophoresis and autoradiography (for methodological details see Jeffreys et al., 1990). The derived internal map of a- and t-type repeats is also shown.
enzyme site, thus defining two classes of MVR: those cut by \textit{HaeIII}, designated \textit{a}-type; and those not cut by \textit{HaeIII}, designated \textit{t}-type. In conjunction with a constant \textit{Hinfl} site present in every repeat unit this variant restriction site can be utilized to map the dispersion pattern of the two MVRs within PCR amplified alleles (Jeffreys et al., 1990). End labelled PCR amplified MS32 alleles are partially digested with both \textit{HaeIII} and \textit{Hinfl} (separate reactions), electrophoresed side by side on an agarose gel, dried and autoradiographed (Figure 3.1C). For the \textit{Hinfl} digest a complete ladder of products derived from every repeat unit is seen, whilst for the \textit{HaeIII} partial digest, bands are only observed for those positions with MVRs incorporating the \textit{HaeIII} restriction site, \textit{i.e.} \textit{a}-type repeats. The observed pattern allows the exact number of repeats to be determined, each repeat to be designated as either \textit{a}-type or \textit{t}-type and a complete binary code for the dispersion pattern, or MVR map, of each allele to be established (Jeffreys et al., 1990). Mapping can be initiated from either end of the allele, allowing alleles up to about 5kb to be mapped in their entirety. This method has been used to show that extreme levels of internal variation exist within MS32 alleles and that alleles of a similar size shared by different individuals can have widely differing allelic structures and presumably widely diverged ancestral origins (Jeffreys et al., 1990). Very large numbers of alleles may be unambiguously identified, many more than the 100 or so alleles maximally distinguished using standard Southern blot allele length analysis.

\textit{Homozygote excess at minisatellite loci.} Calculations of match frequencies at minisatellite loci for forensic applications are based on estimated allele frequencies under the assumption that the relevant population is at Hardy-Weinberg equilibrium. However, data derived from a number of loci have shown that an apparent homozygote excess exists (Lander 1989; Budowle et al., 1991b) and this has been used to suggest that the populations concerned are not homogeneous, but divided into discrete subpopulations (Lander 1989; Cohen, 1990). A number of factors could be contributing to the apparent homozygote excess; these include population substructuring, as well as the limiting resolving power of agarose gel electrophoresis and the occurrence of undetectable or null alleles. Statistical re-analysis of the data based on analysing only information derived from heterozygous individuals has demonstrated that no significant departure from Hardy-Weinberg equilibrium exists (Devlin et al., 1990). Furthermore, theoretical calculation of the level of substructuring required to produce the observed departures have shown these to be very high and inconsistent with the known demographic origins of the populations (Chakraborty and Jin, 1992). The existence of undetectable null alleles has been demonstrated for at least two minisatellite loci (Armour et al., 1992) and it seems likely the majority of the apparent excess of homozygotes is due to the inability to resolve closely spaced alleles (Devlin et al., 1990; Chakraborty and Jin, 1992).

\textit{This work.} This chapter describes the use of single molecule PCR and variant repeat mapping in the analysis of two apparently homozygous individuals at the human minisatellite MS32. I demonstrate that on the basis of allelic structure both individuals are in fact true heterozygotes, one with identical length alleles and the other with alleles differing in size by a single repeat unit. The contents of this chapter have been published (Monckton and Jeffreys, 1991a).

\textbf{Results}

Screening of MS32 across large panels of unrelated people revealed two individuals (AS89 and MACH, of Pakistani and Chinese descent respectively) who by \textit{Sau3AI} restriction digestion and Southern blot analysis appeared to be homozygous for an approximately 2.5kb allele (work performed by A.J. Jeffreys, I. Patel, R. Neumann and myself; data not shown). PCR amplification of MS32 from total genomic DNA revealed single discrete bands on an ethidium bromide stained gel for both individuals (Figure 3.2A). To determine if these...
Figure 3.2. Separation of minisatellite alleles in pseudohomozygotes by single molecule dilution and PCR.

(A) MS32 alleles amplified from total genomic DNA. For each individual 20ng blood DNA was PCR amplified in a 10μl reaction using the nested primers C1 and D corresponding to the flanking region of MS32. Amplified alleles were electrophoresed through a 1% agarose gel in 1x TAE and visualised by ethidium bromide staining. Lane λ = lambda HindIII marker DNA, 0 = zero DNA negative control, M = amplified alleles from individual MACH, A = amplified alleles from individual AS89.

(B) MS32 isoallele separation by single molecule dilution. For each individual, genomic DNA was diluted in 5mM Tris-HCl (pH7.5) in the presence of 0.1μM PCR primers. Ten 10μl PCR reactions containing either 6, 3, or 0 pg DNA were amplified for 28 cycles using the minisatellite flanking primers A plus B. Products were detected by Southern blot hybridization with an MS32 repeat unit probe. Positive (+) and negative (-) PCR reactions are indicated.
individuals were indeed true homozygotes I attempted to map the internal structures of these alleles. Internal mapping of these alleles from total genomic DNA produced an ambiguous autoradiograph with widely differing band intensities, presumably reflecting a composite internal map derived from two comigratory but discrete alleles (Figure 3.3A, lanes marked T). Since the alleles were inseparable by standard agarose gel electrophoresis, the individual alleles were separated by single molecule dilution (SMD) of genomic DNA (Jeffreys et al., 1988a, 1990; Ruano et al., 1990) and recovery with PCR (Saiki et al., 1988). For each individual, total genomic DNA was diluted to 6pg/μl (two haploid genome equivalents per μl) and 3pg/μl (one haploid genome equivalent per μl). One microlitre aliquots of both dilutions of DNA were used to seed ten 10μl PCR reactions with the outside flanking primers A and B (as well as ten zero DNA controls). After 28 cycles of amplification a 5μl aliquot was removed, electrophoresed on a 1% agarose gel and Southern blot hybridized with an MS32 repeat unit probe (Figure 3.2B). Internal nested primers, 32-C1 plus 32-D and 32-C plus 32-D1, (primers C1 and D1 are derivatives of C and D that incorporate a 5' extension containing an EcoRI restriction site, allowing end specific labelling of PCR products) were used to re-amplify the alleles from each of the presumptive single molecule positive reactions up to a level visible on an ethidium stained gel, a further 25-28 cycles. All four alleles were completely MVR mapped from both ends by HaeIII and Hinfl restriction analysis using DNA derived from at least three separate single molecule reactions for each allele (Figure 3.3B).

Individual AS89 was found to have two alleles of identical size, containing 71 repeat units, but of widely differing internal structures. This individual is therefore a true heterozygote at this locus. Analysis of the second individual MACH, of Chinese origin, revealed this person to be a compound heterozygote at this locus, each allele having widely diverged internal structures and a length difference of one repeat unit (allele lengths 72 and 73 repeat units).

Discussion

This method for internal mapping from single molecules does not involve cloning of single amplified molecules (minisatellites are frequently unstable on cloning in E. coli), but samples the average properties of the entire amplified pool of PCR products. For each allele at least three separate single molecule amplifications were performed and in no case was any MVR map discrepancy found between them. As noted in previous experiments (Jeffreys et al., 1990), it appears that Tag polymerase induced misincorporation errors are not a problem with single molecule minisatellite mapping. Finally, in each case the superimposition of the two separated alleles gave rise to the same composite map as derived from total genomic DNA, confirming that the single molecule alleles recovered were the true alleles present in each individual.

Single molecule dilution and PCR recovery is a quick and efficient method for separating comigratory alleles of the same locus. Moreover, it has several further advantages over cloning direct from a genomic library, requiring a very minimal quantity of input DNA, as well as a suitability for isolation of products unstable in a bacterial host. Likewise, it offers advantages over cloning from a PCR derived library in that it has very low probability of incorporating polymerase errors and removes the potential problem of heteroduplex-derived chimeric "alleles". It also assumes no knowledge of internal sequence difference and thus allows isolation of uncharacterised polymorphisms. The limiting factor in the applicability of this method is the maximum size of amplifiable products, approximately 5kb for amplification of minisatellites up to bands detectable on ethidium bromide stained gels.

Internal mapping of minisatellites greatly enhances the effective resolution of an already highly informative set of markers, as demonstrated here by its ability to distinguish pseudohomozygotes from true
Figure 3.3. Internal mapping of PCR amplified MS32 alleles.

(A) Internal map autoradiographs. Cl plus D PCR products, labelled at the 3' end, were partially digested with Hinfl (F) or HaelIII (H), electrophoresed on a 1.2% agarose gel, dried and autoradiographed. Lanes marked A and B are single molecule derived alleles, lanes marked T are internal maps derived from total genomic DNA. See Jeffreys et al., 1990 for detailed methods.

(B) Encoded maps of all four single molecule derived alleles. Where a = repeat unit cleaved by HaelIII and t = repeat unit not cleaved by HaelIII.
homozygotes. These results establish directly that individuals with two closely spaced alleles may not be
distinguished as heterozygotes by standard Southern blot length analysis, providing supportive evidence for the
conclusion that the homozygote excess observed at minisatellite loci is largely an artifact of agarose gel
electrophoresis (Devlin et al., 1990; Chakraborty and Jin, 1992). The alleles involved are relatively short
(~2.5kb) for MS32 yet still indistinguishable by agarose gel electrophoresis. Larger alleles are likely to be even
less well resolved with larger differences remaining undetected. This illustrates well the quasicontinuous nature of
minisatellite length alleles, the limitations of length analysis via agarose gel electrophoresis and the resultant
inability to define precisely allelic state.

The further demonstration that alleles of the same length may have widely differing internal structures
highlights the problems associated with studying length variation in terms of population evolution. The
convergent nature of minisatellite allele length evolution impedes the effective use of length analysis in studies
of population divergence, since alleles of the same size cannot be assumed to have the same ancestral origin.
Internal variation however, has a hugely increased potential for distinguishing alleles and in addition allows
alleles of differing lengths to be related by zones of internal structural identity, which can thus be assumed to
share a recent common ancestor (Jeffreys et al., 1990, 1991b). The application of internal variation analysis to
allelic structure and population divergence is discussed further in Chapter 6.
Chapter 4

SINGLE STRANDED MOBILITY POLYMORPHISMS OF MINISATELLITE ISOALLELES

Summary
Over recent years a variety of techniques have been developed for the analysis of base substitutional polymorphisms in genomic DNA, most based on the use of PCR and polyacrylamide gel electrophoresis. With the exception of standard RFLP analysis, all require either knowledge of sequence differences, or are limited to the scanning of small segments of DNA derived from clones or PCR products. Minisatellite isoallelism, ie minisatellite alleles which are of indistinguishable length, or of the same length but different internal sequence, is a common problem associated with minisatellite length analysis. Internal sequence variation can be used to distinguish such isoalleles, but present systems require detailed knowledge of its sequence basis. In addition, minisatellite alleles are often too large to be analysed by conventional approaches to the identification of base substitutions. In this chapter I describe the analysis of minisatellite isoalleles by the use of single stranded mobility polymorphisms (SSMP) on agarose gels. SSMP can be used to distinguish minisatellite isoalleles after PCR amplification, or in total genomic DNA, by standard Southern blot analysis.

Introduction
Methods for detecting single base substitutions. A number of systems have been developed for detecting single base substitutions in genomic DNA and all have associated advantages and disadvantages. Of these DNA sequencing is ultimately the most powerful, but has several practical limitations. Sequencing can be performed on individual clones, highly impractical for population screens, or directly on PCR amplified products (Gyllensten and Erlich, 1989) which pre-requires sequence information and is limited to relatively short genomic fragments. Standard RFLP analysis can be used to scan quite large DNA fragments, but the majority of base substitutions do not affect restriction enzyme sites (Bowcock and Cavalli-Sforza, 1991; Nickerson et al., 1991) and the whole approach is relatively inefficient. Where sequence differences have been previously characterised they may be detected in PCR products by RFLP analysis, hybridization with allele specific oligonucleotide probes (Saiki et al., 1986) and oligonucleotide ligation analysis (Nickerson et al., 1991), or by direct PCR analysis using allele specific primers (Newton et al., 1989; see Chapter 5). A number of recent approaches utilizing PCR have been developed to identify new single base substitutions, these include denaturing gradient gel electrophoresis (DGGE, Sheffield et al., 1989), chemical mismatch analysis (Montandon et al., 1989) and single stranded conformational polymorphism (SSCP, Orita et al., 1989). All of the PCR based approaches are limited by the maximum size of DNA sequence amplifiable, and those utilizing polyacrylamide gel electrophoresis are limited by the maximum size of product that may be resolved by such a system (<1000bp). DGGE analysis using total genomic DNA is also possible, via heteroduplex analysis (Myers et al., 1985) and...
blot analysis (Krolewski et al., 1992), but both methods still rely on the use of polyacrylamide gel electrophoresis. Of the systems developed for the identification of new polymorphisms ie DGGE, chemical mismatch and SSCP, neither has a 100% detection rate under a single set of conditions; being dependent on the actual base substitution, its specific sequence context and the exact experimental conditions employed.

One of the most widely used approaches is SSCP, whereby PCR amplified DNA is labelled, restriction digested, denatured and electrophoresed on a non-denaturing polyacrylamide gel (Orita et al., 1989). Under non-denaturing conditions a single stranded DNA molecule will hybridize with itself to produce a unique, sequence specific, 3-dimensional conformation, which directly effects its electrophoretic mobility. Thus, the mobility of a single stranded DNA fragment is dependent not only on its length, but also on its sequence, such that DNA segments of the same length, but different sequence, can sometimes be distinguished on the basis of their single stranded electrophoretic mobility. This technique can be used to detect the majority (estimated around 80%) of base substitutions in short (<1000bp) fragments, provided that a range of electrophoretic conditions and restriction enzymes are used (Orita et al., 1989).

Minisatellite Isoallelism and Internal variation. Minisatellite isoalleles are defined as minisatellite alleles indistinguishable by standard Southern blot length analysis. The inability of agarose gel electrophoresis to separate closely spaced minisatellite alleles is a recognised limitation of standard length analysis (Devlin et al., 1990; Chakraborty and Jin, 1992; see Chapter 3) and I have already demonstrated that pseudohomozygous individuals containing two closely sized alleles can be mis-scored as homozygous using such a system. In addition, the convergent nature of minisatellite length evolution means that alleles of the same length can arise through independent allele lineages (Jeffreys et al., 1990) and that pseudohomozygous individuals containing two identically lengthed alleles, but differing in internal structure, can be identified (Monckton and Jeffreys, 1991a; Chapter 3). The existence of frequent internal variation in minisatellite alleles has been well documented (see Chapter 1, p8); this, accompanied by the more recent demonstration that internal variation adds a powerful new dimension to minisatellite analysis, allows huge numbers of alleles to be identified on the basis of their internal structure (Jeffreys et al., 1990, 1991; see Chapters 3 and 6). Unfortunately, internal variation analysis requires either direct sequence analysis (Owerbach and Aagaard, 1984; Gray and Jeffreys, 1991) or detailed knowledge of the sequence variation (Jeffreys et al., 1990, 1991; see Chapters 3 and 6). For the majority of minisatellite loci such detailed sequence data does not exist and furthermore, where it does, the actual variation observed is often not amenable to the types of analysis previously utilized, ie base substitutions do not affect restriction sites or internal MVR variation is too extensive to allow the design of consensus MVR-PCR primers (Jeffreys et al., personal communication). Thus a more general approach to minisatellite internal variation analysis would prove useful, in particular with respect to studying minisatellite isoalleles. As described above, a variety of systems for detecting base substitutions in genomic DNA already exist, but they are severely limited in their application to minisatellite analysis by their basis on PCR and use of polyacrylamide gel electrophoresis. Many minisatellite alleles are too large to be PCR amplified and the vast majority are too large for effective resolution on polyacrylamide gel systems.

This work. In this chapter I describe the analysis of minisatellite isoalleles by the use of single stranded mobility polymorphisms (SSMP) detected by agarose gel electrophoresis. I demonstrate that minisatellite isoalleles containing multiple sequence differences can often be separated on the basis of their single stranded mobility on agarose gels. This type of analysis can be applied not only to PCR amplified alleles, but also to total genomic DNA via standard Southern blot analysis. Thus the technique is limited in length only by

Chapter 4 Page 2
the size of single stranded DNA that may be efficiently prepared. Part of this work has been presented (Monckton and Jeffreys, 1991b).

**Results**

**Single stranded mobility polymorphisms of PCR amplified MS32 alleles.** As described in Chapter 3, I have previously identified two individuals who by standard Southern blot analysis of genomic DNA appear to be homozygous for an approximately 2.5kb MS32 allele, but who are in fact both true heterozygotes. Individual AS89 is heterozygous for two identically sized alleles (71 repeats) and MACH is a compound heterozygote having alleles differing in length by a single repeat unit (29bp, 72 and 73 repeats). Assuming each difference in repeat type between isoalleles represents at least one site of base substitution, then the number of sequence differences between them can be determined. The existence of further sequence variation not affecting either of the two restriction sites assayed in MVR mapping cannot be ruled out and thus this number represents the minimum number of sequence differences. The two pairs of AS89 and MACH isoalleles differ by at least 23 and 31 base substitutions respectively.

Internal mapping by restriction analysis is a relatively tedious and technically demanding method. In contrast, SSCP is relatively easy to perform, very sensitive (capable of detecting single base changes in small PCR amplified fragments) and requires no knowledge of sequence differences (Orita et al., 1989). The pseudohomozygous individuals described above provide ideal substrates for testing new methods of distinguishing minisatellite isoalleles that do not involve knowledge of sequence differences. However, the relatively small MS32 alleles involved (~2.5kb) are still too large to be effectively resolved by standard SSCP analysis (Katrina MacKay, personal communication).

In an attempt to reproduce the SSCP effect for large fragments containing multiple base substitutions, PCR amplified MS32 isoalleles were specifically labelled, heat denatured and electrophoresed on a non-denaturing agarose gel. For both individuals single molecule separated alleles and mixed alleles amplified from total genomic DNA were labelled on either and/or both strands using the flanking primers C1 and D1. Flanking primers C1 and D1 incorporate 5' extensions containing an EcoRI restriction site. Digestion with EcoRI, followed by end filling with α-32P-dATP allows strand specific labelling of C1/D1 amplification products. Labelled PCR products were heat denatured in formamide and electrophoresed through a 1% agarose gel at room temperature. For both individuals the isoalleles were clearly resolved by this type of analysis (Figure 4.1). MS32 repeats show a marked strand bias in nucleotide content, with one strand comprising 45% C and the complementary strand only 17% C. Interestingly, for both pairs of isoalleles the C-rich strand gave sharper, more discrete bands and clearly resolved the two pairs of isoalleles. In contrast, the G-rich strand, produced less discrete, broad ‘fuzzy’ bands. The two MACH isoalleles (differing by a 29bp length change and multiple sequence changes) were resolved with the G-rich strand, but the AS89 isoalleles (multiple sequence differences, but same length) were not. A further experiment using mixed DNA samples showed that all four isoalleles from both individuals could be effectively resolved by PCR-SSCP of the C-rich strand (data not shown). Heat denaturation at 100°C in standard loading dye and alkaline denaturation (1M NaOH) was also tested, but resulted in increased DNA degradation producing blurred bands (data not shown).

**SSMP of MS32 alleles on Southern blot analysis of total genomic DNA.** In order to determine if SSMP could also be applied to total genomic DNA, six samples were analysed using a similar technique. For each sample 10µg of genomic DNA was digested with Mbol and 5µg loaded onto an agarose gel as normal (ie double stranded DNA). The remaining 5µg was heat denatured in formamide before loading (ie
MS32 alleles were PCR amplified using primer pair C + D, for primer sequences see Chapter 3. Primer derivatives C1 and D1 contain EcoRI extensions which allow strand specific labelling of products. Primer C1 labels the G-rich strand and primer D1 labels the C-rich strand. 100pg of each end labelled PCR product (double stranded (ds) and single stranded (ss)) were electrophoresed on a 37cm 1.2% agarose gel in 1x TBE buffer at 150V for 16 hours, after which the gel was dried and autoradiographed overnight. Samples ran as single stranded DNA were dissolved in an 80% formamide buffer and heated to 80°C for 3 minutes prior to loading. A and B = alleles separated by single molecule dilution (see Chapter 3). T = Alleles A+B amplified from total genomic DNA.
single stranded DNA). The samples were electrophoresed, blotted and hybridized with an MS32 repeat unit probe as for normal minisatellite analysis (Figure 4.2). Two of the samples (CAWE and DAMO) were blood DNAs derived from normal individuals, heterozygous at MS32 and produced four single stranded bands (two sharp and two broad bands), two from each allele as expected. The four other samples were DNAs derived from individuals apparently homozygous for MS32 alleles, including individual MACH, previously demonstrated as heterozygous by internal mapping and PCR-SSMP analysis. MACH produced a four band single stranded pattern paralleling the results obtained by PCR-SSMP. The three remaining apparent homozygotes (GOMC, CEPH 133411 and CEPH 3701) had MS32 alleles too large to be efficiently amplified by PCR (>5kb), thus preventing direct restriction mapping or SSMP analysis of amplified alleles. The apparent homozygotes GOMC and 133411 produced a three band single stranded pattern (two sharp and one broad) indicating that these individuals are also true heterozygotes containing alleles either too closely sized to be resolved by electrophoresis, and/or of differing internal sequence structure. The fourth apparent homozygote, 3701, produced only a two band single stranded pattern, suggesting that this person may be a genuine MS32 homozygote. Subsequent analysis of all four apparent homozygotes by MVR-PCR (Jeffreys et al., 1991; see Chapter 6) has confirmed the results obtained here. Interestingly, for all samples a mixture of sharp and fuzzy bands were obtained, with the isoalleles being preferentially resolved by the sharper bands, presumably reflecting the same strand specificity as observed for MS32 PCR-SSMP. The signals obtained from the two single stranded CEPH samples were noticeably weaker, with an increased background smear, than their double stranded signals and the single stranded samples of the other DNAs. The CEPH DNAs represent a major laboratory resource and are frequently used, undergoing multiple rounds of freeze/thawing and it seems likely that the weaker signals represent a reduced single stranded DNA size, presumably a result of DNA nicking, for the CEPH DNAs compared to their double stranded size and the single stranded size of freshly prepared DNA.

**PCR-SSMP at MS205:** Segregation analysis. MS205 is a hypervariable human minisatellite with a length heterozygosity of 96% and an allele size range of approximately 2-5kb, as determined by Southern blot analysis of Hinfl digested human DNA (Royle et al., 1992). The narrow allele size range observed allows all alleles to be PCR amplified and a restriction mapping based analysis, similar to that employed at MS32 (see Chapter 3), has revealed sequence variation within the basic 54bp repeat unit and extensive variation of MVR dispersion patterns within alleles (Armour et al., manuscript in preparation). Southern blot analysis of the 40 CEPH families probed with MS205 revealed 12 apparently homozygous parents and one family for which both parents appeared to share two alleles of the same length and for which the majority of children appeared homozygous (Royle et al., 1992, unpublished data). All of the apparently homozygous parents were re-typed by PCR-Southern blot analysis and by PCR-SSMP and with two exceptions, all were demonstrated to be genuine allele length heterozygotes, with allele length differences of 1-2 repeats (50-100bp, for examples see Figure 4.3). Their original mis-scoring as homozygous demonstrates the inaccuracy of standard Southern blot analysis, exacerbated by use of an inappropriate restriction enzyme (Mbol, which incorporates an extra 2.4kb of flanking sequence, increasing allele size and reducing resolving potential). By SSMP analysis all of the heterozygotes produced at least three single stranded bands, two sharp bands and one or two broad bands; very closely spaced alleles often having only one broad band (eg individuals 3501 and 10201, Figure 4.3). MS205 displays a similar level of nucleotide content strand bias as MS32, with a 57% C-rich strand complementing a 19% C strand (Armour et al., manuscript in preparation). This strand asymmetry may be reflected in the variation in single stranded band width and resolving power observed for the two strands of single MS205 alleles. Of the two
Figure 4.3. MS205 apparent homozygotes: double stranded and SSMP-PCR analysis.

For each individual 100ng of genomic DNA was amplified with PCR primers 205-A and 205-B, for 16 cycles, with a 67°C annealing temperature and a 10 minute extension time in a 10μl reaction. Half of the reaction was loaded as double stranded (ds) DNA and half heat denatured in 80% formamide at 80°C for 3 minutes and loaded as single stranded (ss) DNA. Samples were electrophoresed on a 37cm 1.2% agarose gel in 1x TBE at 150V for 16 hours and Southern blot hybridised with an MS205 repeat unit probe. Primers 205-A and 205-B are unique sequence primers located in the flanking DNA of MS205 allowing amplification across the tandem array (Armour et al., manuscript in preparation), sequences are: 205-A 5’-CTGTGCGGTCACGGCAGGCTGGAC-3’ and 205-B 5’-AGCGGCAGCCCAGACTCCACATGG-3’.
apparent length homozygotes neither could be demonstrated as heterozygous by SSMP (eg 3502, Figure 4.3). For family 66 both parents were typed by PCR-Southern blot analysis (Figure 4.4A) and both parents and all the children were typed by PCR-SSMP (Figure 4.4B). Neither of the shared alleles had resolvable double stranded length differences, but one pair of isoalleles could be resolved by SSMP and its Mendelian segregation is demonstrated (Figure 4.4B). Strand specific oligonucleotide hybridization was used to demonstrate the informative band as deriving from the C-rich strand (data not shown). Subsequent restriction digest based internal mapping has revealed that both pairs of isoalleles are in fact different (data not shown; Armour et al., manuscript in preparation). Based on differences in internal map structure the longer, separable pair of isoalleles (both probably 53 repeats) contain at least 16 base substitutions, whilst the shorter (both 42 repeats), inseparable alleles incorporate a minimum of only 3 base differences.

SSMP analysis for MS1. MS1 is the most variable human minisatellite isolated to date, with a >99% heterozygosity and a 5% per gamete germline mutation rate (Wong et al., 1987; Jeffreys et al., 1988b). It consists of tandem arrays, 140 to >2,500 repeats, of a highly variable 9bp repeating unit and smaller alleles may be PCR amplified in their entirety (Wong et al., 1987; Gray and Jeffreys, 1991). The two most common human MVRs displays extreme strand asymmetry with a 55% C-rich strand complemented with either a completely C-devoid strand or an 11% C strand (Gray and Jeffreys, 1991). Southern blot analysis of MS1 in the 40 large CEPH families revealed four individuals apparently homozygous for large MS1 alleles (alleles >9kb, Jeffreys et al., 1988b). Genomic double stranded and SSMP Southern blot analysis of DNA from these individuals revealed two of these people to be heterozygous for a short allele (<2.5kb) not detected on the original blot (data not shown). No large discrete single stranded bands were observed for any of the CEPH DNAs, but a high background smear was seen. A heterozygous control DNA did produce two discrete single stranded bands and two small single stranded bands were observed for the heterozygotes, suggesting the main reason for lack of discrete large single stranded bands was DNA nicking.

Four small PCR amplifiable MS1 length isoalleles, previously identified as different by sequencing (Gray and Jeffreys, 1991), were analysed by PCR-SSMP Southern blot analysis. The PCR products were hybridized separately with a repeat unit probe (Figure 4.5A) and the two oligonucleotide PCR primers used to amplify the segment (Figure 4.5B, C). Double stranded length analysis demonstrated that all four alleles were essentially indistinguishable. However, SSMP analysis allowed at least three of the alleles to be clearly resolved from the others with the C-rich strand. The G-rich strand produced no clear differences between the alleles.

SSMP analysis of unique sequence DNA. SSMP analysis has been successfully applied to the detection of base substitutions between relatively large minisatellite isoalleles. If SSMP analysis could also be applied to unique sequence DNA then it may allow the rapid screening for polymorphisms of DNA fragments much larger than practical with current methods. To this end SSMP analysis was attempted at two unique sequence loci, one containing known polymorphisms and the other of unknown polymorphic content.

SSMP analysis was performed on a PCR amplified 1.2kb fragment flanking the human minisatellite MS1 (see Chapter 8) in 6 unrelated individuals; no mobility differences were detected in either strand (data not shown). DNA amplified from a 350bp region flanking MS32 (see Chapter 5) from three individuals, known to be homozygous (A), heterozygous (A/B) and homozygous (B) for a single base substitution were also analysed by SSMP and by standard SSCP analysis using polyacrylamide gel electrophoresis. SSMP analysis produced no detectable differences (data not shown), whilst standard SSCP analysis demonstrated the known base substitution and detected a further one (see Chapter 5).
Figure 4.4. Segregation analysis of MS205 isoalleles.

(A) Double stranded PCR amplified MS205 alleles. The MS205 alleles were amplified from the two parents (01 and 02) of CEPH family 66 and double strand (ds) Southern blot hybridized as detailed in Figure 4.3.

(B) SSMP analysis of PCR amplified MS205 alleles. The MS205 alleles were amplified from CEPH family 66 and single stranded (ss) Southern blot hybridized as detailed in Figure 4.3. The two single stranded bands from each allele are indicated and the double stranded and single stranded derived genotypes are also shown.
Figure 4.5. SSMP analysis of PCR amplified MS1 isoalleles.

100pg of gel purified small (~1.2kb) MS1 isoalleles PCR amplified from four unrelated individuals were electrophoresed (ds and ss) on a 37cm 1.2% agarose gel in 1xTBE for 16 hours at 150V and Southern blot hybridised with: (A) Oligolabelled double stranded repeat unit probe (10ng of the small allele PCR amplified from individual 1); (B) Kinase labelled oligonucleotide probe MS1-A, detecting the G-rich strand; and C, Kinase labelled oligonucleotide probe MS1-B, detecting the C-rich strand. PCR amplifications were performed using primers MS1-A and MS1-B on 100ng of genomic DNA in a 20µl PCR reaction with a 63°C annealing temperature and a 5 minute extension time for 28 cycles. Primers A and B are unique sequence primers located in the flanking DNA of MS1 and allow amplification across the tandem array (Gray and Jeffreys, 1991), sequences are: MS1-A 5’-GCTTTTCTCTGATGAAGCGCTTGATG-3’ and MS1-B 5’-AAGAGGGCATATGGCAACCCCATGGAAGG-3’.
Discussion

Apparent homozygosity is a common problem associated with standard Southern blot minisatellite length analysis, as demonstrated here by, a) the failure to detect small length differences between MS205 alleles (83% of apparent homozygotes at MS205 were demonstrated to be true length heterozygotes by PCR length analysis) and b) the non-detection of some small alleles at MS1, which can result in individuals being erroneously scored as homozygous, or which can appear as 'null' alleles in pedigree analysis (for two further extreme examples see Armour et al., 1992). PCR amplification often increases the resolving potential of length analysis (due to the incorporation of less flanking DNA, producing smaller products), but has limitations of its own, requiring sequence knowledge to design amplimers and often fine tuning of reaction conditions to produce high yields of detectable authentic products (Jeffreys et al., 1988a). PCR amplification is particularly useful for detecting very small alleles which cannot be detected by Southern blot analysis of genomic DNA(Armour et al., 1992), but can generate nulls of a new kind where alleles too large to amplify exist, a common phenomena at most minisatellite loci. Internal structure analysis can be used very effectively to increase the resolving potential of minisatellites (Jeffreys et al., 1990, 1991; Monckton and Jeffreys, 1991; Armour et al., personal communication; Neil et al., personal communication), but requires detailed sequence knowledge and is not applicable to many loci. Furthermore, the large size of most minisatellite alleles renders them immune to the usual approaches used for detecting base substitutions.

Isoalleles from pseudohomozygotes of the human minisatellite MS32 have previously been identified and the individual alleles separated by single molecule dilution and PCR recovery (Monckton and Jeffreys, 1991; Chapter 3). I have demonstrated that these isoalleles may also be distinguished, very simply, on the basis of their single stranded mobility on non-denaturing agarose gels. Multiple sequence differences and/or small length differences give rise to significant changes in mobility, allowing isoalleles to be clearly resolved. SSMP can be applied not only to PCR amplified alleles, but also directly to restriction digested total genomic DNA representing a significant advance in the analysis of large non-amplifiable minisatellite alleles. At MS32 this type of analysis has been used to increase the already very high length heterozygosity, with the demonstration that three out of four apparent length homozygotes could be revealed as true heterozygotes. The remaining apparent homozygote has since been demonstrated to be a true homozygote by MVR-PCR analysis.

SSMP analysis at MS205 has demonstrated the expected normal Mendelian inheritance of this type of polymorphism and in addition, the increased resolving power of SSMP in distinguishing isoalleles has extended the linkage information that may be derived from what was a relatively uninformative family. In the family studied, two pairs of isoalleles existed, only one pair of which could be distinguished by SSMP. The separable alleles differed by at least 16 base substitutions, whilst the non-separable alleles were highly related, differing at only three positions. Thus the resolving potential of SSMP, not surprisingly, would appear to correlate with the actual extent of sequence divergence present. The two apparent length homozygotes were not resolved by SSMP analysis and in addition the close length heterozygotes were not more clearly resolved by SSMP. Relative to MS32, MS205 has a lower variability in allelic structure with the existence of a few relatively common alleles (only 19 different alleles in a sample of 26 alleles so far analysed). Furthermore, most alleles, although different, share extensive regions of internal homology and can be arranged into groups of aligned alleles, all of which are assumed to share a recent common ancestor (Armour et al., manuscript in preparation). As such it seems more likely that in many cases closely sized alleles may be highly related, incorporating a minimal number of sequence differences and hence less amenable to SSMP analysis.
In the majority of cases small MSI PCR amplified isoalleles could also be satisfactorily separated by SSMP analysis. Unfortunately the poor single stranded size of the CEPH DNAs did not allow a satisfactory test of MSI isoallelism in total genomic DNA, although considering the results obtained with the small alleles and the control DNA it seems likely that with suitable samples genomic-SSMP could be successfully applied to MSI.

Minisatellite isoallele SSMP studies are limited by the general paucity of homozygotes, pseudo or otherwise, available for examination. Ironically, it seems probable that for less variable minisatellites with lower heterozygosities apparent isoalleles are more likely to be identical or closely related and less amenable to SSMP analysis. Conversely, for the highly variable loci the more likely it will be that apparent isoalleles are in fact different and will contain multiple sequence differences which will enable them to be distinguished using SSMP.

The application of SSMP analysis to unique sequence DNA has been tested in the PCR amplifiable flanking regions of two minisatellites. In neither case, one of which contained two known base substitutions, were SSMPs observed. It seems probable that the low density of base substitutions found in unique sequence DNA is the limiting factor, rather than SSMP being a purely tandem repeat phenomena. Besides, ample satisfactory systems already exist for the identification of base substitutions in small fragments of unique sequence DNA.

For all the minisatellite loci studied by SSMP an extreme strand bias in terms of band width and resolving power has been observed; in all cases the C-rich strand has been sharper and of greater use. The G-rich strand almost always produces broader, fuzzier bands and often fails to resolve isoalleles. It is probable that in most cases identically sized alleles are not separated by the G-rich strand, with closely sized alleles being only poorly resolved. G-rich sequences are known to be capable of forming atypical secondary structures, including quadruplexes, in vitro (Sen and Gilbert, 1988). It is possible that G-rich single stranded molecules adopt a number of such closely related stable conformations, dictated by the G-rich nature of the molecule, under the conditions used for SSMP, resulting in the observed broad bands. Such stable conformations could be less prone to sequence changes, not affecting its basic G-rich nature, producing the observed inability to resolve isoalleles.

The sensitivity of SSMP analysis is difficult to assess, since without detailed sequence analysis of all alleles involved the true level of variation between alleles can only be estimated. Furthermore, the exact nature of differences is also variable, ie repeat unit sequence differences and/or length differences, with the relative contribution of each factor being un-assignable without extensive controlled studies. Nevertheless, as a general rule it seems probable that the larger the number of differences, the more likely isoalleles are to be separated by SSMP. In reality the formulation of strict rules governing applicability are impractical and an empirical approach for each locus would seem more rational.

In this chapter, I have demonstrated that many minisatellite isoalleles can be distinguished by differences in their single stranded mobility on non-denaturing agarose gels, with the new technique termed SSMP. Incidentally, I prefer the term mobility polymorphism, rather than conformational polymorphism, since such a system produces information only on the relative mobilities of two DNA fragments. Electrophoretic mobility of a DNA fragment is a product of both its length and its sequence dependent conformation; as yet no direct information regarding conformation may be derived from a mobility measurement. The technique can be very simply applied to PCR amplified alleles, or directly to restriction digested total genomic DNA, with no requirement for sequence knowledge, or any specialised chemicals or equipment. SSMP is applicable to all hypervariable minisatellite loci tested so far, with the major limitation being the requirement for nick free DNA.
when analysing large alleles directly from total genomic DNA. SSMP can be used to distinguish isoalleles in pseudohomozygous individuals, increasing informativeness in individual identity and linkage analysis, as well as comparing isoalleles from unrelated individuals. Compared to minisatellite loci, applications to unique sequence DNA would appear to be limited, probably a result of the minimal number of sequence changes found in such regions.
Chapter 5

BASE SUBSTITUTIONAL POLYMORPHISMS IN THE FLANKING DNA OF MS32 AND THEIR APPLICATIONS IN MVR-PCR

Summary

The recent development of minisatellite variant repeat (MVR) mapping via the polymerase chain reaction (MVR-PCR) as a digital approach to DNA typing promises to be an exciting new molecular tool for the analysis of individual identification and relationship testing, with very great potential in the field of forensic biology. MVR-PCR also offers the prospect of unravelling many of the questions surrounding minisatellite biology, including the generation and maintenance of the extreme levels of variation observed at some loci. The system assays the dispersion patterns of MVRs within minisatellite arrays and to date has been successfully applied to the human minisatellite MS32 (D1S8), for which two major classes of MVR have been identified. Using MVR specific amplimers and a specific primer located in the DNA flanking the minisatellite it is possible to generate a ladder of PCR products corresponding to the position of each MVR along an allele. The procedure can be applied to single separated alleles producing a code for the distribution of the two MVRs within the array, or it may be applied directly to total genomic DNA to produce a diploid digital code derived from the superimposed maps of the two individual alleles. Previously it has been shown that the level of phenotypic variation discernible between unrelated individuals using this method is enormous and that this ultravariability is based on the extreme variability in the MVR dispersion pattern of different alleles. Current methods for examining allelic variability involve either direct analysis of physically separated alleles or extraction of allelic haplotypes from pedigrees. In this chapter the identification of three common polymorphisms in the DNA flanking the most variable 5' end of MS32 is described, together with their use in developing allele specific flanking primers that may be used to map single alleles directly from total genomic DNA in heterozygous individuals (allele knockout). I go on to show how a similar approach may be used to obtain highly discriminatory MVR information from mixed DNA samples, of obvious forensic importance.

Introduction

MVR-PCR at MS32. The hypervariable human minisatellite MS32 has previously been demonstrated to consist of interspersed tandem arrays of two types of minisatellite variant repeats (Jeffreys et al., 1990; see Chapter 3). These two MVRs differ by a single base substitution that creates or destroys a HaeIII restriction enzyme site (designated a-type, cut by HaeIII and t-type, not cut by HaeIII) and coupled with a constant HinfI site the dispersion patterns of these two classes of MVR may be assayed by restriction analysis of PCR amplified alleles (Jeffreys et al., 1990; Chapter 3). This method has been used to demonstrate extreme levels of variation between MS32 alleles, even distinguishing length isoalleles in unrelated individuals (Jeffreys et al., 1990) and pseudohomozygotes (Monckton and Jeffreys, 1991a; Chapter 3). The method is however limited by
the maximum size of allele that may be PCR amplified (~5kb); furthermore the process is slow, laborious and technically demanding. Recently an improved, technically much simpler, method for assaying the same site of internal variation has been developed based on the use of MVR specific PCR primers (MVR-PCR, Jeffreys et al., 1991b). The theoretical basis for MS32 MVR-PCR is outlined in Figure 5.1 and in summary consists of using MVR specific amplimers and a specific primer located in the DNA flanking the minisatellite to generate a ladder of PCR products corresponding to the position of each a-type and each t-type repeat (Jeffreys et al., 1991b). The procedure can be applied to single separated alleles to give a binary code for the distribution of a-type and t-type repeats within the array, or it may be applied directly to total genomic DNA to produce a diploid ternary code derived from the superimposed maps of the two individual alleles. Each rung on a diploid code ladder may be coded as 1 (both a-type at that position, a/a), 2 (both t-type at that position, t/t) or 3 (heterozygous, a/t). The process is simple, rapid and gives unambiguous easily interpretable information in a digital format ideal for computer based analysis and databasing. The level of variation discernible using this method is enormous and the technique has great potential for use in individual identification, paternity testing, studies of minisatellite evolution and human population genetics (Jeffreys et al., 1991b; Tamaki et al., 1992).

*Null* repeats during MVR-PCR. During MVR-PCR at MS32 some repeat units fail to amplify with either the a-type or t-type specific primers, indicating the existence of rare MVRs that contain additional sequence variation preventing priming with the MVR specific TAG-A and TAG-T primers (Jeffreys et al., 1991b). During amplification of single alleles, these so-called 'null', or O-type repeats, present as gaps in the MVR ladder, whilst in diploid mapping they may be reliably identified (>90% efficiency) by intensity differences of a/O versus a/a and t/O versus t/t positions (Jeffreys et al., 1991b; Tamaki et al., 1992). These null repeats bring in three more coding states in diploid mapping: heterozygous a/O, code state 4; heterozygous t/O, code state 5; and homozygous null, O/O, code state 6. Runs of apparent null repeats also arise beyond the end of short alleles, but these nonexistent nulls can be reliably identified in both single allele and diploid MVR-PCR (Jeffreys et al., 1991b). Null repeats do not represent a significant problem for individual identification, but their correct identification is essential for kinship testing (Tamaki et al., 1992). Three null repeats have been sequenced and found to share the same single base deletion, defining a new class of repeat, N-type (Tamaki et al., 1992). N-type repeats may also be assayed by MVR-PCR using an N-type repeat specific MVR primer (32-TAG-N) and have been found to make up the majority of O-type repeats (>84% of all O-type repeats, Tamaki et al., 1992). One further type of variant repeat has also been sequenced, designated J-type, but it is rare having been identified in only two Japanese alleles (Tamaki et al., 1992). The remaining O-type repeats, not amplifiable with the N- or J-type specific primers, are termed U-type (unamplifiable) and are thought likely to represent a number of independent rare variants (Tamaki et al., 1992).

Allelic variation at MS32. The observed variation seen for diploid codes (no two people among 500 unrelated individuals so far typed share the same diploid code) is based directly upon the massive variation of individual alleles. The estimate for the minimum number of distinguishable alleles present in current Caucasian populations is around 6800 (see Chapter 6). The true number of different alleles is certainly in excess of this and may be as high as $10^8$ for the total world population (based on known mutation rate and population size, Jeffreys et al., 1991b; see Chapter 6). Allele mapping is providing remarkable insights into the evolution of minisatellites and the generation of new length alleles with, for the first time, preliminary evidence for the role of unequal interallelic exchange, or interallelic gene conversion, in the generation of new mutant alleles (Jeffreys et al., 1991b; see Chapters 6 and 7). We also predict that allele mapping will prove a valuable tool in the
Figure 5.1. The principles of minisatellite repeat coding.

(A) Principle of digital coding. Minisatellite alleles consisting of interspersed arrays of two variant repeat units termed a-type (shaded boxes) and t-type (open boxes). Individual alleles can be encoded as a binary string extending from the first repeat unit. In total genomic DNA, a corresponding ternary code of both superimposed alleles can be generated. At each repeat unit position, the alleles can be both a-type (code 1), both t-type (code 2), or heterozygous with one a-type and one t-type repeat (code 3).

B) MS32 repeat unit and MVR-PCR primers. The consensus 29bp repeat unit sequence of human minisatellite MS32 (D1S8) showing the polymorphic site which generates HaeIII cleavable (a-type) repeats and HaeIII-resistant (t-type) repeats. 32-TAG-A and 32-TAG-T are variant repeat specific oligonucleotides terminating at this polymorphic site. Each primer consists of 20nt minisatellite repeat sequence (bold) preceded by a 20nt 5' synthetic non-minisatellite extension identical to the TAG amplimer.

(B) The principle of MVR-PCR. Illustrated for a single allele amplified using primer 32-TAG-A. 1. At low concentration of primer, 32-TAG-A will anneal to approximately one a-type repeat unit per target minisatellite molecule and extend into the flanking DNA. 2. Amplimer 32D primes from the flanking DNA, creating a sequence complementary to TAG. 3. These DNA fragments terminating in 32D and the TAG complement can now be amplified using high concentration of 32D and TAG amplimers, to create a set of PCR products extending from the flanking 32D site to each a-type repeat unit. Use of primer 32-TAG-T at stage 1 will create a complementary set of products terminating at each t-type repeat unit.

This figure was adapted from Jeffreys et al. (1991b).
analysis of human population divergence through the generation of allele groupings from which it may prove possible to derive both allele and human population lineages (see Chapter 6).

**Allelic analysis by MVR-PCR at MS32.** The structure of individual MS32 alleles using MVR-PCR can be presently approached in two ways: First, by pedigree analysis of diploid codings and second, by mapping of individual separated alleles. Using family groups it is possible to derive incomplete allele maps from father, mother, single child trios and total unambiguous allele maps from father, mother and two children who share one allele in common. The use of such family groupings is however limited by availability and by the high de novo mutation rate of ~1% per gamete at this locus (see Chapter 7). Alternatively, individual alleles from one person may be separated on the basis of size, using restriction digestion and preparative agarose gel electrophoresis. This approach is time consuming, tedious and requires reasonably large amounts of DNA (minimum around 5|g total genomic DNA) plus the need for a preliminary experiment to determine allele sizes. Moreover this approach is difficult for individuals with closely sized alleles and pseudohomozygous individuals. Some of these problems may be obviated by single molecule dilution (SMD) and PCR recovery (Monckton and Jeffreys, 1991a; Chapter 3), but this procedure has its own limitations, the main one being that it is applicable only to relatively small alleles that may be amplified in their entirety.

**This work.** In this present study I used single stranded conformational polymorphism (SSCP) analysis, DNA sequencing and inter-species sequence comparisons to identify three common polymorphisms in the flanking DNA of MS32. The sequence information thus gained was used to design PCR based diagnostic tests for allelic state and, through the use of allele specific primers, haplotype specific MVR-PCR of MS32 alleles in heterozygous individuals (ie 'knockout' of one allele). I also show that haplotypic primers may be used to obtain unambiguous individual specific diploid codes, or unambiguous single allele codes, from mixed DNA samples, of obvious potential in forensic applications.

**Results**

MS32 MVR-PCR analysis is directed from a unique sequence primer (32-O, 32-D or 32-B) located in the 5' flanking sequence of MS32 into the minisatellite array (Figure 5.2). The original |l clone containing MS32 includes only a further 425bp of DNA 5' to the first minisatellite repeat unit. This region was previously sequenced in the human clone (Wong et al., 1987) and partially sequenced in a selection of primates (Gray and Jeffreys, 1991). To search for polymorphisms in this region in humans, primer 32-OR was designed and used in conjunction with 32-B (Figures 5.2 and 5.3) to amplify the 348bp of DNA immediately flanking the most variable and unstable end of the minisatellite which is analysed in MVR-PCR.

**Identification of three common polymorphisms in the flanking DNA.** PCR amplification followed by restriction digest analysis of this region from 12 unrelated Caucasian individuals revealed a *HinII* restriction site dimorphism in this region (designated as Hf* for presence of the *HinII* restriction site and Hf for absence of the *HinII* restriction site). This region showed no polymorphisms using the restriction enzymes *BglII, Ddel, Fnu4HI* or *AluI* in the same 12 unrelated individuals. Direct DNA sequencing of this region from PCR amplification products from a heterozygous individual (Hf*/Hf) and a single molecule separated Hf allele from a second heterozygous individual (Monckton and Jeffreys, 1991b; Chapter 3) revealed the polymorphism as a C (presence of *HinII* restriction site, Hf*) to T transition (absence of *HinII* restriction site, Hf) at position 143 (Figure 5.4, Table 5.1).

PCR-SSCP analysis (Orita et al., 1989) of the entire flanking region (32-OR to 32-B) in 8 CEPH parents homozygous for Hf* revealed another common polymorphism (see segregation analysis of family 1416, Chapter 5 Page 3
Figure 5.2. Diagrammatic representation of the MS32 5'-flanking region, showing polymorphic sites and PCR primers.

Filled circles represent polymorphic base substitutions, open squares non-polymorphic restriction sites and filled squares polymorphic restriction sites. Arrows indicate PCR primers. PCR primer sequences are: 32-OR 5'-TCACGGTTGAAACCACCCACCCACACCAATCTT-3', 32-H2AR 5'-GTGCACTTCCCACCCCTCCAGCC-3', 32-H2C 5'-TGATGCTCGTTCCCGGATAC-3', 32-D2 5'-CGACTCGAAGATGAGCGAGATG-3', 32-D 5'-CGACTCGAGATGAGCGAGCAATG-3', 32-H1C 5'-TGATGCTCGAAGAGAGAAATAC-3', 32-H1B 5'-TGGTGCTGCAAGAAGAG-3', 32-NR 5'-AGTAGCCAAATCCAGAATAGC-3' and 32-B 5'-TAAGCTCTCCATTCCAGTCTTCTG-3'.
Figure 5.3. PCR assays for the three polymorphic sites identified in the flanking region of MS32.

(A) Segregation analysis of the Hump1 polymorphism for CEPH family 1416. SSCP analysis top and PCR assay bottom (genotypes for each individual are shown, GG/GC/CC). PCR-SSCP analysis of the flanking DNA amplified with 32-OR and 32-B was performed using the method of Orita et al. (1989), incorporating α-32P-dCTP during PCR, followed by digestion with *Hinf*1 and electrophoresis through a 5% polyacrylamide, 10% glycerol gel in 1x TBE at 4°C. For the direct Hump1 PCR assay 0.1μl of 32-OR - 32-B PCR product was reamplified using the nested primers 32-H1B and 32-NR for 28 cycles with an annealing temperature of 55°C and an extension time of 2 minutes. 5μl of this amplification was digested with *Bsp*1286I and resolved by gel electrophoresis through a 3% NuSieve GTG, 1% Sigma TypeI agarose gel in 1x TBE and the products visualised by ethidium bromide staining. Zero DNA controls (0) and φX174 HaeIII size markers (Φ) are also shown.

(B) Segregation analysis of the Hf polymorphism for CEPH family 1331. 348bp of immediate flanking DNA was amplified using primer pair 32-OR plus 32-B for 30 cycles with an annealing temperature of 69°C and a 2 minute extension. 2μl of PCR products were digested with *Hinf*1 and resolved by electrophoresis as above. All individuals produce a constant 163bp product. Individuals homozygous for the Hf allele (-- ) produce a product of 199bp. In individuals homozygous for the Hf+ allele (++) the 199bp band is further digested to give bands of 141 and 58bp. Heterozygous individuals (Hf+/Hf-) produce all four bands (+-).

(C) Segregation analysis of the Hump2 polymorphism for CEPH family 1421 and four unrelated individuals (1-4). Genotypes for each individual are shown (CC/CT/TT). Hump2 analysis was achieved using primers 32-OR, 32-H2C, 32-H2AR and 32-B at final concentrations of 0.5, 0.5, 2 and 1μM respectively in a single tube assay. PCR was performed with an annealing temperature of 67°C, an extension time of 2 minutes for 30 cycles and the products resolved by agarose gel electrophoresis as above.
Figure 5.3A). Direct DNA sequencing of the PCR products amplified from individuals homozygous for the two forms and their heterozygous father showed the polymorphism to be due to a C to G transversion at position 80, designated Humpl (HUMan Primate variant 1, alleles H1C and H1G) (Figure 5.4, Table 5.1). Further sequence comparisons between the sequences obtained here and those obtained previously revealed another polymorphic site within this flanking region, a C to T transition at position 241, designated Hump2 (HUMan Primate variant 2, alleles H2C and H2T) (Figure 5.4, Table 1).

Primate sequence comparisons. Direct comparisons of the flanking region between the cloned human sequence (Wong et al., 1987) and those previously obtained for Chimp, Gorilla and Orang-utans (Gray and Jeffreys, 1991) allowed the derivation of a great ape/human ancestral sequence for this region (using Orang-utans as the outgroup) (Figure 5.4). Nine sites of sequence divergence exist between man and the primate ancestral sequence and all three of the described polymorphic sites so far identified are contained within this group (Table 5.1). I reasoned that the observed differences between the cloned human sequence and the derived ancestral sequence were likely to be due to mutation events that occurred subsequent to the human - great ape split, approximately 6-8 million years ago (Koop et al., 1986). I further reasoned, assuming a fixation time of around 1 million years and a random timing for the generation of new alleles within that 6-8 million years, that approximately 1/7 (ie 1-2) of the sites would have arisen in the past 1 million years, and would thus be unlikely to have progressed to fixation and could still be polymorphic within the present human population. This type of analysis not only produces an estimate for the number of likely polymorphic sites but also direct information as to their probable location. Most significantly it allows the prediction of easily assayable restriction enzyme sites that differ between the human clone and the primate consensus. Obviously the success of this approach is highly dependent on the initial human sequence obtained, since if the chromosome from which the human sequence was gained carries the ancestral allele at a genuinely polymorphic site then such a site will not be identified by this type of analysis. Of the nine sites of sequence divergence identified, six produced changes in commonly available restriction enzyme sites (BglII, BspMl, Hinfl and Xbal, see Table 5.1) and all were assayed in 20 unrelated individuals amplified with primer pair 32-OR and 32-B. Other than the previously identified Hinfl polymorphism none of the six sites examined were found to be commonly polymorphic. The base substitutions at sites 80, 94 and 241 do not affect recognition sequences for any commonly available restriction enzymes. Sites 80 and 241 were previously shown to be polymorphic by SSCP and sequence analysis (Humpl and Hump2), whilst sequence analysis of seven amplified human alleles and the human clone has not revealed the persistence of the ancestral allele at position 94.

Assays for the polymorphisms and heterozygosity analysis. As a simple restriction site dimorphism the Hf polymorphism was very easily typed by standard PCR amplification (using primer pair 32-OR and 32-B) and subsequent Hinfl digestion (Figure 5.3B). Typing of this polymorphism across the 80 parents in the CEPH panel of families and across 101 unrelated Japanese individuals showed a heterozygosity level of 31% for this polymorphism in both populations (Table 5.2).

Unfortunately neither Humpl nor Hump2 created or destroyed restriction enzyme sites within the flanking region and thus an alternative approach to determining allele status at these polymorphic sites was required. For Hump2 a single tube four primer PCR assay was developed (Figure 5.3C). Two opposing primers specific for the two alternative alleles were created, 32-H2C (for amplification from the H2C allele) and 32-H2AR (for amplification from the H2T allele), and used in conjunction with the universal primers 32-OR and 32-B (see Figure 5.2). An individual homozygous for the H2T allele produces a 259bp band corresponding to the
<table>
<thead>
<tr>
<th>Human Clone</th>
<th>Ancestral Sequence</th>
<th>Human Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAAGCTCTCCATTTTCCAGTTTTCTGAAAAATTTGTGTAGAATTTTTTGTGTAATAATTTTTTGTGCTTGAAGAAAGATA</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>CACTCAACATAAGTTTAATTTTCTCAGCAAGGCAATTTTTCTTCCTCTTAGAAGGTGCAGCTCAGATGAGACAGCATG</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>CAGAGCAGACCTGAACAGGAGGAAAGGTTCTGTCTCTGATACTGAGCCCTCTACTGATGCTGTTCCGTAT</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>CGGTAGGGTTGAGCTGCAGTCATAGCTTAAAATCGAGTTGCTACTTTAAAAGAGACGAGGTTGATGACCAAGGGACGG</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>GGGTAGAGTGGTTGGGAGGAGGTGTT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 5.4. Primate sequence comparisons for the MS32 5' flanking region.**

The sequence of the human clone between primer pair 32-OR and 32-B is given in full (Wong et al., 1987). The human-African Ape ancestral sequence was derived from the human, Chimp, Gorilla and Orang-utan sequences (Gray and Jeffreys, 1991), using Orang-utans as the outgroup. Human sequence variants were determined in this study by direct sequencing of PCR products; single stranded template DNA was generated by asymmetric PCR (Gyllensten and Erlich, 1988) and sequenced in the presence of the detergent NP-40 by the dI-deoxy chain termination method as previously described (Bachman et al., 1990) using T7 polymerase (Pharmacia). Positions of variation only are indicated in bold. N's in the ancestral sequence represents sequence not known.
Table 5.1. Human/primate ancestral sequence variant sites in the MS32 flanking region.

<table>
<thead>
<tr>
<th>No</th>
<th>Position</th>
<th>Human Clone</th>
<th>Human/Ancestral Sequence</th>
<th>Human Clone/Ancestral Restriction Site Differences</th>
<th>Polymorphism in Caucasians</th>
<th>Polymorphic Locus Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80</td>
<td>C</td>
<td>G</td>
<td>none</td>
<td>+</td>
<td>Hump1</td>
</tr>
<tr>
<td>2</td>
<td>94</td>
<td>G</td>
<td>A</td>
<td>none</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>127</td>
<td>C</td>
<td>A</td>
<td>XbaI +/-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>143</td>
<td>C</td>
<td>T</td>
<td>HinI +/-</td>
<td>+</td>
<td>Hf</td>
</tr>
<tr>
<td>5</td>
<td>197</td>
<td>G</td>
<td>T</td>
<td>HinI +/-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>207</td>
<td>A</td>
<td>G</td>
<td>BspMI +/ -</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>241</td>
<td>C</td>
<td>T</td>
<td>none</td>
<td>+</td>
<td>Hump2</td>
</tr>
<tr>
<td>8</td>
<td>309</td>
<td>G</td>
<td>A</td>
<td>BglII +/-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>319</td>
<td>C</td>
<td>T</td>
<td>BglII +/-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

*Only 7 chromosomes have been analysed for this locus.

Table 5.2. MS32 flanking polymorphism allele frequencies.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Allele</th>
<th>Caucasian Frequency</th>
<th>Caucasian Number</th>
<th>Japanese* Frequency</th>
<th>Japanese* Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hump 1</td>
<td>G</td>
<td>0.69</td>
<td>55</td>
<td>0.80</td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.31</td>
<td>25</td>
<td>0.20</td>
<td>41</td>
</tr>
<tr>
<td>Hf *</td>
<td>+</td>
<td>0.81</td>
<td>129</td>
<td>0.81</td>
<td>163</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>0.19</td>
<td>31</td>
<td>0.19</td>
<td>39</td>
</tr>
<tr>
<td>Hump 2</td>
<td>C</td>
<td>0.59</td>
<td>94</td>
<td>0.91</td>
<td>184</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0.41</td>
<td>66</td>
<td>0.09</td>
<td>18</td>
</tr>
</tbody>
</table>

* Some CEPH samples were Hf typed by A. MacLeod. All Japanese samples were typed by K. Tamaki.
PCR product from primer pair 32-H2AR and 32-B, as well as the 364bp internal control band derived from the universal primers 32-OR and 32-B. In contrast an individual homozygous for the H2C allele produces a 142bp band corresponding to the PCR product from primer pair 32-H2C and 32-OR, as well as the 364bp internal control band. Heterozygous individuals (H2C/H2T) produce all three bands. Typing of the Hump2 polymorphism across the 80 parents in the CEPH panel of families and across 101 unrelated Japanese individuals showed heterozygosity levels of 48% and 16% respectively (Table 5.2).

Unfortunately the Hump1 polymorphic site lies in a very A/T rich region of DNA (26% G/C in the 50bp surrounding Hump1) and an alternative strategy was required to assay this site. The mismatched primer 32-HIB primes just 5' to the Hump1 polymorphism and forces the incorporation of a 3' terminal G rather than the A present in genomic DNA. Use of this primer during low stringency PCR allows incorporation of this transition into resulting PCR products (see Figure 5.5). This forced insertional mutation creates or destroys an easily assayable Bsp1286I restriction enzyme site dependent on allelic state at the Hump1 locus (H1G derived products amplified with 32-HIB are cut by Bsp1286I). Unfortunately the low annealing temperature required to ensure the A/T rich 32-HIB primer incorporating the terminal mismatch primes efficiently, prevented the direct use of total genomic DNA as a PCR template. Thus a preliminary amplification of the entire flanking region with primer pair 32-OR plus 32-B (as used in the Hf assay) was required to generate seed DNA for use in a nested 32-NR (32-NR primes just 5' of 32-OR and acts as a nested primer directed into the 5' flanking DNA of MS32) to 32-HIB amplification (Figure 5.2). Simple genotyping of this polymorphism was then achieved by Bsp1286I digestion and agarose gel electrophoresis (Figure 5.3A). Typing of the Hump1 polymorphism across 80 parents in the CEPH panel of families and across 101 unrelated Japanese individuals showed heterozygosity levels of 43% and 32% respectively (Table 5.2).

Knockout MVR-PCR. These flanking polymorphisms can be used to map individual MS32 alleles from total genomic diploid DNA by the use of allele specific primers located in the flanking DNA. PCR primer 32-D2 spans the site of the Hf polymorphism and was used as an allele specific MVR-PCR primer (see Figure 5.5). Using 32-D2 as the fixed primer in the flanking DNA it was possible to amplify only MS32 alleles linked to the Hf* site, ie to 'knockout' the amplification of the Hf- linked allele. For heterozygous individuals (Hf*/Hf-) it was possible to obtain the allele map from the Hf+ allele directly from total genomic DNA (using primer 32-D2) and for the Hf- linked allele by subtraction of the Hf+ allele from the diploid code derived from a standard MS32 MVR-PCR using a universal flanking primer (32-D, 32-O or 32-B) (Figure 5.6).

PCR primer 32-H2C can also be used as an allele specific MVR-PCR primer; using this as the fixed primer in the flanking DNA it is possible to knockout H2T linked alleles and amplify only MS32 alleles linked to H2C. As with the Hf polymorphism in heterozygous individuals (H2C/H2T) it is possible to obtain the allele map from the H2C linked allele directly from total genomic DNA (using primer 32-H2C) and for the H2T linked allele by subtraction of the H2C allele from a standard MS32 MVR-PCR (Figure 5.6).

Similarly the Hump1 specific primer 32-H1C can also be used for knockout MVR-PCR in heterozygous individuals (Figure 5.6). Using these three polymorphisms for allele knockout, over 200 new alleles have been mapped from unrelated heterozygous individuals (Chapter 6).

Haplotype analysis of flanking DNA polymorphisms. Haplotypic analysis of the polymorphisms to each other and to the minisatellite alleles may be achieved in a variety of ways. Pedigree analysis is the simplest, and has been applied to the three flanking polymorphisms and the minisatellite array for 40 CEPH families. Haplotypes of each flanking polymorphism with respect to the minisatellite array can be
Figure 5.5. Sequence of the Hump1 and Hf flanking polymorphisms and associated PCR primers.

(A) The Hump1 polymorphism. The genomic sequence is shown, with the position of the polymorphic G/C transversion, the allele specific primer 32-H1C and the PCR assay primer 32-H1B. Note the 3' terminal mismatched G incorporated into 32-H1B.

(B) The sequence of the PCR products derived from amplification with 32-H1B. Note the forced incorporation of the G base into the product (bold), which allows the two Hump1 alleles (also bold) to be assayed by Bsp1286I digestion.

(C) The Hf polymorphism. The genomic sequence is shown, with the position of the polymorphic C/T transition and the position of the two primers 32-D and 32-D2. At high annealing temperature (>65°C) 32-D2 acts as an allele specific primer.
Figure 5.6. Knockout MVR-PCR.

For each of the flanking polymorphisms (Hump1, Hf and Hump2) three unrelated individuals were chosen who were heterozygous for the polymorphism. Each individual was analysed by MVR-PCR using either the universal flanking primer 32-O (O) to generate the diploid code from both alleles or the allele specific flanking primer (32-H1C, 32-D2 or 32-H2C) to generate coding from a single allele. MVR-PCR products extending to a-type repeats (A) or t-type repeats (T) were resolved by agarose gel electrophoresis and Southern blot hybridisation using 32P-labelled MS32 as probe. The 10th repeat unit on the MVR-PCR ladder is arrowed to show registration of single allele and diploid codes. MVR-PCR with the fixed flanking primers 32-O, 32-H2C or 32-D2 was carried out using an annealing temperature of 69°C and an extension time of 5 minutes for 18 cycles, with all other procedures as previously described (Jeffreys et al., 1991b). Knockout MVR-PCR using the flanking primer 32-H1C was performed with an annealing temperature of 64°C for five cycles and 60°C for 13 cycles, again with all other procedures as previously described.
directly achieved by knockout MVR-PCR. PCR assays for the direct haplotyping of the flanking polymorphisms to each other are currently being developed. Results for the haplotyping of the Humpl, Hf and Hump2 polymorphisms in the parents of the 40 CEPH families are presented in Table 5.3. Significant linkage disequilibrium exists between all the polymorphic sites, but in only case is the observed disequilibrium absolute (although no Caucasian examples have been observed, one Hf'C - Hf' allele has been identified in a Japanese individual). Approximately 70% of Caucasian individuals are heterozygous for at least one of the variant Humpl, Hf and/or Hump2 sites and may have single alleles mapped by knockout MVR-PCR.

**Applications to mixed DNA samples.** As previously described MS32 MVR-PCR is likely to have major applications in forensic science (Jeffreys et al., 1991b), an application for which mixed DNA samples are often encountered, eg mixed victim and assailant’s blood in violent attacks, vaginal swabs in rape cases, mixed semen samples in multiple rape cases and mixed partner/rapist semen samples. Previously it has been shown that ambiguous diploid codes may be derived from mixtures of DNA down to approximately 10% admixture, and that particularly in cases where a pure sample of one of the DNAs, eg victim, is available, a high level of exclusionary power is achieved (on average 99.9993% of false suspects excluded using information contained in the first 50 repeats, Jeffreys et al., 1991). Even in cases where neither DNA from a mixed sample is available in a pure form, valuable information to exclude false suspects may still be derived. However, mixtures of DNA below 10% and mixtures of two or more DNAs are less amenable to standard MVR-PCR analysis. Depending on the genotypes of the flanking polymorphisms in the two DNAs in a mixed sample it may prove possible to use a flanking primer specific to only one individual to generate individual specific MVR code. To investigate the potential forensic applications of knockout MVR-PCR two, of many possible, mixed DNA scenarios were simulated: 1, the mixture of a Hf'C homozygous assailant with a Hf'T homozygous victim, allowing use of the Hf'C specific primer 32-Hf'C to selectively amplify only the assailant’s alleles, thus deriving the assailant’s diploid code; and 2, the mixture of a Hf+/Hf' heterozygous assailant with a homozygous Hf' victim allowing use of the Hf' specific primer 32-D2 to specifically amplify only one of the assailant’s alleles. Two suitable individuals were identified, ie a Hf'T/Hf'T, Hf'/Hf ' victim’ and a Hf'C/Hf'C, Hf'+/Hf' 'assailant' and DNA mixtures from 1:1 to 1:200 (victim : assailant) made and MVR-PCR analysis performed with the appropriate primer combinations (see Figure 5.7). Allele specific primer 32-Hf'C can be used to unambiguously amplify only the assailant’s alleles down to mixtures of at least 1:10 (150ng of victim DNA : 15ng of assailant DNA). For the 1:50 mixture only the assailants diploid code was seen, but some variation in band intensity was observed as the lower limit for the quantity of input DNA was approached (only 3ng of specific input DNA, below 10ng of input DNA MVR priming becomes stochastically variable with a resulting variation in band intensity, Jeffreys et al., 1991b). Below 1:50 mixtures, extra cycles of PCR were required to produce detectable levels of product, with resulting increased background signal derived from mispriming from the victim’s DNA (ie the aberrant low efficiency amplification of the victims DNA with the incorrect allele specific flanking primer); as a consequence unambiguous information was no longer derived. Nevertheless it may be possible to derive an ambiguous code at mixtures far lower than possible using standard MVR-PCR, especially if enough material is available to permit multiple amplifications allowing derivation of a consensus code if stochastic loss of PCR products is observed for very small starting amounts of DNA. The Hf' specific primer (32-D2) shows less allele specificity than the Hump2 allele specific primer, but it does allow the derivation of single allele codes for mixtures down to 1:2. Primer 32-D2 was not initially designed as an allele specific primer but fortuitously spanned the Hf polymorphism (Figure 5.5). An alternative primer designed specifically to access the
Table 5.3. Caucasian haplotype frequencies for the Hump1, Hf and Hump2 polymorphisms.

<table>
<thead>
<tr>
<th>Hump1 - Hf - Hump2 Haplotype</th>
<th>Expected number</th>
<th>Observed number*</th>
<th>Observed frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>C - T</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C - C</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G - T</td>
<td>9</td>
<td>23</td>
<td>0.14</td>
</tr>
<tr>
<td>G - C</td>
<td>12</td>
<td>8</td>
<td>0.05</td>
</tr>
<tr>
<td>C + T</td>
<td>16</td>
<td>30</td>
<td>0.19</td>
</tr>
<tr>
<td>C + C</td>
<td>24</td>
<td>12</td>
<td>0.08</td>
</tr>
<tr>
<td>G + T</td>
<td>37</td>
<td>13</td>
<td>0.08</td>
</tr>
<tr>
<td>G + C</td>
<td>53</td>
<td>74</td>
<td>0.46</td>
</tr>
</tbody>
</table>

* $\chi^2$ (7df) = 75, a significant deviation from a null hypothesis of random association, $p<0.001$.

For Hump1 to Hf: $\chi^2$ (1df) = 14.40, a significant deviation from a null hypothesis of random association, $p<0.001$.

For Hump1 to Hump2: $\chi^2$ (1df) = 21.30, a significant deviation from a null hypothesis of random association, $p<0.001$.

For Hf to Hump2: $\chi^2$ (1df) = 16.46, a significant deviation from a null hypothesis of random association, $p<0.001$. 
Figure 7.7. The application of knockout MVR-PCR to mixed DNA samples.
Legend to Figure 5.7 (page 6b). The application of knockout MVR-PCR to mixed DNA samples.

MVR-PCR of mixtures of two DNAs using allele specific primer 32-D2 or 32-H2C as flanking primers in an otherwise standard MVR-PCR reaction. Individual X ('victim') was a Hf' homozygote (Hf'/Hf') and a H2T homozygote (H2^T/H2^T) and individual Y ('assailant') was a Hf heterozygote (Hf'/Hf') and a H2C homozygote (H2^C/H2^C). Mixtures of DNA from X and Y were prepared using a fixed amount of X (150ng) and decreasing amounts of Y (150ng down to 0.75ng). The most dilute samples of Y (1/100 + 1/200) were given a further 2 cycles of PCR to increase the yield of product to detectable levels.

Figure 5.8. The efficiency of single allele codes in excluding individuals based on comparison with their diploid codes.

Single allele codes extending over at least 50 repeat units were established for 411 different MS32 alleles (349 Caucasian and 62 Japanese). Each allele was then compared with the diploid code of each of 408 unrelated individuals (excluding the individual from whom the allele was derived), giving 167,688 allele/individual comparisons in total. For each comparison, repeat unit positions which excluded the allele as having come from the individual were identified; for example, an allele with a t-type repeat unit at a given position could not have come from an individual homozygous for a-type repeats at that position. The frequency distribution of the total number of exclusions over the first 50 repeat units is given for all allele/individual comparisons. This analysis was performed by A. J. Jeffreys using software written by A.J.J. and MS32 allele and diploid code databases generated by A.J.J., K. Tamaki, A. MacCleod, D. Neil, M. Allen and myself.

Chapter 5 Page 6c
Hf polymorphism should amplify more selectively and allow derivation of unambiguous codes at lower levels than achievable with 32-D2. New Hf⁺ and Hf⁻ specific flanking primers have been designed and are currently being evaluated.

The power of using single allele codes to identify individuals based on comparisons with their diploid MVR-PCR code was also assessed. Each of the 411 different alleles in the current allele database was used to screen the diploid database of 408 unrelated Caucasian and Japanese individual codes. For each comparison the single allele code was checked for positions of exclusion with the diploid code, for example a single allele with an a-type repeat at a given position could not have originated from an individual homozygous for t-type repeats at the same position; the number of exclusions per false suspect is plotted in Figure 5.8. 99.87% of false suspects were excluded using information from the first 50 repeats, with a mean of 10.7 exclusions per case. However, many of the alleles in our database were derived from mother-father-single child trios and thus contain some ambiguous positions (Jeffreys et al., 1991b); this situation does not accurately reflect the circumstances likely to arise in genuine forensic applications where the code of the allele under test will have been generated unambiguously by knockout MVR-PCR. We therefore repeated this analysis using 235 completely mapped alleles and, as expected, the level of exclusion rose slightly to 99.9%, with a mean of 11.3 exclusions per case. The power of exclusion for any one allele though was not uniform with the majority of alleles excluding all false suspects (96.11% and 96.60% respectively for total and unambiguous allele databases), with the major loss in overall exclusionary power being due to a limited subset of alleles with poor discriminatory power. Those alleles which failed to exclude greater than 99% of false suspects were found upon examination to be 'a' rich alleles (ie almost completely comprised of a-type repeats, data not shown). Nonetheless, even the worst unambiguous allele still managed to exclude greater than 95% of false suspects, an exceptionally high level for the worst case scenario of one allele of one locus. In summary more than 98.5% of single alleles exclude more than 99% of false suspects.

Using the Hf and Hump2 haplotype frequencies derived from the analysis of the 40 CEPH families (160 haplotypes) an approximate estimate for the number of mixed DNA samples to which unambiguous diploid or single allele mapping could theoretically be applied using the Hf and Hump2 discriminatory system can be calculated (see Table 5.4; this analysis assumes the mixes are of sufficient quality and in reasonable proportions to allow unambiguous MVR-PCR to be performed). It can be seen that in approximately 25% of cases an unambiguous diploid code would be derivable from a mixed DNA sample, and in up to 50% of cases either diploid code, or single allele, information would be recoverable. Use of the Hump1 polymorphism in this type of analysis should further improve the proportion of mixed DNA scenarios to which MVR-PCR based identification could be applied.

Discussion

Thus far three common polymorphisms have been identified in the immediate 348bp of DNA flanking the minisatellite locus MS32. For each polymorphic site rapid and reliable PCR based tests for allelic state have been developed and allele frequencies in two major populations determined. Each locus appears to be at Hardy-Weinberg equilibrium, whilst significant, but not absolute, linkage disequilibrium exists between sites (a more detailed evaluation of haplotype data in studies of minisatellite mutation processes and recombination is made in Chapter 7). The use of such polymorphic sites to design allele specific primers has been demonstrated as well as their use in single allele or knockout MVR-PCR. With a combined heterozygosity in the flanking DNA of in excess of 70%, the large scale mapping of separate alleles in large numbers of unrelated individuals becomes
Table 5.4. Theoretical estimation of the level of information obtainable from mixed DNA samples using the Hf and Hump2 allele specific primers in MVR-PCR.

<table>
<thead>
<tr>
<th>Assailant's Haplotypes</th>
<th>f (%)</th>
<th>V</th>
<th>+C</th>
<th>+C</th>
<th>+C</th>
<th>+C</th>
<th>-C</th>
<th>-C</th>
<th>+T</th>
<th>+T</th>
<th>+T</th>
<th>-T</th>
<th>-T</th>
<th>-T</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>28.9</td>
<td>5.4</td>
<td>28.9</td>
<td>15.5</td>
<td>0.3</td>
<td>2.7</td>
<td>1.4</td>
<td>7.2</td>
<td>7.7</td>
<td>2.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+C</td>
<td>8.4</td>
<td>1.6</td>
<td>8.4</td>
<td>4.5</td>
<td>0.1</td>
<td>0.8</td>
<td>0.4</td>
<td>2.1</td>
<td>2.2</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+C</td>
<td>5.4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
<td>0.4</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+T</td>
<td>28.9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2*</td>
<td>1</td>
<td>2*</td>
<td>0</td>
<td>1</td>
<td>2*</td>
<td>2*</td>
<td>2*</td>
<td>2*</td>
</tr>
<tr>
<td>+T</td>
<td>8.4</td>
<td>1.6</td>
<td>8.4</td>
<td>4.5</td>
<td>0.1</td>
<td>0.8</td>
<td>0.4</td>
<td>2.1</td>
<td>2.2</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-C</td>
<td>15.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-C</td>
<td>0.3</td>
<td>2*</td>
<td>1</td>
<td>2*</td>
<td>2*</td>
<td>0</td>
<td>1</td>
<td>2*</td>
<td>2*</td>
<td>2*</td>
<td>2*</td>
<td>2*</td>
<td>2*</td>
<td>2*</td>
</tr>
<tr>
<td>-C</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-C</td>
<td>2.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-C</td>
<td>0.8</td>
<td>0.1</td>
<td>0.8</td>
<td>0.4</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-T</td>
<td>1.4</td>
<td>2*</td>
<td>1</td>
<td>2*</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2*</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-T</td>
<td>0.4</td>
<td>0</td>
<td>0.1</td>
<td>0.4</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+T</td>
<td>7.2</td>
<td>2*</td>
<td>2*</td>
<td>1</td>
<td>2*</td>
<td>2*</td>
<td>1</td>
<td>2*</td>
<td>0</td>
<td>1</td>
<td>2*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+T</td>
<td>2.1</td>
<td>0</td>
<td>0.4</td>
<td>2.1</td>
<td>1.1</td>
<td>0</td>
<td>0.2</td>
<td>0.1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+T</td>
<td>7.7</td>
<td>2*</td>
<td>2*</td>
<td>1</td>
<td>1</td>
<td>2*</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-T</td>
<td>2.2</td>
<td>0</td>
<td>0.4</td>
<td>2.2</td>
<td>1.2</td>
<td>0</td>
<td>0.2</td>
<td>0.1</td>
<td>0.6</td>
<td>0.6</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-T</td>
<td>2.1</td>
<td>0</td>
<td>0.6</td>
<td>2*</td>
<td>2*</td>
<td>1</td>
<td>2*</td>
<td>2*</td>
<td>1</td>
<td>2*</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-T</td>
<td>0.6</td>
<td>0</td>
<td>0.1</td>
<td>0.6</td>
<td>0.3</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
<td>0.2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The upper figure is the number of assailant’s alleles for which code could be derived (where information on both alleles is recoverable, * = at least one allele separately recoverable and * = diploid code only recoverable) and the lower figure is an estimate for the percent likelihood of this scenario being encountered. Haplotype frequencies (f) are based on a sample of 160 Caucasian chromosomes (Table 5.3).
feasible, with obvious potential for the generation of large allele databases, allele groupings and possible derivation of allele and human lineages (see Chapter 6). Mapping of more alleles and concurrent haplotyping of the flanking polymorphisms should shed more light on the mutation processes involved in maintaining ultravariability at this locus (see Chapter 7). It will also help to assess the extent to which interallelic exchange is involved in the generation of new alleles, and to determine whether or not a local recombinational hotspot is indeed present at this locus (Jeffreys et al., 1991; see Chapters 6 and 7). The identification of additional polymorphisms in the flanking DNA of MS32 will further increase the proportion of individuals heterozygous for at least one of the flanking sites, increasing both the number of single allele maps directly obtainable and providing more flanking DNA markers for the detailed analysis of the molecular processes operating at this highly mutable locus.

The existence of additional unknown flanking polymorphisms which affect ‘universal’ flanking primers (32-O, 32-D and 32-B) could lead to inadvertent allele knockout during MVR-PCR (as originally found for the flanking primer 32-D2) and the generation of incorrect diploid codes. However, such knockout of an allele will produce an apparently homozygous pattern devoid of heterozygous (a/t) positions; such patterns are easily identified and such apparently homozygous individuals can be retested with other flanking primers to check for true homozygosity (or possibly heterozygosity for a null MS32 allele carrying a deletion of flanking DNA and flanking primer sites, though no such allele has been identified).

A preliminary study of the potential forensic applications of knockout MVR-PCR in analysing mixed DNA samples has also been described, although a more rigorous and extensive study is needed to confirm the full scope of such applications. The optimization of PCR primer allele specificity and the characterisation of additional polymorphisms should increase the proportion of mixed DNA samples to which MVR-PCR can be applied. The application of knockout MVR-PCR to multiple mixed DNAs has not been tested directly, but they too should prove tractable, although the probability of obtaining unambiguous codes will decrease as the number of DNAs involved increases. Knockout MVR-PCR under some circumstances can be used to obtain information for mixtures containing as little as 1% admixture of DNA; this represents a considerable improvement over other techniques such as Southern blot hybridization using single locus hypervariable probes. Mixed DNA samples also occur in analytical contexts other than forensic medicine, eg monitoring of transplant success in bone-marrow transplants, and such situations should also prove amenable to the same techniques.

I have also investigated the potential use of primate consensus sequences to pin-point sites of potential variation in present day human populations. Although unsuccessful in further increasing the number of polymorphic sites found in this investigation, an initial analysis would have identified the three sites now known to be polymorphic in this region. It should be noted that where primate sequence information already exists it may be used to more rapidly target potentially polymorphic sites in humans.
Summary
Restriction mapping of small PCR amplifiable MS32 alleles has previously shown that the dispersion patterns of the two MVRs are highly diverse, allowing large numbers of alleles to be identified and distinguished. Although most of a small collection of alleles mapped in this way were unique, many shared clear zones of haplotypic identity and could be arranged into groups of aligned alleles. These data showed clear evidence for polarity in allelic variation, with a limited number of 3' haplotypes and extreme divergence in the 5' termini. Recently a rapid PCR based method for assaying the same site of internal variation has been developed (MVR-PCR), with mapping directed in from the variable 5' end and which can be applied to alleles of any size. This method has allowed us to map large numbers of alleles in both pedigrees and unrelated individuals. In addition, we have also been able to link MS32 flanking polymorphisms to the minisatellite internal haplotypes and, through computer analysis, align large number of alleles into related groups. The use of these data in studies of individual identification, minisatellite evolution and population analysis are discussed.

Introduction
Allelic variation at MS32. Haelll restriction based mapping of small MS32 alleles has revealed extreme levels of internal variation in the dispersion patterns of a- and t-type repeats, distinguishing many length isoalleles (Jeffreys et al., 1990; Monckton and Jeffreys, 1991a; Chapter 3). The method is though technically limited, but nonetheless, reasonable numbers of small alleles have been mapped by this process (Jeffreys et al., 1990; and unpublished) and very valuable information derived from them.

Although the majority of alleles are unique, many, of diverse length, share zones of haplotypic identity (ie regions of MVR code that are alignable and presumed to derive from a common ancestral allele) and can be arranged into groups of aligned alleles, which are assumed to share a recent common ancestor (Jeffreys et al., 1990). Such alignments have shown clear evidence for polarity in MS32 internal variation with the major haplotypic zones of identity localised to the 3' ends of alleles. Interallelic variation is largely confined to the extreme 5' end of alleles and is assumed to arise through the action of a local mutational hotspot (Jeffreys et al., 1990).

MVR-PCR at MS32. As described in Chapter 5 MVR-PCR has recently been developed to assay the same sites of internal variation as assayed in MS32 restriction based mapping (Jeffreys et al., 1991b; see Chapter 5). Mapping is applicable to alleles of any size and is directed in from the previously identified ultravari able 5' end (Jeffreys et al., 1991b). Single allele MVR haplotypes can be derived from family groups by pedigree analysis and from unrelated individuals using physically size separated alleles, or by knockout MVR-PCR (Jeffreys et al., 1991b; Chapter 5). The three polymorphisms identified in the 350bp of DNA flanking the 5' end of the MS32 minisatellite have a combined heterozygosity in excess of 70%, thus allowing the
widespread application of knockout MVR-PCR to produce single allele codes from unrelated individuals (Chapter 5).

This work. The extreme levels of variation observed for MVR-PCR diploid codes is based directly on the variation in individual alleles. This chapter describes the application of MVR-PCR to obtaining large numbers of individual single allele MVR maps and their associated flanking haplotypes from both family groups and unrelated individuals. The applications of such data to the study of individual identification, minisatellite evolution and population analysis are discussed. This work forms a central theme of the overall laboratory research project and was conducted by a number of individuals: A.J. Jeffreys, A. MacLeod, D.L. Neil, K. Tamaki, M. Allen and myself. Part of this work has been published (Jeffreys et al., 1991b; Tamaki et al., 1992).

Results and Discussion

Pedigree analysis. Diploid MS32 MVR-PCR codes, extending for at least 50 repeats, were derived from the parents and offspring of all 40 families in the CEPH panel. Diploid codes from both parents and two children of differing genotypes sharing one allele in common were used to unambiguously deduce all four parental MVR haplotypes (see Figure 6.1). The sibling codes obtained in large families were used to confirm the authenticity of all codes typed. This procedure was repeated for 7 large Bangladeshi families, 8 small Fanconi anaemia families of Germanic origin and 60 British single child-mother-father trios. For single child family trios, and some small families, unambiguous MVR-haplotypes were derived for all four parental alleles, except for those positions where all individuals were heterozygous a/t; in such cases the relevant map positions could not be deduced and were scored as unknown (?). PCR based typing of the three flanking polymorphisms in all the CEPH parents and at least two children of differing genotype allowed haplotyping of the Humpl, Hf and Hump2 polymorphisms to the minisatellite MVR map (see Table 5.3 for flanking haplotype summary).

Unrelated individuals. Single allele MVR maps were derived from some Caucasian individuals by physical size selection of MboI digested genomic DNA. Allele knockout was used on many remaining unrelated individuals. Large numbers of unrelated Japanese individuals, plus some unrelated Melanesians and Papua New Guineans were typed for the Hf flanking polymorphism. Heterozygous individuals had single allele maps derived by knockout MVR-PCR (see Chapter 5). Each individual was typed with the Hf" specific flanking primer (32-D2), to generate a single allele code and a universal flanking primer (32-D or 32-0) to generate a standard diploid code. The single allele code of the unknown (Hf" linked) allele was deduced by subtraction of the known (Hf" linked) allele from the diploid code. Some Japanese individuals were also haplotyped for the other two flanking polymorphisms (Humpl and Hump2).

Allelic diversity. With three major coding states (a-, t- and O-type) single allele mapping over the first 50 repeats of MS32 has the theoretical capability to distinguish $7 \times 10^{23}$ ($3^{50}$) different alleles. From both pedigree analysis and unrelated individuals we have coded a total of 608 alleles, of these 591 were different from each other and 570 were unique, being observed only once. For Caucasians 395 alleles have been mapped, 383 of which are different and 372 unique. One allele has been sampled 3 times and 10 have been sampled twice, giving a very low maximum allele frequency of $3/395 = 0.0076$. Japanese samples show a similar level of variation, with 148 alleles mapped, 144 different, 140 unique and a maximum allele frequency of $2/1148 = 0.0135$. Assuming all alleles are equally rare a Poisson analysis indicates the existence of a minimum of 6800 Caucasian and 2600 Japanese alleles (A.J. Jeffreys, personal communication). However, using the world population size ($5 \times 10^9$) and the known mutation rate (~1%, see Chapter 7) we can predict the existence of $>10^8$ different MS32
<table>
<thead>
<tr>
<th>Individual</th>
<th>Alleles</th>
<th>Diploid code</th>
</tr>
</thead>
<tbody>
<tr>
<td>father 10401</td>
<td>A B</td>
<td>33331 33131 33123 33111 11332 11333 31113 23131 22331 33133..</td>
</tr>
<tr>
<td>mother 10402</td>
<td>C D</td>
<td>11233 23251 14331 15112 35555 55454 55555 55666 66666 66666..</td>
</tr>
<tr>
<td>children:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10406,07,11,12</td>
<td>A C</td>
<td>11313 31331 31321 13113 34445 44544 44444 54454 55554 44454..</td>
</tr>
<tr>
<td>10408</td>
<td>A D</td>
<td>11331 33341 34131 14113 13332 33331 33333 23454 55554 44454..</td>
</tr>
<tr>
<td>10404,05</td>
<td>B C</td>
<td>33223 23232 13323 32113 34555 54445 55445 55444 54444 54544..</td>
</tr>
<tr>
<td>10409,10</td>
<td>B D</td>
<td>33221 23251 15133 35113 31232 33123 23332 22444 55444 54544..</td>
</tr>
</tbody>
</table>

Derived single allele maps

<table>
<thead>
<tr>
<th>Allele</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>?aaaaa aaaaa taata aaaaa aaat aataa aaaaa taata tttta aataa..</td>
</tr>
<tr>
<td>B</td>
<td>?tttt tttta ttata atatt tttaa aattt aatatt taaat tttaaa tttaa ttaat..</td>
</tr>
<tr>
<td>C</td>
<td>?aatat ttata aattta ataat t0000 00000 00000 00000 00000 00000..</td>
</tr>
<tr>
<td>D</td>
<td>?aatta ttt0a a0aaa a0aat atttt ttata ttttt tt000 00000 000000..</td>
</tr>
</tbody>
</table>

Figure 6.1. MVR-PCR pedigree analysis.

Complete reconstruction of allele maps using the large sibship of CEPH family 104. The children show four different diploid codes, corresponding to the four possible combinations of parental alleles. The mother unusually contains two short alleles, resulting in coding state 6(00) beyond the end of the longer allele. The resulting haplotypes (alleles C and D) terminate in a string of "null" (nonexistent) repeats. Haplotypes were extracted from the diploid codes of the father, mother and each child using software written in VAX BASIC V3.4. The alleles are defined: allele A, the paternal allele inherited by child one; allele B, the paternal allele not inherited by child one; allele C the maternal allele inherited by child one; and allele D, the maternal allele not inherited by child one. For a family with a single child, the four parental haplotypes were extracted sequentially along each position of the diploid code. For each position, the code of the father, mother and child were noted and checked in a look-up table to determine whether exclusions exist and if not, to determine the repeat unit type transmitted or not from each parent to the child. For example, if the father is 1(aa), mother 3(aa) and child 3(at), then no exclusions exist and the repeat units at that position on each allele is given by allele A, a; allele B, a; allele C, t; allele D, a. In contrast, codes 1(aa) + 2(tt) + 5(00) would give a paternal mutation/exclusion. For families with more than one offspring, the incomplete haplotype of each parental allele was determined separately for each child as described above for the single child family. Incomplete haplotypes deduced from all children were then compared to identify matching alleles and to deduce which parental allele had been transmitted to each child. The consensus haplotype of each allele was then determined from the incomplete haplotypes deduced from each child, thereby removing all uncertain positions. Finally, the diploid code of each individual was compared with the code predicted from the two constituent alleles, as a final check to ensure full concordance of all diploid codes and haplotypes.

This figure was adapted from Jeffreys et al. (1991b).
alleles world-wide. For the other populations investigated the sample sizes were too small to make reliable estimates, although preliminary data suggests they too have high allelic variability.

**Diploid code diversity.** The large numbers of different alleles present in the population means that the number of potential diploid codes is extremely large. Assuming random association of different alleles we can predict the existence of a minimum of $23 \times 10^6$ possible combinations of allele pairs in the Caucasian population, the majority of which will produce unique diploid codes. This huge diploid variability is reflected in current data; out of 500 unrelated individuals digitally typed no two individuals share the same diploid code. In fact, with the exception of two individuals, all individuals can be distinguished on the basis of their code over the first 19 repeats. The two individuals for whom diploid code identity extends the furthest into MS32 (25 repeats, see Figure 6.2A) have atypical diploid codes, with both individuals containing a short allele (indicated by the run of hemizygous positions (code states 4 and 5) beyond code position 21). For one individual the two alleles have both been mapped by knockout MVR-PCR. The short allele is from a group of short alleles relatively common in Japanese populations (see below). This allele is almost completely homogeneous for $a$-type repeats and is probably shared by both individuals (consistent with the diploid codes seen). In both cases the longer allele is also relatively homogeneous, comprising largely of $a$-type repeats. The two individuals differing at the least number of positions (5) in the first 50 repeats, both also contain $a$-rich, relatively homogeneous alleles (Figure 6.2B). Homogeneous $a$-rich MS32 alleles, particularly short ones, probably represent the greatest limitation to diploid code diversity and it seems probable that such alleles provide the greatest potential for identifying two unrelated individuals with shared diploid codes. It should be noted that the two pairs of individuals identified here still differed at at least 5 code positions, each one representing an absolute exclusion.

**MS32 MVR-PCR heterozygosity.** Diploid coding presents a highly objective method for distinguishing homozygotes, compared to standard Southern blot analysis (Devlin et al., 1991; Chakraborty and Jin, 1992; see Chapter 3, page 2); genuine MS32 homozygotes will have diploid codes restricted to homozygous codes 1 (a/a), 2 (t/t) and 6 (O/O). So far, only four such MS32 MVR-PCR homozygotes have been identified out of 500 unrelated individuals typed, giving an observed heterozygosity of 99.2%. These individuals are unlikely to be heterozygous for a polymorphism in the flanking DNA knocking out the flanking primer site since all showed the expected single band by conventional Southern blot analysis (data not shown). In contrast, the majority (8/10) of apparent MS32 homozygotes previously identified by Southern blot analysis have been demonstrated to be heterozygous for similar or identical length alleles by diploid coding. From the massive allelic variability observed, it would seem likely that the majority of MS32 homozygotes represent recent consanguinity; from the number of different alleles observed, the true heterozygosity in outbred individuals should be in excess of 99.9% (A.J. Jeffreys, personal communication).

MVR-PCR diploid coding does present the potential for incorrect assignment of homozygosity under two scenarios: firstly, heterozygosity for a polymorphism in the flanking DNA knocking out the site for the flanking primer could produce an apparently homozygous diploid code (but really hemizygous); or secondly an individual containing two alleles of differing length but sharing exactly the same 5' haplotype would also appear homozygous. Polymorphisms in the flanking DNA knocking out the flanking primer site are likely to be the most common cause of pseudohomozygosity, but can easily be investigated by repeating the coding with an alternative flanking primer which should produce the same homozygous code in true homozygotes; this is now our recommended procedure for all apparent homozygotes encountered. The existence of an individual heterozygous for a large deletion removing all the flanking primer sites cannot be discounted, although no such
Figure 6.2. The most similar MS32 diploid codes.

(A) The two most similar diploid codes reading in from the 5' end. The diploid codes for two unrelated Japanese individuals (J1 and J2) are shown with positions of difference indicated (*). For individual J2, single allele maps are also shown.

(B) The two most similar diploid codes extending over the first 50 repeats. The diploid codes for two unrelated Caucasian individuals (C1 and C2) are shown with positions of variation indicated (*). For individual C2, single allele maps are also shown. Note that the two alleles are very similar.

individual has yet been identified. Individuals homozygous for two different length alleles, but with identical 5' haplotypes, have not yet been encountered and it is probable they will be very rare. Although many alleles share large zones of identity, only one point of difference between them is needed to demonstrate heterozygosity. This effect is exaggerated in diploid mapping with coding from a fixed flanking primer directly into the most variable end of the minisatellite array. Even alleles differing only by a single repeat unit insertion/deletion will be thrown out of alignment and produce many heterozygous positions (eg individual C2, Figure 6.2B).

Allele alignments-structural analysis. Although the vast majority of alleles are different, many show clear zones of relatedness (Jeffreys et al., 1990). Using both computer based staggered alignment and dot matrix matching procedures we were able to align ~69% of alleles into 61 different groups containing from 2 to 63 alleles (for details and examples see Figure 6.3). The remaining 31% of alleles failed to align obviously with any other alleles in our current database using the methods outlined (see legend to Figure 6.3). A significant proportion of alleles fall into one of two large groups (eg Group A, Figure 6.3), each containing around 10% of all alleles mapped and suggesting the existence of two relatively ancient haplotypes. As previously noted (Jeffreys et al., 1990), regions of alignment show polarity, with variation largely localised to the extreme 5' end of alleles. This is most clearly demonstrated for groups of small alleles mapped in their entirety, where variation is almost exclusively restricted to the 5' terminus (eg Groups F, G, H, I and J, Figure 6.3). Nonetheless, this terminal variability is observed for all allele groups, including those containing large incompletely mapped alleles, providing further support for the existence of a mutational hotspot operating at the 5' end of MS32 (Jeffreys et al., 1990) which does not appear to be dependent on allele size. As also noted earlier (Jeffreys et al., 1990), regions of internal tandem duplication can be observed, confirming a role for either unequal sister chromatid exchange or replication slippage in minisatellite evolution. Internal differences within regions of haplotypic similarity mostly include small insertion/deletions and very short patches of repeat unit 'conversion' (ie switching of a- and t-type repeats) that do not change repeat copy number or disrupt alignments. Alignments often break down toward the 3' end of the mapped region (the majority of alleles exceed 50 repeat units and are
### Figure 6.3

#### Group B

<table>
<thead>
<tr>
<th>Race</th>
<th>H1</th>
<th>H2</th>
<th>MVR-haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>G + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>G + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>G + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>G + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>G + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
</tbody>
</table>

#### Group C

<table>
<thead>
<tr>
<th>Race</th>
<th>H1</th>
<th>H2</th>
<th>MVR-haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>J + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>J + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>J + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>J + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>J + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>J + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>J + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>J + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>J + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>J + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
</tbody>
</table>

#### Group D

<table>
<thead>
<tr>
<th>Race</th>
<th>H1</th>
<th>H2</th>
<th>MVR-haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>G + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>G + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>G + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>G + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
<tr>
<td>G + C</td>
<td></td>
<td></td>
<td>aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa</td>
</tr>
</tbody>
</table>

Chapter 6 Page 4b
Figure 6.3.

**Group E**

<table>
<thead>
<tr>
<th>Race</th>
<th>H1</th>
<th>HH</th>
<th>HE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>9</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

**Group F**

<table>
<thead>
<tr>
<th>Race</th>
<th>H1</th>
<th>HH</th>
<th>HE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

**Group G**

<table>
<thead>
<tr>
<th>Race</th>
<th>H1</th>
<th>HH</th>
<th>HE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

**Group H**

<table>
<thead>
<tr>
<th>Race</th>
<th>H1</th>
<th>HH</th>
<th>HE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>1</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

**Group I**

<table>
<thead>
<tr>
<th>Race</th>
<th>H1</th>
<th>HH</th>
<th>HE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>1</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

**Group J**

<table>
<thead>
<tr>
<th>Race</th>
<th>H1</th>
<th>HH</th>
<th>HE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Chapter 6 Page 4c
Figure 6.3. Examples of groups of aligned MS32 alleles.

For each allele its ethnic origin (Caucasian (C), Japanese (J), Bangladeshi (B) or Papua New Guinean (P)), flanking haplotype (Hump1 (H1, alleles H1^C (C) and H1^G (G)), Hf (alleles Hf^+ (+) and Hf^- (-)) and Hump2 (H2, alleles H2^C (C) and H2^T (T))) and MVR haplotype are shown. The major group specific, portion of each haplotype shared amongst members of an alignable group is shown in red. Positions of divergence are shown in black and other colours show regions of subgroup homology. Gaps (--) have been introduced to improve alignments. Some alleles have been mapped using single offspring and therefore show uncertain positions (?). Some regions of internal tandem duplication are indicated with arrows (--->) above the allele. The ends of short alleles (<-) and the unknown haplotype of long alleles beyond the mapped region (....) are shown. Alleles derived from homozygous individuals (*) are shown and two pairs of closely related alleles showing switching of flanking haplotypes are also shown (Δ and *).

Allele alignments. Alleles showing shared zones of haplotypic identity were identified by A.J. Jeffreys using two approaches: (a) the MVR haplotype of each allele was compared pairwise with every other allele and analysed for the proportion of matching repeat positions and the proportion of matches which were a-type repeats. For each pair of alleles, comparisons were repeated for alleles misaligned up to ± 10 repeat units out of register. Allele pairs which showed a high proportion of matches, but a relatively low proportion due to a-type repeats (homogeneous a-rich alleles often show a misleadingly high match frequency with each other due to the presence of large tracts of a-type repeats) were likely to represent bona fide matches (Jeffreys et al., 1991b). (b) Pairwise dot matrix analysis of each allele with every other allele was also performed searching for perfect 9 repeat matches. Match points consisting of more than five a-type repeats were discarded. Allele pairs showing a cumulative match score of greater than 20 matches over the best two diagonals were preselected (A.J. Jeffreys, personal communication); the authenticity of selected matches and the final alignment of allele groups were made by eye, with gaps inserted to improve alignments.
only partially mapped by MVR-PCR), but whether this represents a complete loss of homology is unclear. This is particularly common in the two very large ancient haplotypes and may be due to the action of large, but rare, internal deletion events (Jeffreys et al., 1990). Alternatively, a loss in homology toward the 3' end of the mapped region may represent the product of an internal interallelic recombination or interallelic conversion event. However, without knowledge of the full 3' haplotypes these alternatives are indistinguishable. The small scale terminal divergences most frequently observed do not contain enough information to determine if any of these have interallelic origins.

Allele alignments-flanking haplotypes. In general, groups of related alleles also share flanking haplotypes, confirming the authenticity of the allele alignments. Nonetheless, some clearly related alleles show exchange of flanking markers, providing direct evidence for interchromosomal exchange operating in this region (see Figure 6.3 and 6.4 for examples). The very high relatedness of these allele pairs (ie the alleles within a pair are assumed to share a recent common ancestor), compared to the inter-pair divergence (ie alleles from two different pairs are assumed not to share a recent common ancestor), suggests that the exchange events observed are mostly independent events. If we consider this region in terms of expected recombination frequencies, then using the genomic average of 1cM per Mb and the distance involved (~300bp) this region should have a genetic distance of ~3x10^{-6}cM. Thus we would expect to observe an allele to exchange flanking markers approximately once every 330,000 generations. Applying the known mutation rate of 1% (Chapter 7), two alleles separated by this many generations should have accumulated ~3,300 mutations. It seems unlikely that alleles diverged by this many mutation events would remain obviously related. This would indicate a very high level of interallelic exchange operating over a relatively small distance and suggests the observed variation hotspot and presumed mutation hotspot may also be a recombination hotspot. Whether these observed changes represent true interallelic recombination events, as opposed to local interallelic conversion events, cannot be deduced at present and awaits the placing of MS32 in a wider haplotypic framework. The use of these data to estimate the recombination frequency operating over this region is discussed in Chapter 7.

Allele alignments-population analysis. As well as its use in minisatellite evolution, allelic analysis may also reveal information relevant to human population studies. The two largest groups contain alleles from all populations studied so far, suggesting these haplotypes are indeed ancient, pre-dating the human radiation (~50,000 years ago, see Cann et al., 1987). Frequently, subgroups of alleles sharing a further region of 5' homology can be identified, presumably reflecting more closely related alleles. This is particularly apparent for the two ancient haplotypes, which contain many, often population specific, subgroups (eg Group A, Figure 6.3). The majority of groups though, contain far fewer alleles (most groups contain less than 10 alleles) and are population specific; 75% contain only Caucasian or Japanese alleles (eg Groups D and G, Figure 6.3 and Figure 6.5). However, it should be noted that considering the extreme allelic variation observed, the sample sizes obtained so far are still relatively small and it is probable that some groups may expand and incorporate alleles of widening origins. Nonetheless, many groups are likely to remain population specific and may have value as population specific markers. For instance group D (Figure 6.3) is a large Caucasian-only group, containing no Japanese alleles and would so far appear to be a good marker of current Caucasian populations. Allelic haplotypes apparently specific for one or more populations can have two major origins, either: firstly, they represent new haplotypes arising after population divergence, or secondly they represent ancient haplotypes differentially lost in some populations. Obviously, incorrect interpretation of the above two possibilities could produce misleading conclusions and distinguishing these two possibilities probably represents the major
Figure 6.5. Population content of the groups of aligned alleles.

For each group the number of alleles derived from Caucasian (blue), Japanese (yellow) and other (Bangladeshi, Papua New Guinean and Melanesian, red) individuals is shown. The groups were derived as in Figure 6.3, and the examples shown in Figure 6.3 (A-I), are also indicated.
challenge to the effective use of MVR maps in population analysis. The large groups of ancient haplotypes show very ragged alignments, with zones of haplotypic alignment often being relatively small and showing multiple switching of flanking haplotypes. However, the population specific groups, even some quite large groups, are far less ragged displaying large zones of identity and almost exclusively terminal divergence, with very little switching of flanking haplotypes. These results suggest that the distinction between ancient and new haplotypes may not be so difficult to make, with alleles mutating apart at the rate of one mutation about every 50 generations (~1000 years). Besides which, the loss of major haplotypes from a population is only likely to affect populations having gone through a major bottle-neck; such a limitation applies equally to other standard genetic markers currently used in population analysis.

Remembering the relatively small sample numbers involved, a few tentative preliminary observations can still be made. It has already been noted, from diploid code analysis, that Japanese individuals contain a higher proportion of short alleles (<50 repeats); 14% of alleles are short, compared to about 3% in Caucasians (Jeffreys et al., 1991b). This present survey revealed six groups of short alleles, four Japanese specific groups containing two to five alleles (e.g. Groups G, H and I, Figure 6.3), one group containing two Caucasian and one Bangladeshi allele (Group J, Figure 6.3) and one almost exclusively Japanese group containing 13 Japanese alleles and one Caucasian allele (Group F, Figure 6.3). Of the 25 small Japanese alleles so far mapped only three did not fall into an alignable group. Thus, the five groups of short Japanese alleles identified represent the major proportion of short alleles identified in Japanese populations (~88%), with the majority (>50%) falling into a single group (Group F). Interestingly, the one Caucasian allele in group F is identical to a Japanese allele, however both alleles are comprised almost exclusively of a-type repeats (apart from two 3' terminal t-type repeats) and have different flanking haplotypes. It would seem probable that these alleles are an example of convergent evolution, arising through independent lineages, rather than a direct genealogical relationship. The only other example of an allele shared across a population boundary is again amongst three alleles comprised exclusively of a-type repeats (two Caucasian and one Japanese; data not shown) and once more, likely represent convergent evolution.

Although only a few Bangladeshi and Papua New Guinean alleles have been mapped, their distribution amongst allele groups appears to suggest certain population ancestry. Not surprisingly, only one small population specific group of Papua New Guinean alleles has been identified (containing only 3 different alleles). Interestingly however, other than those alleles in the two large ancient haplotypes (3/17), Bangladeshi alleles are almost all (12/17) found in otherwise Caucasian specific groups (e.g. Groups B and E, Figure 6.3). Whether these preliminary results are truly representative of genuine population lineages awaits confirmation with the typing of further alleles.

Conclusions

MVR-PCR is a powerful technique for studying allelic structure and the allelic diversity revealed by it at MS32 is considerable. This allelic diversity forms the basis of the even larger diploid code diversity, highlighting the massive potential resolving power of a single MVR-PCR test in individual identification. Analysis of allelic structure through the use of allele alignments to produce haplotypic groups has confirmed the previously observed polarity in variation and revealed insights into the potential mechanistic basis of mutation with data suggesting the existence of a mutation and/or recombination hotspot operating over the 5' terminus of MS32 (discussed in more detail in Chapter 7). With such a high mutation rate, high variability, ease of allelic definition and the ability to group related alleles, the potential for the analysis of recent population divergence is obvious. Such allele alignments have already been used in tentative population analysis, although a more
thorough evaluation of the applications in this area requires a more comprehensive survey of allelic diversity in a wider range of populations.
MINISATELLITES, MUTATION AND INTERALLELIC RECOMBINATION

Summary

Ever since the initial discovery of large numbers of hypervariable minisatellites in the human genome it has been speculated that such sequences may play a functional role in chromosome synapsis and/or meiotic recombination. Initial findings provided circumstantial evidence to support this hypothesis, but more recent molecular data have suggested that unequal homologous exchange is not a common feature of minisatellite mutation. Minisatellite internal variation is now being used to reveal the true level of allelic diversity present at some loci and study in detail the changes occurring during minisatellite mutation. MVR-PCR pedigree analysis was used to detect and study de novo mutation events at the hypervariable human minisatellite MS32. In the CEPH panel of 40 large families, 7 mutation events were detected and characterised by MVR-PCR. Interestingly, all of these changes involve small additions of repeat units, with an extreme bias in recipient and donor breakpoints toward the 5' terminus of MS32 and all are totally conservative with respect to the recipient allele. The mutational hotspot so revealed correlates well with the previously observed polarity in variation. Two of the events characterised appear, for the first time, to provide evidence for the action of unequal interallelic exchange. These results, in addition to the haplotypic analysis of flanking polymorphisms, suggest the existence of a recombination and/or conversion hotspot operating over the 5' terminus of MS32. Detailed sequence analysis of one, apparently simple, mutation event has revealed a greater complexity than anticipated and suggests a role for unequal interallelic conversion in minisatellite evolution. Finally, a speculative mechanistic model, based on the double strand break repair model of homologous recombination, is proposed to account for the observed properties of minisatellite mutation events.

Introduction

Minisatellites as recombinators. The tandemly repeated nature of minisatellite DNA and its often high copy number variability are properties highly suggestive of being a consequence of unequal exchange between misaligned alleles. The idea that minisatellite variability may be a by-product of a functional recombinagenic role appeared at first to be very attractive. This intuitive basis for a possible recombinagenic role appeared to be supported by a high degree of sequence similarity between the first few minisatellite repeat sequences obtained and the recombination signal of E. coli (Jeffreys et al., 1985a). The repeat unit sequence of each minisatellite, although different, appeared to share an apparently conserved G/C rich core region of about 10 to 15bp and it was speculated that minisatellites may be acting as promoters of recombination in the human genome, mediated through the action of a recombinagenic core (Jeffreys et al., 1985a). As more minisatellite loci were cloned an increasing divergence from the original core was noticed (Wong et al., 1987) and new core consensus sequences were proposed (Nakamura et al., 1987a, 1988). Although the majority of minisatellite loci do have a high G/C content, some minisatellites containing A/T rich repeats, appear to have no obvious similarity to the core
(Vergnaud et al., 1991). The relevance of the core sequence remains in considerable doubt; although it is far from an absolute requirement of minisatellites, so many loci contain a close-to-perfect match with the core that a sequence specific component of minisatellite instability is still a significant possibility, rather than the core being an artifact arising through ascertainment bias due to isolation of minisatellites by cross hybridization.

A significant bias in the genomic distribution of human minisatellites exists. In situ hybridization (Royle et al., 1988) and linkage mapping (Nakamura et al., 1988; Armour et al., 1990) revealed a tendency for minisatellites to cluster in the proterminal regions of human chromosomes and/or toward the end of genetic linkage maps (also assumed to be the proterminal regions of human chromosomes). The subtelomeric regions of the human genome are precisely those regions known to have a high recombination rate and to be the sites for the initiation of chromosome synopsis and pairing during meiosis (Solari, 1980; Laurie and Hulten, 1985). Furthermore, hybridization of minisatellite core probes appeared to show preferential localization to chiasmata in human bivalents (Chandley and Mitchell, 1988), once again suggesting a link between minisatellites and recombination.

If minisatellites are really important as functional components of the recombination/chromosome synopsis/pairing machinery then we might expect them to be highly conserved between species. However, PCR based analysis of a number of human hypervariable minisatellites has revealed that the cognate loci in primates, as closely related as Chimps and Gorillas, are often short and monomorphic (e.g. Gray and Jeffreys, 1991; Armour et al., 1992). These results argue against, but do not preclude, a conserved functional role for minisatellite sequences.

Pedigree analysis of five hypervariable human minisatellites in the CEPH panel of families has revealed significant rates of mutation to new length alleles for some loci (Jeffreys et al., 1988b). As expected, the mutation rate observed correlates well with measured heterozygosity, with the most variable human minisatellite MS1 having an allele length heterozygosity of ~99% and a mutation rate of ~5% per gamete, consistent with a neutral mode of evolution (Jeffreys et al., 1988b). The size changes observed were apparently equally distributed between gains and losses consistent with a role for unequal exchange (assuming such a process produces reciprocal products) and/or replication slippage, which may both be expected to produce gains and losses at equal frequency. Paternal and maternal mutation rates were approximately equal, which considering the difference in postzygotic cell divisions between mature spermatocytes and oocytes (~400 and ~24, respectively (Vogel and Rathenburg, 1975)) suggest that mutation is restricted to one stage of gametogenesis, possibly meiosis; this tends to argue against a role for mitotic recombination and replication slippage which might both be assumed to be dependent on the number of mitotic divisions involved (Jeffreys et al., 1988b).

A necessary consequence of a simple unequal exchange model as the basis for minisatellite variability is the exchange of flanking markers in de novo mutation events. This prediction has been tested directly in a limited number of instances. Wolff et al., (1988) characterised in detail a single de novo mutation event at the human minisatellite YNZ222. The two parental alleles contained three and four copies of a 70bp repeat and the derived mutant two copies. Sequencing across the minisatellite array and immediate flanking DNA of both parental alleles revealed three base substitutional polymorphisms within 20bp of the tandem array on one side and three more polymorphisms around 2kb upstream on the other side. The derived mutant was parental on both sides of the tandem array for the three repeat allele, thus ruling out simple unequal interallelic exchange. No internal variation between repeat units was identified, and thus tandem repeat breakpoints could not be defined. A study of much more distal flanking markers in 12 mutation events identified at another hypervariable human
minisatellite (MS1) in the CEPH panel of families demonstrated that the frequency of exchange of these distal flanking markers was not significantly elevated in mutant alleles (Wolff et al., 1989). More recently Vergnaud et al. (1991) have isolated a very unusual hypervariable human minisatellite. The CEBl locus, located in the proterminal region of chromosome 2, has a very high, but male specific, mutation rate of ~15% per gamete (female mutation rate ~0.3% per gamete). Size changes were biased toward small increases (72%), but with no obvious dependence on allele size and with no evidence for an increase in the exchange rate of distal markers.

Thus far, only limited, and often conflicting evidence concerning minisatellite mutation processes has been obtained. The data available suggest that simple unequal interallelic recombination does not appear to be the major mechanism of variation operating at all minisatellite loci. Nonetheless, interallelic exchange may still be important for a subset of minisatellites and/or mutation events and, for all of the events described above, the possibility for the involvement of more complex mechanisms, such as unequal interallelic conversion, could not be ruled out.

General recombination mechanisms. The ease of manipulation of fungi and the ability to dissect, via tetrad analysis, all four duplexes present in a meiotic cell, has greatly facilitated the study of meiotic recombination and conversion. As such, the majority of information relating to these processes has been obtained in fungal systems, mainly in the yeast Saccharomyces cerevisiae. Since the initial proposal for the mechanistic basis of recombination (Holliday, 1964), two major competing models of meiotic recombination and conversion have been described. The model proposed by Meselson and Radding (1975) is based strongly on the original Holliday model and involves initiation by a single stranded nick on one chromosome. This model accounts for many of the observed genetic properties of meiotic recombination and conversion, but does require certain constraints to fully fit the data (see Szostak et al., 1983). One of the most significant deviations from in vivo results is that the Meselson-Radding model predicts the initiating chromosome to be the donor in unequal conversion events, whereas this has in fact been demonstrated to be the reverse (see Szostak et al., 1983). To accommodate some of the genetic anomalies, the double strand break repair (DSBR) model was proposed and modified by Szostak et al. (1983; Sun et al., 1991a; described in Figure 7.1). Since its inception the DSBR model has gathered extensive experimental support, including the demonstration of physical intermediates such as double strand breaks (DSBs) (Sun et al., 1989; Cao et al., 1990) and reciprocal 3′ single stranded overhangs (Sun et al., 1991a) associated with known yeast recombination hotspots. The yeast Arg4 recombination hotspot has been extensively characterised and the initiation site (IS) localised to a 14bp poly-A tract (Schultes and Szostak, 1991). Deletion of the IS, when homozygous, reduces recombination by around 75% and, when heterozygous the deletion always acts as the donor of information, often deleting the intact IS in the initiating recipient (Schultes and Szostak, 1991). The site of the associated DSB does not however map exactly to the IS sequence but is actually positioned approximately 100bp upstream (Schultes and Szostak, 1991). Whether this poly-A tract is a specific recognition site for a double strand endonuclease which then cleaves at a distal site, or merely a low melt domain allowing access of such an enzyme has yet to be determined. Analysis of meiotic sister chromatid exchange (SCE) in yeast has revealed that it is initiated at the same sites and involves the same DNA repair proteins as meiotic recombination and would appear to occur via the same mechanism (Sun et al., 1991b). In summary, meiotic recombination would appear to be a highly controlled process with the involvement of specific sequences, specific proteins and progressing through defined pathways.

DSBR and recombination mechanisms in humans. Virtually nothing is known of meiotic recombination mechanisms in higher eukaryotes, largely due to the inavailability of the elegant meiotic systems.
Figure 7.1. The Double Strand Break Repair model of meiotic homologous recombination and gene conversion.

(a) A double strand break (DSB) is made in one duplex. (b) Extensive 3' overhanging single stranded tails are generated on both sides of the break by exonucleases. (c) One 3' end invades the homologous duplex displacing a D-loop. (d) The D-loop is enlarged by repair synthesis and annealed to the second 3' end single stranded DNA. (e) Repair synthesis from the second 3' end takes place and two Holliday junctions are formed. (f) Resolution of the two junctions by cutting strands 1-4 or 2-3 can give rise to non-crossover products (g) or by cutting strands 1-3 or 2-4 crossover products (h). Newly synthesised DNA is shown hatched. This figure was adapted from Sun et al. (1991).
available for the dissection of such processes in lower eukaryotes. Experiments designed to attempt to assess the role of DSBR in recombination have revealed a second major mechanism of DSBR operating in mitotic mammalian and yeast cells (Lin et al., 1984; Fishman-Lobell et al., 1992). It has been observed that DSBs introduced in vitro into mitotic cells are preferentially repaired via a non-conservative single stranded annealing (SSA) process mediated through short regions of homology (Lin et al., 1984; see Figure 7.2A). The SSA process is intramolecular and results in the loss of genetic material and obviously cannot be the mechanism by which meiotic recombination generally occurs. Nonetheless, it may have relevance to minisatellite biology and will be discussed later in the context of large deletions.

As also observed in yeast, targeted integration of DNA into mammalian chromosomes is facilitated by linearisation of DNA to reveal free double stranded ends, consistent with a role for DSBs in recombination. Recent experiments have demonstrated the existence of thymus specific DSBs present in vivo in new born mice located within the V-(D)-J joining region of the T-cell receptor δ locus (Roth et al., 1992). Although somatic recombination events within the immune system are probably not typical of general meiotic recombination events, these results do show that DSBs can play a functional role in at least some aspects of mammalian recombination in vivo.

MS32 deletion mutants and allelic variation. Study of rare large deletion events in bulk sperm and blood DNA by single molecule analysis and HaellI restriction based internal mapping at MS32 has revealed that the majority of such events do not involve interallelic exchange (Jeffreys et al., 1990). In addition, many such mutants were found to be mosaic in sperm DNA indicating a premeiotic origin, although large deletion events are rare and may not be typical of the majority of mutation events at MS32.

As demonstrated in Chapter 6, allelic variation at MS32 is very high and although the majority of alleles are unique, many share zones of haplotypic identity and can be arranged into groups of aligned alleles, which are likely to share a recent common ancestor (Jeffreys et al., 1990, 1991b; Chapter 6). Such alignments have shown clear evidence for polarity at MS32, with interallelic variation largely confined to the extreme 5' end of alleles and assumed to arise through the action of a local mutational hotspot (Jeffreys et al., 1990, 1991b; Chapter 6).

This work. The major mutational processes operating at MS32, or indeed any other minisatellite, are presently unknown. In this chapter, the analyses of seven mutation events identified by MVR-PCR pedigree analysis are presented and accompanied by haplotypic analysis of flanking polymorphisms in both mutant and non-mutant alleles. These results are discussed with respect to the role of interallelic recombination in minisatellite evolution and to a possible mechanistic basis. MVR-PCR pedigree analysis was performed by A.J. Jeffreys, A. MacLeod, D.L. Neil, K. Tamaki, M. Allen and myself. Part of this work has been published (Jeffreys et al., 1991b).

Results

Pedigree analysis and mutant identification. MS32 MVR-PCR pedigree analysis of parents and offspring in the CEPH panel of 40 large families was performed and the allelic haplotypes of all four parental alleles was determined as described in Chapter 6. During this process seven children were identified who produced
Figure 7.2. The Single Stranded Annealing model of non-conservative double strand break repair.

(A) The basic SSA model. (a) Initiation from a double strand break in a DNA duplex. (b) 5' exonucleases degrade the 5' strand to reveal 3' single stranded overhangs. (c) Regions of intramolecular homology (thin stripes) anneal and non-homologous DNA is excised (arrowed). (d) DNA synthesis (thick stripes) and ligation lead to gap filling and regeneration of duplex DNA.

(B) The SSA model applied to minisatellites. (a) Initiation from a double strand break within the tandem array. (b) 5' exonucleases degrade the 5' strand to reveal 3' single stranded overhangs. (c) Tandem repeats anneal. (d) DNA synthesis and/or ligation lead to regeneration of duplex tandem repeat DNA with concommitant loss of repeats and generation of a central heteroduplex domain.
diploid maps showing multiple parental exclusions. The CEPH pedigrees have been typed with over 1000 DNA polymorphisms and correct parentage is beyond dispute. Thus parental exclusions in the diploid codes of these children indicates that a new mutant allele has been inherited. In each case exclusions were specific to one parent only and by subtraction of the correctly inherited non-mutant allele from the diploid code, the structure of the mutant allele could be deduced (see Figure 7.3). MS32 MVR-PCR pedigree was also performed for eight Bangladeshi, seven small Fanconi anaemia families and 60 British mother-father-single child trios. No mutation events were detected in any of these pedigrees. From the total set of families (total of 360 children, 720 meiosis) we can calculate an MS32 MVR-PCR detectable mutation rate of 7/720 = 0.97% per gamete.

**Mutant analysis:** The structure of each mutant allele and both parental alleles, as revealed by MVR-PCR analysis, are shown in Figure 7.4 (except mutant a, see Figure 7.3), along with the probable origins of repeat units in the mutant allele and possible breakpoints. The properties of each mutation event and possible breakpoints are summarised in Table 7.1 and Figure 7.5.

**Mutant a** is a paternally inherited size increase of 13 repeat units which appears, for the first time, to show direct evidence for interallelic exchange in its generation. Defining the recipient as allele 2 (the major proportion of the mutant allele) then it would appear that this allele received a direct donation of the 5' terminal 11 repeats of allele 1 into its first repeat, with no loss of information from the recipient. However, such a mechanism does not explain the origin of the two extraneous a-type repeats (underlined in Figure 7.3) which may be the result of a small patch of conversion repair at the site of the mutation event. Alternatively, a mechanism involving donation of material internal to that so far mapped in allele 2, in a sister chromatid or intramolecular event cannot be discounted. Typing of the flanking markers showed the father (141601) to be homozygous for the Hf and Hump2 polymorphisms, but heterozygous and thus informative, for the Hump1 locus. Pedigree analysis of this marker in the whole family showed that this flanking site was not exchanged in the 141605 mutation event (see Figure 5.3, Chapter 5). This may be indicative of a local conversion event within the MVR haplotype or a genuine recombination event accompanied by conversion effecting the flanking DNA; the differentiation of these two possibilities awaits the placing of MS32 in a wider haplotypic framework.

In summary the structure of the 141605 mutant suggests an interallelic origin, but not a simple unequal exchange, although this has yet to be formally established.

**Mutant b** is a maternally inherited addition of a single repeat somewhere in the first 11 repeats of allele 3. This event could be the result of a simple unequal SCE or slippage, or possibly (if the untyped first repeat of allele 4 is a-type) simple unequal interallelic exchange.

**Mutant c** is another very simple maternally inherited addition of a single repeat with a recipient breakpoint located in the first three repeats of allele 3. Again, this event would appear to be the result of a simple unequal SCE or slippage, or possibly (if the untyped first repeat of allele 4 is a-type) simple unequal interallelic exchange.

**Mutant d** is a maternally inherited addition of a single repeat unit for which the recipient breakpoint lies within repeats 11 to 25 of allele 3. This event appears to be entirely intraallelic, probably arising by simple unequal SCE or slippage.

**Mutant e** is a paternally inherited addition of three a-type repeat units into the first two repeats of allele 2. However, allele 2 does not contain three potential donor a-type repeats in this region, therefore ruling out a simple unequal SCE or slippage event. A suitable donor site does exist in the other allele and the resultant
Figure 7.3. Analysis of MS32 mutant alleles detectable in pedigrees by MVR-PCR.

Example of a CEPH pedigree showing a child with a mutant allele. Maps of parental alleles 1-4 were deduced from 7 non-mutant offspring (not shown). Comparison of the diploid code of child 141606 with the parents shows 4 specifically paternal exclusions (p) plus 3 ambiguous exclusions (e) which do not indicate the parental origin of the mutant allele. There are no maternal exclusions, and thus the child has inherited a mutant paternal allele and non-mutant maternal allele. The diploid code of the child is compatible with the child having inherited maternal allele 3 but not 4. Subtraction of the code for allele 3 from the diploid code of the child yields the code for the mutant paternal allele (2M). Comparison of the mutant allele with paternal alleles 1 (blue) and 2 (red) indicates that this allele commences with the code of allele 1 and then switches to the beginning of the code of allele 2 after two a-type repeats of unknown origin (underlined). This allele therefore appears to have arisen by unequal crossing over between the two paternal alleles, as indicated, with possible cross-over sites marked X. During standard MVR-PCR the first minisatellite repeat does not usually produce a detectable signal and is not scored in diploid coding, and thus each allele starts with an unknown repeat unit (n).

This figure was adapted from Jeffreys et al. (1991b).
Figure 7.4. MVR maps of mutant and progenitor alleles for the CEPH MS32 mutation events.

For each mutation event the mutant allele (M) is shown in red, along with the two parental alleles (paternal, 1 + 2 or maternal, 3 + 4) and the individual inheriting the mutant allele. Mutant and parental alleles are aligned to show possible origins of all repeats in the mutant (shown in red) and possible exchange points (x) for unequal exchange events. For mutant c allele 4 is short. For mutant g the resultant tandem duplication in the mutant is indicated (red arrows).
Table 7.1. Summary of the properties of the CEPH MS32 mutants.

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Parental origin</th>
<th>Individual inheriting the mutant allele</th>
<th>Change in repeat unit copy number</th>
<th>Detected on Southern blot *</th>
<th>Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Paternal</td>
<td>141606</td>
<td>+13</td>
<td>+</td>
<td>inter-allelic</td>
</tr>
<tr>
<td>b</td>
<td>Maternal</td>
<td>134505</td>
<td>+1</td>
<td>-</td>
<td>intra-allelic?</td>
</tr>
<tr>
<td>c</td>
<td>Maternal</td>
<td>134606</td>
<td>+1</td>
<td>-</td>
<td>intra-allelic?</td>
</tr>
<tr>
<td>d</td>
<td>Maternal</td>
<td>142106</td>
<td>+1</td>
<td>-</td>
<td>intra-allelic</td>
</tr>
<tr>
<td>e</td>
<td>Paternal</td>
<td>142409</td>
<td>+3</td>
<td>+</td>
<td>inter-allelic</td>
</tr>
<tr>
<td>f</td>
<td>Paternal</td>
<td>1329405</td>
<td>+2</td>
<td>+</td>
<td>inter-allelic?</td>
</tr>
<tr>
<td>g</td>
<td>Maternal</td>
<td>140808</td>
<td>+34</td>
<td>+</td>
<td>intra-allelic</td>
</tr>
</tbody>
</table>

* Results of Southern blot analysis of AluI digested genomic DNA (Armour et al., 1989a).

Table 7.2. Disequilibrium measures for the MS32 5' flanking polymorphisms.

<table>
<thead>
<tr>
<th>Disequilibrium measure</th>
<th>Hump1 to Hf</th>
<th>Hf to Hump2</th>
<th>Hump1 to Hump2</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>-0.05</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>$\chi^2$ of D</td>
<td>13.5</td>
<td>15.4</td>
<td>22.0</td>
</tr>
<tr>
<td>D'</td>
<td>0.36</td>
<td>0.54</td>
<td>0.53</td>
</tr>
<tr>
<td>Q</td>
<td>13.4</td>
<td>21.8</td>
<td>15.3</td>
</tr>
</tbody>
</table>

All figures are based on the allele and haplotype frequencies derived from 160 Caucasian chromosomes (Tables 5.2 and 5.3, Chapter 5).

The disequilibrium measure D is a frequency dependent measure,

$$D = h \cdot p_A q_A$$

where,

- h = observed frequency of haplotype A A
- $p_A$ = frequency of allele A at locus 1
- $q_A$ = frequency of allele A at locus 2
- $p_B$ = frequency of allele B at locus 1
- $q_B$ = frequency of allele B at locus 2
- n = number of haplotypes sampled

and

- $p_A < q_A < 0.5$

For the null hypothesis of total equilibrium D is asymptotically distributed as a $\chi^2$ with one degree of freedom,

$$\chi^2 = (D_n)^{1/2} / (p_A p_B q_A q_B)^{1/2}$$

All the $\chi^2$ values obtained show a significant deviation with p<0.001.

D' is a frequency independent disequilibrium measure based on D,

$$D_{\text{max}} = p_A q_B$$
$$D' = D / D_{\text{max}}$$

The normalized disequilibrium statistic Q is also approximately distributed as a $\chi^2$ under the null hypothesis of total equilibrium with one degree of freedom,

$$Q = n (D^2 / p_A q_A + D^2 / p_A q_B + D^2 / p_B q_A + D^2 / p_B q_B)$$

All the Q values obtained show a significant deviation with p<0.001.

For full details of D, D' and Q, see Hedrick (1987).
<table>
<thead>
<tr>
<th>Mutant</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
<th>50</th>
<th>55</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recipient</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>f</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>Donor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>f</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
<td></td>
<td>xx</td>
</tr>
</tbody>
</table>

**Figure 7.5. CEPH MS32 mutant breakpoints.**

Possible locations (x) of unequal exchange points on the recipient and donor alleles (the recipient being defined as the allele comprising the majority of the allele). For presumptive intra-allelic (sister chromatid) unequal exchange, the donor and recipient alleles are identical. Note, MS32 alleles average 300 repeats in length.
mutant could be simply explained by a simple unequal interallelic recombination event. Unfortunately, the father (142401) is homozygous and therefore uninformative, for the three flanking polymorphisms identified so far.

**Mutant f** is another paternally inherited small addition of two repeats into the 5' terminus of one allele. From the MVR-PCR derived allele maps a simple unequal SCE or slippage event would appear to explain the structure of the derived mutant with no evidence for the involvement of the other allele (however see below for a more detailed analysis of this event).

**Mutant g** is the least typical of the mutation events so far identified comprising a relatively large internal tandem duplication of 34 repeat units (986bp, estimated allele length change from Southern blot analysis was ~1kb (Armour et al., 1989)). The breakpoint for the recipient allele cannot be precisely defined but lies within 14 to 19 repeats of the 5' terminus. The mutant structure is consistent with a simple unequal SCE or slippage event.

**Detailed analysis of CEPH mutant f.** From the MVR-PCR analysis this event appeared to be a simple unequal SCE or slippage event involving addition of two repeats somewhere into the six terminal 5' repeats of one allele, with no evidence for the involvement of the other allele. In an attempt to define precisely the breakpoints and origin of the two repeats inserted, the two parental alleles and the derived mutant were sequenced across the mutation boundaries and into the flanking DNA, in the hope of finding further informative sequence polymorphisms. For each allele a suitable t-type repeat distal to the mutation boundary was selected for sequencing (see Figure 7.6A). Genomic DNA was amplified by MVR-PCR to the point where PCR products could be visualised directly on agarose gels by staining with ethidium bromide; up to 20 repeat rungs on the MVR ladder could be generated (data not shown). Allele 1 was amplified by knockout MVR-PCR using primer 32-D2 from the total genomic DNA of the father (1329401) who is heterozygous for the Hf site (allele 2 is linked to the Hf" site and is not amplified by 32-D2). The other two alleles could not be obtained by knockout MVR-PCR (with primers available at the time), but the relevant t-type repeats occurred at heterozygous positions in the diploid code, thus allowing their isolation from standard diploid MVR-PCR using 32-D. The three required t-bands were excised from the gel, purified and sequenced (Figure 7.6C).

The sequence of the immediate 5' flanking DNA and first two repeats of MS32, as derived from the original λ clone (Wong et al., 1987), is shown in Figure 7.6B. The first repeat unit is defined as that containing the first full site for priming of the TAG-A and TAG-T specific MVR primers, although this position often does not produce a detectable signal by standard MVR-PCR and is therefore not scored in diploid mapping (diploid mapping scoring is started from repeat unit number 2, code position 1). Also shown is a four base pair insertion, termed Δ (designated Δ+, for present and Δ-, for absent). These four bases were present in the original MS32 sequence obtained but were not present in the two unrelated alleles sequenced in this study. In addition, these four bases were not present in all primates sequenced previously (Gray and Jeffreys, 1991) suggesting the ancestral status is Δ-. At present it is unclear whether the Δ site represents a genuine human polymorphism or whether the apparent Δ+ allele is a cloning/sequencing artifact of the original MS32 analysis.

The father (1329401) is heterozygous for both the Hf and Hump2 flanking polymorphism, as revealed by PCR analysis (see Chapter 5) and by direct sequence analysis (Hump2 only) and, as expected, the mutant allele shows no exchange of flanking markers.

As well as the polymorphic base substitution utilized to distinguish a- and t-type repeats in MVR-PCR (here defined as site-1), MS32 repeats also contain another commonly polymorphic C/T transition (site-2,
Figure 7.6. Detailed sequence analysis of MS32 CEPH mutant 1329405.

(A) MVR-maps of the parental alleles (1+2) and the derived mutant (1M). The Hf polymorphism is shown (+/-) and the t-type repeats used for sequencing are indicated in bold.

(B) Detailed sequence of the immediate 5' flanking DNA and the first two repeats. The position of the PCR primer 32-O and the first (code position 0) and second (code position 1) full minisatellite repeats detected by MVR-PCR are indicated, as well as the location of two imperfect cryptic repeats (\(\sim\)). Polymorphic positions (A/G- polymorphic site assayed in MVR-PCR (MVR-site1); C/T additional polymorphic site not assayed in MVR-PCR (MVR-site2)) and the potential polymorphic four base deletion (\(\Delta\)) are also shown.

(C) Summary of detailed sequence analysis of CEPH mutant 1329405. The sequence-derived status of all polymorphic sites in the flanking DNA and first few repeats of the two parental alleles (1+2) and the derived mutant (1M). 100ng samples of total genomic DNA were MVR-PCR amplified in 30\(\mu\)l reactions using the primers described previously (see Chapter 5), with an annealing temperature of 69°C and an extension time of 5 minutes for 30 cycles. Amplified products were electrophoresed through a 3% NuSieve GTG, 1% Sigma TypeI Agarose gel in 1x TBE and visualized by ethidium bromide staining. An appropriate “rung” in the MVR-PCR ladder required for sequencing was excised from the gel and purified by electroelution onto dialysis membrane. The purified PCR product was reamplified with PCR primers 32-D or 32-D2 and TAG using the same cycling conditions as before for a further 12 cycles. The double-stranded PCR product was re-purified by electrophoresis and electroelution and used to generate single stranded template by asymmetric PCR (Gyllensten and Erlich, 1988) and cycle sequenced using Taq polymerase (see Chapter 2).

(D) Pictorial representation of CEPH mutant 1329405. Alleles are aligned to show possible origins of all repeats in the derived mutant. Red arrows, MVR site1-t, site2-T; yellow arrows, MVR site1-a, site2-C; blue arrows, MVR site1-t, site2-C; green arrows, MVR site1-a, site2-T; yellow/green arrows, MVR site1-a, site2-unknown; and red/blue arrows, MVR site1-t, site2-unknown. The allelic state of the Hf, Hump2 (H2) and \(\Delta\) sites are also shown.
defining four repeats, Ca-, Ct-, Ta- and Tt-type), only two base pairs 5' of site-1 (Wong et al., 1987; see Figure 7.6B).

Other than Hump2, the Δ site and the known polymorphic MVR sites-1 and -2, no other sequence variants were detected in either of the two unrelated alleles in either the flanking DNA or the repeat units sequenced. However, significant variation at MVR site-2 was detected (Figure 7.6C), allowing some of the a-type repeats at the beginning of the alleles to be distinguished and a more precise definition of the mutation event to be made. The structure of the derived mutant defines the recipient breakpoint to precisely within the first repeat unit. More interestingly though, the mutation is defined as an expansion of one Ca-type repeat to three Ca-type repeats, incompatible with a simple slippage or unequal SCE mechanism. These results are though consistent with a mechanism involving unequal conversion, either from the other allele (which contains an appropriate donor sequence in the correct alignment, see Figure 7.6D) or from a sister chromatid (with a donor site distal to that sequenced). A speculative mechanism for the action of unequal conversion in tandem arrays and how it may be applied to this mutant is outlined in the discussion.

**Estimation of recombination frequency from:**

a) **Mutant analysis.** If the two mutation events most likely derived from interallelic recombination events (mutants a and e) are indeed genuine recombination events then we can use these figures to estimate an approximate recombination frequency operating over the 5' terminus of MS32. We have observed two apparent recombination events in 720 meiosis allowing us to estimate a recombination frequency of 2/720 = 0.27cM over the first ~400bp of the minisatellite, representing a 675 fold enhancement over the genomic average of 1cM per Mb.

b) **Aligned alleles.** Within the groups of aligned alleles a number of switches of flanking haplotypes have been observed, often between highly related alleles which are assumed to share a recent common ancestor (see Chapter 6 text and Figure 6.4). These data suggests a higher than expected recombination and/or conversion frequency between the minisatellite alleles and the flanking loci and have been used to estimate this rate by A.J. Jeffreys (personal communication). Briefly, this was performed as follows. The number of different alleles sampled and the known mutation rate (~1%) was used to estimate the effective population size (~100,000 for Caucasian populations). This figure was then used in computer simulations of allele coalescence, to estimate the average time at which 69% of alleles would align with at least one other allele (as observed in our data set). The average time to coalescence obtained was approximately 300 generations, and thus each allele in a group can be assumed to share a single common ancestor 300 generations previously. For each group, the most common flanking haplotype was assumed to be the ancestral state and any alleles with flanking switches were assumed to arise from a single recombination event. Using the known flanking haplotype frequencies (Table 5.3, Chapter 5) the recombination frequency producing the maximum likelihood estimate for the number of observed recombination events was calculated. This calculation yields a figure of approximately 0.18 per 300 generation allele lineage, giving a recombination frequency of 0.18/300 generations = 0.06cM, again representing a significant elevation over the genomic average (~150x).

c) **Flanking haplotypes.** As described in Chapter 5 the flanking polymorphisms have all been haplotyped to each other by pedigree analysis in the 40 CEPH families (Table 5.3, Chapter 5) and within a few unrelated Japanese individuals. Linkage disequilibrium between all the loci is significant (all show significant χ² deviations from the null hypothesis of random association, comparing observed to expected haplotype frequencies (see Table 5.3, Chapter 5) and for the disequilibrium coefficient D (Table 7.2)), although surprisingly in no case
is this absolute. The use of this type of haplotypic data in more complex analyses of linkage disequilibrium and the identification of recombination hotspots is a highly controversial subject (see Weir and Hill, 1985; Chakravarti et al., 1985; Hedrick, 1987). In a comprehensive review of disequilibrium statistical analyses Hedrick (1987) recommended the use of the frequency independent D' value, which varies from 1 (total disequilibrium) to 0 (total equilibrium) (see Table 7.2). With D' values of between 0.36 and 0.54 the all three loci would appear to show partial equilibrium with each other. Only one other comprehensive study of closely linked (<1kb) non-coding base substitutions in humans has been made (Nickerson et al., 1991). Unfortunately their raw data, including information on the number of haplotypes observed, is not given, and thus direct comparisons of D' values cannot be made. They used the Q statistic (see Table 7.2) as a measure of disequilibrium and showed that many loci linked by less than a kb displayed partial disequilibrium, with Q values from 6.5 to 70 (Q=1 for total equilibrium). In general they observed decreasing Q values for increasing distances (as expected), although this relationship was not absolute and some very small distances had relatively low Q values. For example two polymorphisms only 52bp apart had a Q value of only 19.6, comparable to the values obtained in this study (Table 7.2). The degree of disequilibrium observed is dependent on a number of factors, including the age of the mutations, the order of the mutations and random genetic drift, as well as the recombination frequency between the loci (see Weir and Hill, 1985; Chakravarti et al., 1985; Hedrick, 1987). The convergent nature of multiple recombination events means that the number of recombination events and hence rate, can only ever be indirectly estimated from population data. For example, the D' value obtained linking Hf and Humpl is the lowest observed at 0.36, even though this is the only pair which does not contain an obligatory recombinant haplotype (in Caucasians at least) and the observed haplotype frequencies could theoretically be derived entirely from drift. As such it would seem unwise to make estimates for the recombination frequency based on this data, other than to say that disequilibrium is not absolute with at least three recombinant haplotypes having been observed (a minimum of four recombination events are required to generate all eight possible haplotypes in a three locus-two allele system assuming a single mutation event for each allele).

Discussion

MVR-PCR pedigree analysis has been used here as a very sensitive tool for detecting de novo mutation events at a hypervariable human minisatellite locus in family groups. Of the seven de novo mutation events detected, only four had previously been identified by Southern blot allele length analysis (Armour et al., 1989a). The remaining three events involving addition of only a single repeat unit (29bp), a change too small to be detected by standard Southern blot length analysis, but producing multiple exclusions in MVR-PCR pedigree analysis.

The polarity in variation observed in aligned alleles is reflected in the location of the mutation events which also show an extreme bias toward the 5' terminus of MS32 alleles (Figure 7.5). All seven mutation events so far characterised have recipient breakpoints in the first 23 repeats (MS32 alleles contain an average of ~300 repeat). In fact, two mutants map exactly to the very first repeat (mutants a and f), with another three potentially doing so (mutants b, c and e). Obviously, with MVR-PCR mapping directed in from the 5' end of MS32 such pedigree analysis is biased towards detecting mutation events at this end. Although standard MVR-PCR diploid mapping extends at least 50 repeats into the minisatellite array and routinely exceeds 60+ repeats, the average length of an allele is ~300 repeats. However, all four of the length change mutations previously identified by Southern blot analysis (Armour et al., 1989a) were also detected by MVR-PCR analysis and also displayed the 5' bias. Thus, it seems likely that the 5' terminal mutation hotspot identified here represents the
major mutational hotspot operating at MS32, although the existence of another hotspot operating internally, or at the 3’ terminus, of large alleles producing very small length changes cannot be ruled out.

Two of the observed mutation events appear to suggest interallelic origins, revitalizing speculation that some minisatellites may indeed be involved in the processes of homologous recombination and/or chromosome synapsis and pairing. If these mutation events are truly interallelic then MS32 would represent an extreme example of a human recombination hotspot. This prediction would appear to be supported by evidence obtained from linking the flanking haplotypes to the aligned alleles, although the use of such population data to derive recombination rates involves numerous assumptions. Ideally, a direct measure of the observed recombination rate in large numbers of pedigrees, or through the analysis of bulk germline DNA using a PCR based assay (see Chapter 9) should provide a more satisfactory measure.

Interestingly, all the events identified so far involve small additions of repeats, suggesting a size increase bias. Such a property is not indicative of a model involving simple unequal exchange generating reciprocal products or replication slippage, both of which might be predicted to cause both size increases and size decreases with equal frequency. In addition, all of the changes would appear totally conservative, with no loss of information in the recipient allele, again incompatible with a simple mechanism of unequal interallelic exchange. Furthermore, at least two events suggest the involvement of interallelic gene conversion. Thus MS32 appears to be acting as a recombination and/or conversion hotspot, but not mediated through a simple unequal exchange process.

Unpublished results from preliminary studies of two other hypervariable human minisatellite loci (MS205, Armour et al. (personal communication) and MS31, Neil et al. (personal communication)), using similar techniques to those employed at MS32 have reflected many of the above observations. Terminal polarity in both allelic variation and germline mutation is common to both loci (MS205, 4/4 mutants terminal, 2 in the first repeat and MS31, 3/3 mutants terminal, 2 in the first repeat), although the size increase bias observed at MS32 is not obviously apparent at these other two loci (MS205, 2 small increases, 1 small deletion and 1 large deletion and, MS31, 2 small increases and 1 large deletion). One interesting event at MS205 would appear to be interallelic with very similar properties to MS32 mutant a, ie a terminal insertion of the end of the homologous allele, with no loss of information in the recipient and no exchange of close flanking markers. Thus, it would appear that the results obtained at MS32 may be at least partially reflective of a significant proportion of human minisatellites.

Based on a limited number of examples from only three loci a very general pattern of minisatellite mutation processes is appearing. The major mechanism emerging seems to involve small additions of repeats, from either the sister chromatid or homologous chromosome, into a terminal region of one allele, with no loss of information in the recipient and which may or may not involve the exchange of flanking markers. Such a mechanism cannot be explained by simple unequal interallelic or sister chromatid exchange. We can therefore conclude that the mutational processes operating at minisatellite loci may be more complex than initially anticipated.

An interesting question that arises is can these observed mutational properties be incorporated into any of the current models of homologous recombination? Wolff et al. (1991) have previously proposed a model of minisatellite mutation involving recombination and/or conversion, based on a simple Meselson-Radding type mechanism (see Figure 7.7). Briefly, they propose that: 1, a Holliday junction forms within the tandem array between two misaligned homologous chromosomes; 2, the junction may be resolved within the array producing
Figure 7.7. The 'Wolff' model of interallelic recombination and/or conversion and minisatellite mutation.

(a) Progenitor alleles. (b) A Holliday junction is formed between the two misaligned alleles. (c) The Holliday junction is resolved to give reciprocal products, one size increase, one size decrease (i.e. no net change); or (d) the Holliday junction moves out of the minisatellite array into the flanking DNA. Since the two alleles are misaligned a single stranded loop out forms in both products, equal to the size of the misalignment. (e) Resolution yields products with single stranded loop outs (conversion products only shown) which may be repaired by excision (leading to size decrease in one allele (the white allele) or no length change in the other allele (the black allele)) or insertion (leading to no length change in one allele (the white allele) or a size increase in the other allele (i.e. terminal insertion in the black allele)).

This figure was adapted from Wolff et al. (1991).
an internal conversion or a simple unequal interallelic recombination; or 3, the Holliday junction migrates beyond the terminus of the array into the flanking DNA. Since the alleles are aligned out of register the migrating Holliday junction will reach the flanking DNA of one allele, whilst still in the repeat DNA of the other allele. Thus a loop out of single stranded tandem repeat DNA forms until the flanking DNA of the other allele is reached and a normal migrating Holliday junction is once again formed (ie an effective deletion, equal to the size of the misalignment, is encountered). Resolution in the flanking DNA will result in interallelic conversion (without exchange of distal flanking markers) or recombination. The loop out can be either excised (producing no length change or a deletion), or repaired (producing an increase equal to the size of the original misalignment or no size change). This model does fit some of the data allowing for simple unequal interallelic exchange (although this has never yet been observed) and terminal unequal recombination/conversion. If one assumes the Holliday junction more often migrates into the flanking DNA, then terminal polarity can also be accommodated. However, unless a directional constraint on branch migration is assumed, then polarity would be directed toward both termini (only unidirectional polarity has been observed so far). Sequences inserted into the terminus of one allele could only be those terminal in the other allele, ie terminal insertion of interstitial sequences cannot be accounted for. Furthermore, assuming equal excision/insertion repair of the single stranded loop out, then this mechanism should produce equal numbers of size increases and size decreases. If, though, a bias toward insertion of the single stranded loop out were assumed, then a size increase bias would also result, although in vitro experiments in mitotic mammalian cells suggests single stranded loop outs are preferentially deleted (Weiss and Wilson, 1987). In addition, fungal experiments have indicated that branch migration beyond large deletions is rare, with a bias toward resolution at the deletion boundary (Hamza et al., 1981), ie a bias against asymmetric migration into the flaking DNA might be expected. In summary, the preliminary model proposed by Wolff et al. (1991) does not accommodate all of the current data, or agree with the predictions of branch migration and loop out repair. Below I propose a highly speculative mechanistic model that more readily accounts for the majority of the observed physical phenomena.

**The Gap Expansion Model of unequal conversion and minisatellite mutation.** This model is based on the DSBR model of homologous meiotic recombination proposed by Szostak et al. (see Sun et al., 1991a; Figure 7.1) and attempts to explain how the observed physical consequences of minisatellite mutation, such as conservative terminal insertion and the possible size increase bias, could be achieved. This model applies to both homologous recombination and SCE, assuming both proceed through the same pathway, and merely differ in the repair template used (although for ease of reading this account discusses consequences in terms of interallelic exchange).

The underlying principle of this model is that minisatellite length changes may be brought about by aberrant repair of DSBs. Thus, the initiating event in minisatellite mutation is proposed as being the generation of a DSB within the tandem repeat array. Such a DSB could be repaired by one of three possible mechanisms: Firstly, the DSB is correctly repaired through a DSBR mechanism where all repeats are precisely aligned, using the homologous chromosome as the template. This pathway would produce equal interallelic recombination and/or conversion with no resultant size change. Secondly, the free ends could diffuse laterally apart, misaligning on the template such that subsequent DNA synthesis and repair results in the de novo generation of new material through unequal conversion, ie a size increase (ie a size increase/gap expansion, Figure 7.8A). Thirdly, the free ends could laterally diffuse together, misaligning on the template such that repair results in the deletion of material (Figure 7.8B).
Figure 7.8. Minisatellite length changes produced by aberrant repair of DSBs.

(a) Initiation by a DSB in one allele (the recipient, white arrows). (b) Exonuclease activity generates 3' overhangs on both sides of the break.

(A) Gap expansion. (c) Lateral diffusion (gap expansion) of the free ends and out of alignment strand invasion into the homologous duplex (the donor, black arrows) displaces a D-loop. (d) The D-loop is enlarged by DNA synthesis and annealed to the second single stranded 3' end. Repair synthesis from the second 3' end takes place and two Holliday junctions are formed. (e) Resolution yields recombinant (not shown) and/or conversion products. The recipient gains repeats in the gap expansion domain (GE), whereas the donor does not change length. Both products may contain heteroduplex domains (hDNA).

(B) Gap collapse. (c) Lateral diffusion (gap collapse) of the free ends and out of alignment strand invasion into the homologous duplex (the donor, black arrows) displaces a D-loop. (d) The D-loop is enlarged by DNA synthesis and annealed to the second single stranded 3' end. Repair synthesis from the second 3' end takes place and two Holliday junctions are formed. (e) Resolution yields recombinant (not shown) and/or conversion products. The recipient loses repeats from the gap collapse site (GC), whereas the donor does not change length. Both products may contain heteroduplex domains (hDNA).
Thus the repair of DSBs within tandem arrays is capable of producing both size increases and deletions and all of the observed MS32 mutants could be interpreted using this basic model of an initiating DSB and subsequent gap expansion (although the two extraneous a-type repeats in mutant a still have no obvious origin). Two events (mutants d and g) would require interstitial DSBs, but the remaining five events could be accounted for with an initiating DSB located in the very first repeat unit; in fact mutants a and f have obligate first repeat breakpoints. An initiating DSB located in the very first repeat unit would have several important implications for minisatellite evolution. Such a DSB may be processed in four ways: Firstly, the DSB is correctly repaired with all repeats precisely aligned producing equal interallelic recombination and/or conversion, possibly progressing into the flanking DNA, with no resultant size change. Secondly, the DSB is repaired by the same mechanism except that the minisatellite free end (ie the end containing the major proportion of the minisatellite tandem array) misaligns at an interstitial position on the homologous chromosome (Figure 7.9A). This pathway would produce terminal insertions of the terminal sequences of the homologous allele, equal to the size of the misalignment, again possibly involving migration into the flanking DNA. Thirdly, the DSB is repaired with both free ends misaligning on the template, producing terminal insertions of interstitial homologous sequences (Figure 7.9B). Branch migration into the flanking DNA may be prevented by loss of homology, or alternatively the single stranded loop out structure may be formed and repaired as proposed in the Wolff model. Fourthly, the minisatellite free end misaligns, slipping in one repeat, to produce a single repeat deletion (Figure 7.9C).

This model allows for the insertion of either terminal or interstitial material from the homologous chromosome into the terminal position of the recipient, with no loss of information in the recipient and with or without the involvement of flanking markers in an unequal recombination and/or conversion event. As described above, the Wolff model cannot account for the conservative insertion of interstitial sequences into a terminal site in the recipient, as observed for mutant f. The gap expansion model proposed above does allow for such a phenomena and its possible application to mutant f is shown (Figure 7.10).

A terminal initiating DSB would explain many of the observed mechanistic properties of minisatellite mutation processes including terminal polarity, conservative insertion and the size increase bias (a terminal initiating DSB can only induce a one repeat collapse), but does not ipso facto explain the origin or position of such a DSB. An attractive possible hypothesis could be to assume the existence of a specific initiation sequence located in the flanking DNA directing DSB formation to a distal site, ie the tandem array, often the first repeat; however, comparisons of minisatellite flanking DNA have not revealed any obvious sequence elements conserved between minisatellites (A.J. Jeffreys, personal communication). If though this were the case, then even in the absence of the minisatellite we may expect the locus to be a recombination hotspot. A minisatellite could theoretically derive from such a recombination hotspot if an initiating, presumably rate limiting, duplication occurs at the site of the DSB, thus acting as a template for further DSB induced gap expansion. If correct, then minisatellites could therefore be viewed as deriving from recombination hotspots, rather than as being recombinators in their own right.

Obviously, if a genuine bias toward small increases at minisatellite loci really exists then this must be balanced by some other process reducing allele lengths, otherwise all alleles would increase in size indefinitely. Such a balancing process could be DSB repair in mitotic cells. In vitro experiments have been used to show that DSBs in mitotic cells are preferentially repaired via a non-conservative SSA mechanism (Figure 7.2A). Presumably, with tandem repeat arrays the potential for SSA mediated collapse repair of an interstitial DSB must be high; also random DSBs occuring in mitotic cells may be expected to lead to large deletions (see Figure
Figure 7.9. Aberrant repair of DSBs in the first repeat of minisatellite tandem arrays.

Continued overleaf.
Figure 7.9. Aberrant repair of DSBs in the first repeat of minisatellite tandem arrays.

(a) Initiation, from a DSB in the first repeat of the tandem array, with processing to reveal 3’ single stranded overhangs. The flanking DNA and first two repeats (arrows) of the recipient allele are coloured green and interstitial repeats blue. The flanking DNA and first two repeats of the donor allele are coloured red and interstitial repeats yellow. Newly synthesized DNA is striped.

(A) Gap expansion from a terminal donor site. (b) Lateral diffusion of the minisatellite free 3’ end. (c) Out of alignment strand invasion of the minisatellite free 3’ end into the homologous chromosome. (d) D-loop expansion promoted by DNA synthesis and annealing of the second free 3’ end, in register. Repair synthesis and ligation to form the double Holliday junction structure, with possible branch migration into the flanking DNA. (e) Resolution to yield recombinant and/or conversion products, only the converted recipient product is shown (including regions of heteroduplex DNA). This mechanism produces insertion of terminal information from the donor chromosome into a terminal position in the recipient, with no loss of information in the recipient. The donor undergoes no length change, but is possibly terminally converted.

(B) Gap expansion from an interstitial donor site. (b) Lateral diffusion of both free 3’ ends. (c) Out of alignment strand invasion of the minisatellite free 3’ end into the homologous chromosome. (d) D-loop expansion promoted by DNA synthesis and annealing of the second free 3’ end, also out of register. Repair synthesis and ligation to form the double Holliday junction structure. Branch migration into the flanking DNA may be prevented by loss of homology. (e) Resolution to yield recombinant and/or conversion products, only the converted recipient product is shown (including regions of heteroduplex DNA). This mechanism produces insertion of interstitial information from the donor chromosome into a terminal position in the recipient, with no loss of information in the recipient. The donor undergoes no length change, but is possibly interstitially converted.

(C) Gap collapse. (b) Lateral diffusion of the minisatellite free 3’ end. (c) Out of alignment strand invasion of the minisatellite free 3’ end into the homologous chromosome. (d) D-loop expansion promoted by DNA synthesis and annealing of the second free 3’ end, in register. Repair synthesis and ligation to form the double Holliday junction structure, with possible branch migration into the flanking DNA. (e) Resolution to yield recombinant and/or conversion products, only the converted recipient product is shown (including regions of heteroduplex DNA). This mechanism produces a terminal deletion of one repeat in the recipient.
7.2B). In concordance with this expectation is the finding that many large deletions are mosaic in bulk sperm DNA, indicating a premeiotic origin, and equally frequent in somatic blood cells (Jeffreys et al., 1990). In addition, Dubrova et al. (manuscript submitted) have recently shown that irradiated (radiation is assumed to induce random DSBs into chromosomal DNA) mice show an increase in minisatellite length change mutation. Although the direction of the size change events could not be determined, these results do demonstrate a probable mutational role for random DSBs. The probability of a random DSB should presumably increase directly with the length of DNA tract involved, thus longer alleles should be more prone to random DSB induced collapse, maintaining a stochastic balancing effect against more frequent small size increases.

Conclusions

Seven de novo germline mutation events have been characterised at the hypervariable human minisatellite MS32. All of these events were totally conservative size increases, with preliminary evidence for interallelic exchange and gene conversion. These data, along with flanking haplotype data, suggest that MS32 may be a recombination and/or conversion hotspot. The mechanistic basis for such a recombinagenic role in relation to minisatellite mutation processes has been discussed and a speculative gap expansion model proposed. Other factors possibly affecting minisatellite mutation and the future direction of minisatellite analysis is discussed further in Chapter 9.
Figure 7.10. Application of the gap expansion model to MS32 mutant 1329405.

This figure shows the possible origin of the 1329405 mutant assuming an initiating DSB in the first repeat of the recipient (allele 1) and repair, with gap expansion, from the homologous chromosome (allele 2). White boxes, unique sequence flanking DNA; green arrows, MVR site1-t; blue arrows, MVR site1-a; yellow boxes, MVR site2-C; red boxes, MVR site2-T; black boxes, Δ - and stripes, newly synthesized DNA. (a) Progenitor alleles. (b) Initiation by a DSB in the first repeat of allele 1. (c) Generation of free 3’ overhangs by exonuclease activity. (d) Gap expansion and out of alignment strand invasion of the minisatellite free 3’ end into the homologous chromosome. (e) D-loop expansion promoted by DNA synthesis and annealing of the second free 3’ end, also out of register. Repair synthesis and ligation to form the double Holliday junction structure. (Branch migration into the flanking DNA prevented by loss of homology?). (f) Resolution to yield the observed converted mutant.
Summary

Dispersed repeats comprise over 10% of the human genome, but have no known function and possibly represent an extreme example of intragenomic selfish DNA. Two major types of dispersed repeats exist in the human genome, these being the highly abundant Alu (> 500,000 copies) and L1 (>100,000 copies) elements, although other less numerous classes of dispersed repeats have also been characterised. Both Alus and L1s are assumed to have spread through the genome through the process of retrotransposition and are termed retroposons. The majority of human Alus and L1s at a given locus are present in the great apes, but some loci are human specific and a few are polymorphic in current human populations. A limited number of genetic diseases have been shown to be caused by de novo transposition events of both L1 and Alu elements into functional genes, suggesting that these elements are still actively transposing through the genome. The current rate of transposition is thought to be very low, but no suitable systems for measuring this rate have been developed. Previously, it has been shown that the frequency of dispersed repeats in the DNA flanking minisatellites is elevated over the genomic average, although the basis and relevance of this apparent clustering has not been determined. In this chapter, attempts to develop a system to measure and characterise de novo transposition events in the DNA flanking a hypervariable human minisatellite are described. The system developed is capable of detecting events as rare as 1 in 10^6 from bulk genomic DNA. A variety of source DNAs have been processed using this system, including germline, somatic and tumour DNA, but so far no insertion events have been detected.

Introduction

Dispersed repeats are a ubiquitous feature of higher eukaryotic genomes and may, on the basis of unit length, be arbitrarily classified into two major groups, the short interspersed repetitive elements (SINES) and long interspersed repetitive elements (LINES) (Singer, 1982). The major SINE in the human genome is the Alu element and the major LINE is the L1 element; both are present in vast numbers and together comprise over 10% of the human genome (see Schmied and Jelinek, 1982).

The Alu element. The most abundant SINE in primate genomes is the Alu element with over 500,000 copies dispersed around the human genome, an average of one element every 5kb (Houck et al., 1979; Jelinek et al., 1980). The Alu consensus sequence is approximately 300bp in length and is a dimer of two imperfect monomers, the left hand monomer containing a ~30bp insertion and separated from the right hand monomer by an A-rich spacer segment (Deininger et al., 1981). Each monomer shows homology to the 7SL RNA gene, from which it is assumed to derive, with the left hand monomer retaining the 7SL RNA polymerase III (pol III) promoter sequence (Ullu et al., 1982). Alu elements show several features suggestive of an RNA transpositional cycle, including a poly-A tail and target site duplication. These results, in addition to the
intragenic pol III promoter in the left hand monomer, lead to the proposal that Alu elements arise through transcription from the pol III promoter, reverse transcription of the transcript and insertion back into the genome at disperse loci, generating new loci, all capable of further rounds of transcription and insertion (Jagadeeswaran et al., 1981). Although the internal Alu pol III promoter sites seem active \textit{in vitro}, \textit{in vivo} transcription would appear to be very low or non-existent in HeLa cells (Paulson et al., 1986).

Alu elements have no known function and their dispersed and apparently non-conserved sequence nature makes it difficult for them to be viewed as selected under the same mechanisms as classical single locus genes. However, the observed patterns of dispersion and variation do not preclude an evolutionary role and various proposals have been put forward as to possible selected roles for Alu and/or other dispersed repeats (for review see Howard and Sakamoto, 1990).

Analysis of Alu element flanking DNAs has been used to suggest that Alu elements preferentially insert into short A/T-rich stretches, possibly mediated through homology with the Alu poly-A tail (Daniels and Deininger, 1985). In some cases it would appear that the poly-A tail of an already inserted Alu element has acted as a target site for subsequent Alu insertion (see Ma et al., 1991) and it has been suggested that the original Alu dimer may have been produced from an insertion of a monomer into the poly-A tail of an active monomer source gene (Quentin, 1992). In addition to very local sequence specificity in integration sites, a grosser level of Alu clustering has also been observed. Using \textit{in situ} hybridization of Alu probes in human metaphase chromosomes, it has been demonstrated that Alu repeats are preferentially localised in the Giemsa negative, or R-, bands (Korenberg and Rykowski, 1988). These results suggest a non-random basis for Alu integration.

The majority of Alus studied appear to be fixed in the human genome with relatively few insertional polymorphisms characterised. A limited number of polymorphisms have been identified, including dimorphisms in the tissue plasminogen activator gene (Friezner-Degen et al., 1986), the human C1 inhibitor locus (Stoppa-Lyonnet et al., 1990), the M1v1-2 locus (although this example may represent a rare variant or a \textit{de novo} somatic event, since it was observed only once in DNA from a B-cell lymphoma (Economou-Pachnis and Tsichlis, 1985)) and three other anonymous loci (Matera et al., 1990; Batzer et al., 1991). To date, only two examples of \textit{de novo} insertions causing genetic disease have been characterised: an insertion into the Factor IX gene in a case of haemophilia B (Vidaud et al., 1989); and an insertion into the neurofibromatosis gene (Wallace et al., 1991). One further example of an Alu insertion has also been partially characterised (Muratani et al., 1991). They have identified an acholinesterasaemic patient who is homozygous for an Alu insertion into the cholinesterase gene; whether this allele represents a polymorphism, rare variant or a recent event coupled with consanguinity remains to be determined. As mentioned in Chapter 1 (pages 2 and 3) internal sequence variation (at a rate equivalent to non-genic DNA, indicating a neutral mode of evolution at the level of the individual Alu) and variation in the length of the poly-A tail have been described, as well as a role for Alus in frequently deleterious non-homologous recombination.

Alu elements are found in the genomes of all primates, but not outside primates and are therefore assumed to have arisen before the primate radiation, some 65 million years ago and accumulated in the human genome since then (see Shen et al., 1991). Hwu et al. (1986) used hybridization techniques to predict the number of Alu elements present in current primate genomes and predicted the existence of around 900,000 elements in humans, compared to \~300,000 in Chimps and \~400,000 in Gorillas; indicating that large numbers of Alus have been inserted/deleted within primate genomes since the great ape split. However, these large interspecies differences do not agree with the results of many other groups, since the vast majority of Alus
studied are fixed in the human population and most are conserved between all the great apes (eg Sawada et al., 1985; Koop et al., 1986). Recent analysis of large numbers of Alu sequences has been used to show that Alus may be classified into a number of subgroups, the individual members of which are assumed to arise from distinct source genes. These results have lead to the current hypothesis that the majority of Alus derive from a limited number of source genes, only one or a few of which are active at any one time (see Shen et al., 1991; Leeflang et al., 1992). The major subgroups are shared amongst the great apes, suggesting that the major source genes were active before the great apes split and indicating that the majority of Alu insertion events are old (see Shen et al., 1991; Leeflang et al., 1992). Of the subgroups recently identified, only one family would appear to be largely human specific, the PV or HS subgroup and estimates at the number of copies of Alus falling within this group are put at around 500 to 2,500 (Matera et al., 1990; Batzer and Deininger, 1991). Significantly this group contains most of the known polymorphic Alus (all of them, except the dimorphic C1 Alu), as well as the de novo Alu insertion in the neurofibromatosis gene (see Leeflang et al., 1992). In addition, this group has been shown to be transcriptionally active in HeLa and Jurkat cell lines and a thyroid carcinoma (Matera et al., 1990). Thus it would appear that Alu elements are still transpositionally active in the human genome, but at a level which is at least an order of magnitude lower than has occurred previously in primate evolution (Shen et al., 1991; Leeflang et al., 1992).

The L1 element. L1 elements are ubiquitous to the genomes of all mammals and, although not as numerous as Alu repeats, L1 elements are generally significantly larger than Alus and comprise an approximately equivalent proportion of the human genome (Adams et al., 1981; Fanning and Singer, 1987). The full length L1 unit is about 6kb in length and present at around 20,000 copies per haploid genome, although an approximately 5 fold number of 5' truncated copies also exist (Adams et al., 1981). As with Alus, L1s show features suggestive of an RNA transpositional life cycle, including a poly-A tail and target site duplication, along with frequent 5' truncation (see Singer, 1982). The full length consensus L1 unit contains two open reading frames separated by a non-coding 14bp spacer, the second and larger open reading frame coding for a protein with a conserved domain showing high similarity to a number of reverse transcriptases, strongly suggesting that L1s are indeed retroposons (Hattori et al., 1986; Fanning and Singer, 1987). Both heterogeneous and discrete low level transcription of L1 sequences have been reported from lymphoblasts, lymphoblastoid cell lines and a variety of other cell lines (Kole et al., 1983; Schmeckpeper et al., 1984; Sun et al., 1984; Skowronski and Singer, 1985). However, in the pluripotent human embryonal carcinoma cell line NTer2D1, high levels of full length L1 transcripts have been detected (Skowronski and Singer, 1985; Skowronski et al., 1988).

Interestingly, L1s also show a similar preference toward integration at A/T-rich sites as seen for Alus (see Kazazian et al., 1988). However, in contrast to Alus, L1s are found preferentially localised in the Giemsa positive bands of human metaphase chromosomes (Korenberg and Rykowski, 1988). Nonetheless, these gross biases in Alu and L1 distribution are not absolute and both are sometimes found in the same regions (eg Ryan et al., 1991).

The majority of L1 elements characterised are conserved between man and other primates (see Scott et al., 1987), but again as with Alu sequences, subgroups of L1 elements have been identified, at least one of which may be human specific (Scott et al., 1987; Jurka, 1989; Hohjoh et al., 1990). Two de novo insertions of L1 elements into the same exon of the factor VIII gene causing haemophilia A have been characterised (Kazazian et al., 1988), along with a rare non-deleterious insertion in an intron of the same gene (Woods-
Samuels et al., 1989). The sequences of these newly inserted L1 elements, along with the majority of L1s cloned from NTera2D1 transcripts (Skowronske et al., 1988), place them in the proposed newest class of active L1 elements (Woods-Samuels et al., 1989; Jurka, 1989; Hohjoh et al., 1990). Use of an oligonucleotide probe specific to one of the newly inserted L1s, on genomic DNA of unrelated individuals revealed up to 10 bands, most of which were polymorphic (Dombroski et al., 1989). All these results indicate that L1 transposition is still an active process within the human genome.

**Other dispersed repeats in the human genome.** Several other less abundant classes of dispersed repeats have also been partially characterised, including transposon like human elements (THE) (Paulson et al., 1985), MER elements I-VI (Jurka, 1990) and retroviral like elements (Mager and Heninhorn, 1984). Of these, THE are the most numerous, at around 10,000 copies per genome and are comprised of a unique internal element of ~1.6kb, flanked by two long terminal repeats (LTRs) (Paulson et al., 1985). The THE LTRs frequently exist (~30,000 copies) as solitary units, previously termed O repeats (Sun et al., 1984). A number of retroviral like elements (RTVLs) have been identified (see Maeda and Kim, 1990), but the most common of these (RTVL-H) exists at only about 1,000 copies per genome (Mager and Heninhorn, 1984). None of these additional repeats have been shown to be polymorphic in humans and no recent insertion events have been observed.

**Minisatellites and dispersed repeats.** Sequence analysis of the DNA flanking a number of hypervariable human minisatellites has revealed a higher than expected frequency of dispersed repeats (Armour et al., 1989b). In less than 6.5kb of flanking DNA sequenced, seven dispersed repeats have been identified (four Alus, one L1, one RTVL LTR, plus one other, apparently very old, degenerate dispersed repeat (see Donehower et al., 1989; Armour et al., 1989b)). The general direction of this relationship is not clear, although in at least one case it can be seen that the dispersed repeat predates the minisatellite expansion; MS32 has expanded from within a retroviral LTR (Armour et al., 1989b). Furthermore, recent sequence analysis of eleven random clones containing GGAA tetranucleotide repeats has shown that seven of these were derived from the expansion of an Alu poly-A tail and one from within an L1 element (Gray and Jeffreys, unpublished data). Whether the clustering of dispersed repeats and tandem arrays has interdependent functional relevance to the generation of either remains an open question. Alternatively their association may merely reflect their occurrence in non-genic regions over which selective pressures are minimal.

**Estimation of the rate of transposition in current human genomes.** From present data it would appear that Alus and L1s are still the most active transposable elements in the human genome, although neither seems to be as mobile as at previous points in primate evolution. Both have approximately equivalent levels of insertional polymorphism and characterised de novo insertion events and both would appear to be operating at roughly equivalent levels. Using some of the data available from studies of Alu elements it is possible to make a very crude estimate as to their current rate of transposition in the germline. If the approximately 1,500 human specific Alu elements (Matera et al., 1990; Batzer and Deininger, 1991) have inserted consistently over the past ~6 million years since the great ape split and assuming a generation time of 20 years then we can predict that Alus are inserting at about 5x10^-5 per genome per generation. If we also assume that integration is random within the 3x10^9bp of the haploid genome then we arrive at a figure of ~2x10^-9 per kb per genome per generation (this calculation assumes a neutral mode of mutation for Alu insertion, under which the mutation rate per gamete equals the fixation rate per generation). Thus we can see that the predicted rate of transposition is very low.
Alternatively, we may use the figures of Hwu et al. (1986), who suggest that ~500,000 Alus have inserted since the great ape split (although it seems highly likely that their data is an overestimate). Using this figure and those used above we arrive at a predicted rate of ~6x10^-7 per kb per genome per generation.

This work. Transposable elements comprise over 10% of the human genome and are involved in genetic disease as both insertional mutagens and mediators of non-homologous recombination. Yet despite their genomic abundance, relatively little is known about the mechanisms of transposition, the nature of currently mobile elements, their source genes, target site preferences or accurate rates of transposition. This is largely due to the relatively low rate of transposition, which has previously thwarted the analysis of large numbers of transposition events through traditional pedigree approaches. The recent advent of PCR though, has opened a new direction in mutation analysis, allowing the recovery of very rare events from bulk genomic DNA. In this chapter I describe attempts to develop a system for measuring and characterising de novo transposition events at a defined locus (DNA flanking a hypervariable minisatellite) in the human genome.

Results

Strategy. Previously, Jeffreys et al. (1990) have used physical size selection and PCR recovery to isolate very rare large deletion mutants of a hypervariable minisatellite from bulk genomic DNA. Briefly, they took bulk genomic DNA from an individual with known allele sizes. They digested this DNA with restriction enzymes that cut the flanking DNA of the minisatellite, releasing the tandem repeat array. Size fractions smaller than the known progenitor alleles of this digested DNA were then prepared using agarose gel electrophoresis. After multiple rounds of agarose size selection all of the progenitor alleles were removed, leaving only rare deletion mutants in the smaller size fractions. Rare deletion mutants were then recovered from these pools using single molecule PCR. This system was shown to be highly efficient and capable of enriching for mutants at least as rare as 1 in 10^6 (Jeffreys et al., 1990). Using an analogous approach, I hoped to develop a system capable of detecting not size decreases, but size increases at a defined locus. Thus the basic strategy is to PCR amplify DNA from a defined locus and size separate the products by agarose gel electrophoresis. Electroelution is then performed to enrich for products larger than the expected size, followed by further amplification, until rare size increase mutants are detected. Due to the 'streaking back' phenomenon of agarose gel electrophoresis (ie the aberrant migration of a fraction of the total DNA molecules at a higher weight than expected) it is anticipated that larger size fractions will be heavily contaminated with the 'normal allele' and that multiple rounds of enrichment and amplification will be necessary. Using this strategy at a defined unique sequence locus, it is hoped to isolate rare insertion events, as well as rare duplication events (although, if such duplication events are detected it will prove difficult to distinguish genuine de novo duplications, present in the original genomic DNA sample, from PCR artifacts generated during amplification).

The locus chosen. The region of DNA chosen to attempt this procedure was ~1.2kb of unique sequence DNA flanking the 5' end of the hypervariable human minisatellite MS1 (Wong et al., 1987). The reasons for choosing this locus were several fold, but ultimately somewhat arbitrary, including: the locus incorporates ~1.2kb of known unique sequence DNA, allowing design of locus specific PCR primers; in view of the association between hypervariable minisatellites and dispersed repeats (MS1 already has an Alu element in the 3' flanking DNA), it may be that regions flanking hypervariable minisatellites are prone to insertion; and the presence of a number of short A/T-rich potential integration sites (the 1.2kb region incorporates 8 independent 10bp windows with at least 90% A/T).
**PCR amplification of the MSI 5' flanking region.** Four PCR primers were designed and synthesized to produce a nested set of primers capable of amplifying ~1.2kb of DNA flanking the 5' side of MS1 (see Figure 8.1). The internal nested primers MS1-3 and MS1-4 incorporate 15bp 5' extensions containing restriction enzyme sites (BamHI, EcoRI, XmaI and SmaI) which should simplify cloning of amplification products, if necessary. Such a nested system should allow the efficient amplification of single molecules, up to levels visible on ethidium bromide stained agarose gels (Jeffreys et al., 1990). The efficacy of single molecule PCR was determined using diluted genomic DNA samples. 39 'single molecule' reactions were prepared using 6pg input genomic DNA (ie ~two haploid genome equivalents) and PCR amplification with the outside primers MS1-1 and MS1-2. Amplification for 30 cycles was performed and the products detected by Southern blot hybridization (Figure 8.2A). Out of 39 reactions, 31 gave products (14 zero DNA controls gave no product) giving an amplification efficiency of 79% per molecule (assuming a Poisson distribution). This experiment was repeated, but rather than detection by blot hybridization, the products of the first round of amplification were used as seed DNA in a nested amplification with primers MS1-3 and MS1-4. After a further 30 cycles of amplification, PCR products derived from single molecules were visible on ethidium bromide stained agarose gels (Figure 8.2B).

**Insertion analysis of sperm DNA.** In order to determine the rate of transposition into our chosen locus, bulk samples of human sperm DNA (ie human germline DNA) were processed using the system outlined above. Five 10μg samples of sperm DNA from one individual were amplified in 100μl reactions using the outside primers MS1-1 and MS1-2, for 12 cycles. The number of cycles used in the reaction was kept to a minimum (to give PCR products below levels visible on ethidium stained agarose gels, ie <1ng) for two reasons: firstly, to ensure that the normal allele (present in vast excess) would not out-compete larger, much rarer, mutant molecules (at low cycle number even large fragments should amplify with reasonable efficiency (Jeffreys et al., 1988a)); and secondly, to prevent product dimerization, which can occur at high product concentration and which would present as artifactual insertion events. Control reactions were also performed, including zero DNA controls, to check for contamination of the reagents, positive controls (using gel purified samples of previously amplified material), to check the fidelity of the amplification process and positive doped samples (ie reactions containing equivalent amounts of DNA as the test reactions and doped with a positive sample), to check against PCR inhibition of the test sample (very high DNA concentrations, as used in the first round of amplification in this procedure and/or gel purified DNA fractions, as used in the latter stages, can sometimes inhibit PCR amplification). After amplification, each sample was doped with 500ng of pAT153-Ddel/BamHI marker DNA, loaded into 25mm wide wells on a 0.8% 20cm agarose gel in 1xTAE and electrophoresed at 30V for 16 hours, alongside marker lanes containing 100ng of pAT153-Ddel/BamHI and λHindIII marker DNAs. After electrophoresis, the marker lanes and an approximately 1mm wide slice of each sample lane, was excised from the gel and stained with ethidium bromide (1μgml⁻¹ in 1xTAE). Sample lanes were not stained directly since UV illumination produces thymidine dimer formation in DNA which prevents subsequent amplification. The positions of the relevant markers were cut into the gel slice (under UV illumination), realigned with the sample lanes and the required size fraction from the sample lane excised. A gel fraction from ~1.4kb to 9.5kb was purified by electroelution onto dialysis membrane (see Chapter 2), ethanol precipitated and dissolved in 10μl of 5mM Tris-HCl (pH7.5). The internal pAT153-Ddel/BamHI marker DNA has an upper fragment size of 1,205bp, compared to the 1,195bp of the normal MS1-1 to MS1-2 product, allowing accurate sizing of the gel fraction required and minimizing contamination with the normal allele. The
Figure 8.1. The MS1 locus and the position of the PCR primers.

PCR primers are shown (arrows), along with the minisatellite tandem array (hatched boxes). Primer sequences are: MS1-1, 5'-GCTGCTGTGCTTTCCGGGTTTC-3'; MS1-2, 5'-ACCCACTGACTCACAGCAAGGG-3'; MS1-3, 5'-GGCGGATCCCGGGAATTCAGTCTAGTGC-3'; and MS1-4, 5'-GCGAATTCTCGGATCCCGGTTTCTCTTATTTGCTGG-3'. Primers MS1-3 and MS1-4 contain 15bp 5' extensions (underlined) incorporating restriction endonuclease sites (BamHI, EcoRI, SmaI and Xmal).

Figure 8.2. Single molecule PCR of the MS1 5' flanking region.

(A) Single molecules recovered by PCR and detected by hybridization. Total genomic sperm DNA was diluted to 6pg/μl in 0.1μM PCR primers (MS1-1) as carrier. Nineteen 10μl reactions with 6pg of input DNA, plus 4 zero DNA controls (0) and a positive control (P) with 300pg of input DNA were amplified for 30 cycles with the outside primers MS1-1 and MS1-2 with an annealing temperature of 63°C and an extension time of 4 minutes. 5μl of each product was electrophoresed on a 1% agarose gel in 1xTAE and Southern blot hybridized with a flanking sequence probe (probe used was gel purified MS1-3 to MS1-4 amplification product). Positive (+) and negative (-) reactions are showns.

(B) Single molecules recovered by PCR and detected by visualisation on agarose gels. Single molecule reactions were amplified to hybridizable levels as above using primers MS1-1 and MS1-2. 0.1μl of these reactions were used to seed second round amplifications using the internal nested primers MS1-3 and MS1-4 for a further 30 cycles of amplification with an annealing temperature of 63°C and an extension time of 4 minutes. 5μl was removed and electrophoresed on a 1% agarose gel in 1xTAE and visualised by ethidium bromide staining. Positive (+) and negative (-) reactions are shown.
size fraction taken should enrich for alleles containing insertions of between 200 and 7,000bp (the consensus Alu and L1s have unit lengths of ~300 and 6,000bp respectively). The level of purification obtained was determined by electrophoresis of a 1μl sample of the purified fraction alongside a marker dilution series and subsequent Southern blot hybridization with a pAT153-DdeI/BamHI probe. Densitometric analysis of the 1.2kb band on the resultant autoradiograph demonstrated enrichment consistently at a level of between 100 to 500 fold (data not shown). A further 1μl sample of the gel purified fraction was used in a cycle titration test, ie a 20μl reaction was set up and 3μl samples removed after 10, 12, 14, 16, 18 and 20 cycles and the products electrophoresed and visualised on an ethidium stained agarose gel. From this analysis the number of cycles required to amplify the normal allele to sub nanogram levels could be determined (generally 12-16 cycles, data not shown). 5μl of the remaining gel purified fraction was then used in a further rounds of amplification (using the empirically derived number of cycles) and size fractionation, as above. For each sample this procedure was repeated two more times with the outside primers and four times with the internal nested primers. No insertion products were detected, either by visualisation on ethidium stained agarose gels or by Southern blot hybridization using a total genomic DNA probe (such a probe should detect any element represented at least 100-1000 times in the human genome ie should detect Alus, L1s, THes and possibly a few other dispersed repeats (Shen and Maniatis, 1980)). Thus it would appear that the samples analysed did not contain insertion events at the locus examined at a level detectable using the system developed.

Recovery from bulk genomic DNA of a rare pseudoinserion allele. In order to determine the minimal level at which rare insertion events could be recovered from bulk DNA, a pseudoinserion allele was generated and used in recovery tests. The 1.8kb pseudoinserion allele was generated by inserting the 657bp NarI fragment of the Tet gene of pBR322 into an MS1-3 to MS1-4 product using the 'PCR cloning' strategy shown (Figure 8.3). The pseudoinserion allele was generated as an MS1-3 to MS1-4 product so that it could not contaminate MS1-1 to MS1-2 reactions, as used for genuine insertion analysis. This allele was diluted to the single molecule level and shown to be single molecule PCR amplifiable (data not shown). The pseudoinserion allele was next used in a doping experiment to test recovery from bulk genomic DNA. Four 100μl reactions each containing 10μg of sperm DNA were set up as before, except two were doped with 1μl of a 2 molecule per μl dilution of the pseudoinserion allele (95% probability that each reaction contains less than 5 molecules) and amplified with the nested primers MS1-3 and MS1-4. The amplification products were size fractionated and gel purified as before and a second round of amplification and size fractionation performed. The products of the second round purification were used in cycle titration test and the products visualised by ethidium staining of the agarose gel and Southern blot hybridization with a pseudoinserion specific probe (Figure 8.4). The pseudoinserion allele was successfully recovered from the single molecule level to hybridization detectable levels from bulk genomic DNA (~3.3x10^6 normal alleles). No pseudoinserion specific products were detected in the undoped control reactions.

Insertion analysis of blood, placental and tumour DNA. Since no insertion events were detected in germline sperm DNA, but the system was proven capable of detecting relatively rare events, it was decided to use a variety of samples for insertion analysis. For transposons to become stably integrated into the human germline, they must be active either during early embryogenesis, before separation of the germline, directly in the germline itself or continuously throughout all somatic and germline development. The fact that high levels of retroposon transcripts are found in human embryonal carcinoma cell lines suggests that the former may be the case. No evidence exists to suggest whether transposons are active in normal human somatic tissue.
Figure 8.3. 'PCR cloning' strategy used to create the MS1 pseudoinserstion allele.

1μg of pBR322 was digested with NarI and the 657bp Tet fragment gel purified. 500ng of gel purified MS1-3 to MS1-4 product was digested with TaqI, which produces the same sticky end as NarI and both fragments gel purified. To prevent self ligation, the 832bp fragment was dephosphorylated with calcium alkaline phosphatase and once again gel purified. 20ng of all three fragments were ligated, in the presence of NarI, to prevent multimeric insertion and the 1,859bp pseudoinserstion allele recovered by PCR amplification (10 cycles of a standard MS1-3 to MS1-4 amplification, see Figure 8.2) and gel purification.
Figure 8.4. Recovery from bulk genomic DNA of a rare pseudointerstion allele.

(A) Ethidium bromide stained gel of pseudointerstion recovery.
(B) Southern blot hybridization of pseudointerstion recovery with a pseudointerstion specific probe.

Duplicate samples of 10μg of sperm DNA, two normal (S1 and S2) and two doped with 1μl of a 2 molecule per μl dilution of the pseudointerstion allele (I1 and I2) were amplified for 12 cycles with the PCR primers MS1-3 and MS1-4 and size fractionated as for the normal insertion analysis (described in the text). This was repeated for a second round of amplification and size fractionation, as before. One quarter of the second round purification products were used to seed a further 20μl amplification and 3μl samples taken after 8, 10, 12, 16, 20, 25 and 30 cycles. Products were electrophoresed on a 1% agarose gel in 1xTAE for 16 hours at 30V and the DNA visualised by (A) ethidium bromide staining and (B) by Southern blot hybridization with a pseudointerstion specific probe (the pBR322 Tet fragment used in its genesis, see Figure 8.3). Positions of the normal allele (N) and the expected pseudointerstion allele (I) are shown. The smaller insertion specific products detected in the later amplifications (B) are presumed N/I heteroduplexes and single stranded DNA.
or not. The Mvli-2 Alu insertion, identified in a B-cell lymphoma, may represent a rare germline variant, a rare somatic mutant amplified by tumour expansion or alternatively insertion may even have occurred during tumour expansion. Tumour progression is accompanied by frequent gross chromosomal rearrangements operating on a positive feedback loop as cell cycle control decays (Holliday, 1989) and it seems possible that under such chaotic conditions transposons may become activated. Previously Armour et al. (1989a) have shown somatic length instability of minisatellite loci as detected by analysing tumour DNA, finding a high level of mutant bands in colonic tumours, although it remains unclear as to whether mutant bands are generated during, or merely amplified by, tumour expansion. Human placent material is derived directly from embryonic tissue, is freely available and may also represent a potential source for identifying de novo transposition. Thus an additional seven sources of DNA were obtained for insertion analysis: two blood DNAs (a kind gift of A.J. Jeffreys), two colonic tumour DNAs (a kind gift of J.A.L. Armour), two placent DNAs (a kind gift of Z. Wong) and one additional sperm DNA (a kind gift of J.A.L. Armour). 10μg of each sample was analysed for insertion events using the same procedure as for the sperm DNA samples described above. No insertion events were detected in any of the samples.

Discussion

The current rate of transposition within the human genome remains an unknown factor, although approximate estimates put the level very low, probably between 10^-9 and 10^-6 per kb per genome per generation. These figures are calculated assuming random sites of integration, an assumption we know is incorrect, and values for individual loci may vary considerably from this, although which loci and, by how much, still remains unknown. During this work a system was developed for identifying rare de novo insertion events at a defined locus using physical size selection and PCR recovery of single molecules. The locus chosen as a potential target site for transposition was a 1.2kb fragment of unique sequence DNA flanking the hypervariable minisatellite MS1. Using an artificial pseudoinsertion mutant allele the system was shown to be very sensitive, capable of detecting single molecule 'mutants' from a background of >10^6 normal alleles. 50μg of human sperm DNA (1.7x10^7 genomes) was analysed for insertion events using this strategy, but no insertion events were detected. These results enable us to put an upper limit on the rate of de novo transposition into this locus of <2x10^-7 in the sample analysed. Further samples from a range of sources, including blood, tumour and placental DNA, were analysed in the same way and once again, no insertion events were detected. In all of these tissues, the upper limit on the rate of de novo transposition into this locus is less than 1x10^-6 per cell. We can therefore conclude that the rate of transposition into our chosen locus is very low, certainly less than 1 in 10^-6. Whether this locus represents an accurate reflection of the rest of the genome remains to be determined. However, J.A.L. Armour (personal communication) has applied the same technique to another locus with similarly negative results. The locus chosen in this case was the factor VIII gene, a known insertional target in recent human evolution (see introduction to this chapter). Thus it seems likely that, as expected, rates of transposition in the human genome are indeed very low, although higher rates at other, as yet undetermined, loci or in other source DNAs cannot be ruled out.

The system developed here for detecting transposition events is very sensitive, capable of detecting single molecule mutants from a high background of normal alleles. Nonetheless, it would appear that in vivo transposition is extremely rare, such that relatively large amounts (maybe as much as 1mg) of input DNA will be required to detect even single molecule mutants. The system employed here is already operating at a maximal level, starting with 100μl reactions and 10μg of input DNA; increasing reaction volumes reduces amplification...
efficiency, as does increasing the DNA concentration. Even if the amount of input DNA can be increased, it may still not be enough to achieve the 100 fold gain that may be required if the rate is as low as $10^{-9}$ per kb per genome per generation. Obviously an increase in input DNA could also be achieved by using multiple reactions, although this approach is already very labour intensive and a significant increase in sample throughput is unrealistic. An alternative approach is to find source DNAs in which the rate of transposition is elevated above that found in normal human tissue, at levels realistically detectable using the system developed here.

As mentioned above, the human embryonal carcinoma cell line NTERa2D1 has reported high levels of retroposon transcription. This cell line is derived from an embryonal carcinoma and is comprised of undifferentiated pluripotent cells, thought to be reflective of very early embryonic tissue (Andrews, 1984). Thus, if retroposon activity is limited to early embryogenesis, as seems likely, then this cell line may represent a good source tissue. This prediction is strengthened by the recent finding that high levels of reverse transcriptase activity can be isolated from this cell line, presumed to derive from translation of the L1 ORF (Deragon et al., 1990) and it would seem reasonable to assume that retroposon activity may be elevated in this cell line. Obviously this cell line is not normal human tissue and may not reflect accurately the transpositional processes operating in normal embryonal development. Nonetheless, if de novo transposition could be detected in such a cell line then it would represent a significant advance in studies of human transposon biology. At the very least, active elements, their source genes, preferred integration sites etc could be determined. (Unfortunately, lack of time has not yet permitted application of the system developed here to such a source DNA, but hopefully this will be attempted at a later date.)
Chapter 9

DISCUSSION

Minisatellite internal variation, allelic diversity and pseudohomozygosity. Minisatellite loci often show very high levels of variation in the number of repeats at a given locus with resulting high allele length heterozygosity. Minisatellite allele length analysis has been, and continues to be, a major investigative tool in many areas of genetic analysis including linkage mapping, kinship testing and individual identification (see Chapter 1). However, due to the vagaries of agarose gel electrophoresis, the necessary assumptions involved in calculating match significance and the heated scientific debate accompanying population statistical analyses, some of the practical limitations in the application of minisatellite length analysis to forensic medicine have recently been highlighted (see Chapter 1). In particular, the apparent homozygote excess has been used to suggest that a significant level of population substructuring exists, impeding the application of the multiplication rule to calculating DNA profile match frequencies (see Chapters 1 and 3).

Variation in allele length is not though the only source of minisatellite variation; all human hypervariable minisatellite loci which have been extensively characterised show additional variation in the dispersion patterns of MVRs. This second level of variability has been previously accessed by direct sequence analysis and by methods of internal mapping based on nucleotide substitutions that affect restriction endonuclease sites (Owerbach and Aagard, 1984; Jeffreys et al., 1990; Gray and Jeffreys, 1991). For the most well characterised human minisatellite MS32, MVR repeat mapping was originally developed by utilizing variant repeats either cut or not cut by the enzyme HaeIII (Jeffreys et al., 1990). This restriction mapping technique has been used to show that the true level of allelic variability at MS32 is substantially greater than measurable by length analysis. Standard Southern blot length analysis yields a maximum of around 100 distinguishable alleles, whereas internal mapping has in theory the capacity to discriminate between many millions of alleles. In addition, internal mapping provides unambiguous digital information, defining each allele as a discrete entity. Formerly, HaeIII restriction mapping at MS32 was used to show that alleles of the same length shared between unrelated individuals often have different internal structures. In this study these results were extended by demonstrating that two individuals identified as apparent allele length homozygotes were in fact true heterozygotes, based on small differences in allele length and/or internal structure (Chapter 3).

This type of mapping, although very informative, does require detailed sequence information to develop the methods for assaying internal variation. Furthermore, not all loci are amenable to the types of analysis employed at MS32, due to either lack of suitable restriction site variability or conversely too much internal variation (Jeffreys et al., personal communication). In contrast, the SSMP approach to detecting internal variation between minisatellite isoalleles requires no information concerning the sequence basis of variation (Chapter 4). SSMP analysis has been successfully applied to all minisatellite loci so far investigated and may even be applied to very large alleles (too large to be PCR amplified), and to loci for which flanking DNA sequence is not available (and hence not PCR amplifiable), through a very simple SSMP procedure using restriction digested total genomic DNA (Chapter 4). Application of SSMP to a number of apparent length
homozygotes has revealed that a significant proportion of these individuals were indeed true heterozygotes. Nonetheless, although SSMP analysis is very general in terms of the loci to which it may be applied, it is very much a practical tool, applicable mainly to the direct analysis of minisatellite isoallelism, in particular, to pairs of isoalleles of specific interest (e.g., to increase the information derived from pseudohomozygotes and/or shared alleles in pedigree based linkage analysis), rather than as a direct measure of minisatellite internal variation, as provided by MVR mapping.

The application of the improved MVR-PCR approach (Jeffreys et al., 1991b) to the internal mapping of single alleles, through both pedigree analysis (Chapter 6) and knockout MVR-PCR (Chapters 5 and 6), has enabled us to extend considerably the data concerning allelic diversity at MS32. Allelic diversity in two major populations has been determined and shown to be very high (corresponding conservatively to 2,500 equally rare alleles in both Caucasian and Japanese populations), with a predicted worldwide diversity of more than $10^8$ different alleles (Jeffreys et al., 1991b; Chapter 6). Once again, the application of internal mapping to more apparent allele length isoalleles has shown that the majority of such individuals are true heterozygotes for identical or closely sized alleles.

The demonstration that the majority of apparent allele length homozygotes (Chapters 3, 4 and 6) are true heterozygotes, reflects the quasicontinuous nature of minisatellite allele length distributions and the convergent nature of minisatellite allele length evolution. These data strengthen the argument that the observed homozygote excess is probably an artifact of agarose gel electrophoresis, caused by the coalescence of closely sized alleles into a single band and the occurrence of small null alleles not detected by conventional Southern blot analysis (Devlin et al., 1990; Armour et al., 1992; Chakraborty and Jin, 1992).

MVR-PCR and individual identification. The MVR-PCR system (Jeffreys et al., 1991b) developed for the internal mapping of MS32 alleles can be applied directly to total genomic DNA to reveal the diploid code derived from the two superimposed single allele maps. As a result of the observed massive allelic variation, MVR diploid codes are highly individual specific and provide an ideal system for personal identification, with obvious potential applications in forensic medicine. The codes generated are in a digital format, ideal for computer based storage, analysis and the dissemination of data and databases between laboratories. In addition, diploid codes provide absolute match criteria, for which match significance can be determined directly from observed code frequencies, with no reliance on assumptions of Hardy-Weinberg equilibrium. MVR-PCR is immune to many of the problems associated with agarose gel electrophoresis such as band shift and overcomes many of the limitations associated with conventional allele length analysis. The development of knockout MVR-PCR (Chapter 5), which allows highly discriminatory codes to be derived from some mixed DNA samples and from admixture ratios far lower than achievable with conventional systems, further extends the number of cases to which MVR-PCR may find applications in forensic medicine.

No two unrelated individuals sharing diploid codes have been identified (over 500 unrelated individuals typed at MS32). Since, however, coding is based on a single locus, siblings stand a one in four chance of having the same diploid code. Thus a battery of MVR-PCR systems operating at disperse loci would be required to enable distinction between siblings and other highly related individuals. Currently, MVR-PCR has been successfully applied to MS32, although a few other loci are presently under development (Neil et al., personal communication; Armour et al., personal communication). The development of suitable systems is hampered by the paucity of cloned loci that fulfil the necessary requirements for the successful application of diploid MVR-PCR, i.e., loci that have: high allelic variability (>95% allele length heterozygosity); suitable internal variation,
allowing PCR definition of two (and preferably two only) variant repeat units; and, the criterion by which most loci fail, constant repeat unit length (abnormal length repeat units will put single allele codes out of register with each other, preventing diploid code analysis). One additional intriguing possibility is the development of an MVR-PCR system for a Y-chromosome specific minisatellite; such a locus would have obvious applications in forensic biology and could prove extremely informative in the analysis of male genealogies. An effectively haploid system such as the Y-chromosome offers the potential to analyse minisatellite mutation in the absence of an homologous chromosome, as well as removing the constant repeat unit length requirement for the successful application of MVR-PCR. Thus far, only one Y-chromosome specific hypervariable minisatellite has been identified and its potential analysis through MVR-PCR is being actively pursued (M. Jobling, personal communication).

MVR-PCR and population analysis. With the high mutation rates, high variability, ease of allelic definition, the ability to identify groups of related alleles and the relative ease of obtaining large amounts of data, the potential for the analysis of recent population divergence using MVR-PCR is obvious. The application of such techniques in studies of population genetic diversity could allow the derivation of nuclear autosomal lineages, as previously produced for maternal lineages using mitochondrial DNA polymorphisms (Cann et al., 1987). These studies would be further complemented by the incorporation of a Y-specific system that would provide a paternal lineage, analogous to the mitochondrial derived maternal line.

Minisatellite mutation processes. The massive allelic variation observed at some minisatellite loci is ultimately based on the mutation of progenitor alleles to form new variant alleles. Mutation rates at some loci are very high and germline de novo mutation to new length alleles has been detected by standard Southern blot based minisatellite length analysis in large families (Jeffreys et al., 1988b; Wolff et al., 1988; Vergnaud et al., 1991). MVR-PCR pedigree analysis has been used here to detect and characterise further germline de novo mutation events at MS32, identifying an increased set of mutants involving size changes too small to be detected by Southern blot length analysis (Jeffreys et al., 1991b; Chapter 7). Internal mapping of these mutants and of mutants characterised at two other minisatellite loci has revealed interesting new insights into minisatellite biology, with for the first time, data indicating a role for unequal interallelic exchange in minisatellite mutation (Jeffreys et al., 1991b; Chapter 7). A role for recombination and/or conversion in minisatellite mutagenesis is supported by haplotypic analysis of polymorphisms flanking the minisatellite array, which show an elevated level of recombination compared to the genomic average (Chapters 6 and 7).

A general mutation process at MS32 would appear to involve small additions of repeat units to the 5' terminus of alleles, from either a sister chromatid or homologous chromosome donor, with no loss of information in the recipient. The terminal (unidirectional) polarity in mutation parallels the previously observed polarity in allelic variation and would appear to be a general phenomenon having now been observed at two other loci (Jeffreys et al., 1990; Jeffreys et al., 1991b; Chapters 6 and 7; Armour et al., manuscript in preparation; Neil et al., personal communication). A subset of the mutation events can be explained by simple unequal sister chromatid exchange or replication slippage, although these events appear simple and do not contain enough information to exclude an interallelic origin. However, at least three events not interpretable using such mechanisms have been identified. A speculative model for unequal recombination and/or conversion has been proposed, based on the DSBR model and involving initiation by a DSB, which can be used to describe all of the observed mutation events (although two repeats in one mutant still have no obvious origin), as well as accommodating the inferred recombinational and/or conversional domain between the minisatellite and flanking
markers. Terminal polarity is an intriguing and unexpected phenomenon and strongly suggests a role for cis acting sequence elements, although possible orientation effects relative to chromosome ends and/or replication origins cannot be ruled out. Relatively few minisatellites have been orientated within large scale physical maps and MS32 is not one of them. A cis acting element directing DSBs to the terminal repeat remains an attractive hypothesis and could also account for the apparent size increase bias (7/7 mutation events at MS32 were size increases, Chapter 7). Previously it has been suggested that minisatellites arise and expand somewhat stochastically, via mechanisms of equal gains and losses, and are probably transient in evolution (Gray and Jeffreys, 1991). If however, a true size increase bias exists then minisatellite amplification (through terminal gap expansion or some other mechanism biased toward size increase) could represent another, as yet unexpected, fuel in the engine of molecular drive (Dover, 1982). Nonetheless, even with a size increase bias catastrophic collapse to a single repeat, and presumably stable, ground state still remains possible and we may yet expect minisatellites to have stochastically limited genomic life-spans.

Although several of the mutants characterised have features consistent with generation through replication slippage, several lines of evidence suggest that replication slippage is probably not a major mechanism of minisatellite mutation. Firstly, as an amplification process it might be expected that the number of mutations would be dependent on the number of mitoses. Despite far more mitoses in male gametogenesis, no increase in male germline mutation rate, over female, is observed at MS32. Nonetheless, other loci do show extreme parental bias in mutation rate, eg the CEB1 locus (15% paternal mutation rate, 0.3% maternal mutation rate) and a slight, but significant, bias (60:40, male to female) has been observed in SLP parentage analysis performed at multiple loci (mainly MS1). Secondly, it is difficult to conceive how terminal polarity could be incorporated into a model based on replication slippage. A directional bias in relation to the replication origin could be invoked, but under such a scenario we might expect a distance mediated effect (ie an increasing mutation rate the further the repeat from the replication origin), increasing mutation rate with allele length, rather than the observed apparently allele length independent terminal variation. Thirdly, in vitro studies of replication slippage suggest that A/T-rich sequences are more likely to slip than G/C-rich sequences, in direct contrast to the observed bias in high G/C content at minisatellite loci (Schlotterer and Tautz, 1992). In addition, it was observed that triplet repeats had lower slippage rates than dinucleotide repeats, suggesting that slippage decreases with repeat unit length (Schlotterer and Tautz, 1992). Since such microsatellites generally have very low mutation rates (<0.05% per gamete, Kwiatkowski et al., 1992), it seems unlikely that slippage could generate the very high mutation rates observed at minisatellite loci with longer repeats. Finally, in vivo studies in yeast suggest, if anything, a bias toward, presumably slippage mediated, deletion rather than a size increase bias at di- and tri-nucleotide repeats (Henderson and Petes, 1992).

If minisatellite instability is a result of cis acting sequence elements then the possibility exists for the loss, through deletion, of this element with a presumed cessation in mutation of the linked allele. If we also assume that, as seems likely (see Szostak et al., 1983), the initiating chromosome is always the recipient of information during conversion events, then such a deleted allele (and hence initiation deficient) could only ever donate its deleted initiator and never be converted back to initiation proficiency. Such a scenario may explain the existence of monomorphic minisatellites and other minisatellite loci which have common alleles. Unfortunately, virtually nothing is known of the biology of monomorphic minisatellites.

Other factors possibly involved in minisatellite mutation directly related to the minisatellite itself include repeat unit sequence, repeat unit length and total allele length, in terms of both physical size and number
of repeat units. As discussed previously (Chapter 7), the relevance of the core sequence remains in doubt, although the G/C bias observed does suggest some sequence contribution. The demonstration that \textit{in vitro} slippage rates are sequence specific shows that sequence can have an effect on tandem repeat instability, although how this relates to minisatellite instability is presently unclear. One possibility is that certain sequences are stabilized or destabilized by specific interactions with proteins, and in support of this viewpoint is data identifying proteins that interact with G-rich tandem repeat sequences (Collick and Jeffreys, 1990; Wahls \textit{et al.}, 1991; Collick \textit{et al.}, 1991) and C-rich tandem repeat sequences (Yamazaki \textit{et al.}, 1992). None of these proteins have been sequenced and their functions and/or relevance to minisatellite biology, if any, remain elusive.

Evidence obtained from mutations detected by Southern blot length analysis has revealed no obvious correlation between allele length and mutation rates (Jeffreys \textit{et al.}, 1988b; Vergnaud \textit{et al.}, 1991) and it would seem unlikely that mechanisms of terminal mutation should be dependent on allele length, assuming a minimum number of repeats are present. It has been suggested that large deletion rates may increase with physical allele size (Chapter 7), but present data relating to rare large deletion events is limited and this hypothesis remains speculative.

With so few minisatellites characterised, and even fewer mutation events analysed, the relevance of repeat unit length remains obscure. Nonetheless, the demonstration that \textit{in vitro} slippage rates are repeat unit length dependent (Schlotterer and Tautz, 1992) suggests that repeat unit length can have a role in tandem repeat instability. As expected, dinucleotide repeats appeared more mutable than triplet repeats, presumably reflecting an increased number of target sites and a reduced diffusion distance required, for misalignment of complementary strands in tandem repeat arrays. Interestingly, the most variable human minisatellite, MS1, has a very short repeat unit length of only 9bp (Wong \textit{et al.}, 1987) and it is feasible that short repeats may be more unstable than long repeats. Furthermore, the two most unstable mice minisatellites cloned to date (Ms6-Hm and Hm-2) have repeat lengths of 5 and 4bp (Kelly \textit{et al.}, 1989; Kelly \textit{et al.}, 1991; Gibbs \textit{et al.}, 1991). Moreover, the highly unstable repeats in fragile-X and DM (see Chapter 1, page 7) are trinucleotides, although their relevance to general minisatellites remains unclear at present.

Even if cis acting factors are not an absolute requirement of minisatellite instability, it seems likely that surrounding sequence and/or general genomic location may have some sort of a role to play. The subtelomeric clustering of minisatellites can be explained by assuming that minisatellites are involved in the recombination process (either as mediators or products), but this does not \textit{ipso facto} explain why subtelomeric regions should display high rates of recombination. The apparent relationship between dispersed repeats and hypervariable minisatellites still has no obvious explanation (see Chapter 8), but the fact that several minisatellites (including MS32) have expanded from within dispersed repeats suggests more than a coincidence. Other factors such as methylation status, proximity to active genes and location in late or early replicating areas may also influence mutation rate, although too few minisatellites have been characterised to evaluate such factors.

In summary it would appear that the factors influencing and, the mutational processes involved in, minisatellite mutation are many and numerous, and more complex than was initially anticipated, although at some loci interallelic exchange would appear to be operating in at least a proportion of mutation events. It seems likely that the specific factors affecting any one minisatellite will be slightly different, with different loci displaying varying mutational and variational spectrums, reflecting the variety of tandem repeat phenotypes currently observed.
Future directions. So far only a few *de novo* mutation events at minisatellite loci have been identified and even fewer of these characterised at the sequence level. Immediate prospects at MS32 include the detailed characterisation of all of the mutation events so far identified (as for the 1329405 mutant), as well its placing in a wider haplotypic framework. The isolation of more flanking DNA is already under way (A. MacLeod, personal communication), via a vectorette walk (Riley *et al.*, 1990) out from the known sequence. The identification of more distal flanking markers will allow the distinction between genuine recombination events and local conversion events, for both *de novo* mutants and groups of related alleles. The isolation of more mutants should provide increased examples upon which to base inferences of mutation mechanisms. More mutants may be derived from further pedigree analysis, although at a mutation rate of about 1% per gamete, this approach remains relatively inefficient. Additional PCR based strategies for isolating rare length change mutants from bulk genomic germine DNA are currently being developed (A.J. Jeffreys, personal communication). Preliminary results from amplifying multiple aliquots of up to 100 input molecules suggest that mutants may be detected and isolated by such a procedure and initial data would appear to confirm the previously observed size increase bias (~75% size increases). The feasibility of further assays attempting to recover and measure directly interallelic recombinants (and/or flanking convertants) using allele specific flanking primers is also being assessed (A.J. Jeffreys, personal communication). Previously, somatic length change mutants have been identified in tumour DNA (Armour *et al.*, 1989a), although none of these mutants have been characterised using internal mapping. The detailed analysis of such somatic mutants, and further somatic mutants potentially derived from bulk genomic DNA using the PCR based strategies described above, should reveal whether mutation properties such as terminal polarity and size increase bias are confined to the germline. Preliminary results suggest that the mutational properties observed at MS32 are at least partially reflected at other loci, the analysis of more minisatellites should help to establish the generality of these findings.

Virtually nothing is known of the mechanistic basis of homologous recombination in mammalian genomes. If however, minisatellites are directly involved in such processes then the study of minisatellites may provide a direct link into the analysis of recombination. As mentioned above, at least four minisatellite specific DNA binding proteins have been identified, although their relevance to minisatellite biology or general DNA repair and/or recombination is unknown. In yeast, a number of DNA repair enzymes have been characterised and many shown to effect levels of DSBR and recombination (*eg* see Cool and Malone, 1992). Several human diseases thought to arise from defects in DNA repair have been identified and the genes for some of these have now been cloned (see Hoeijmakers and Bootsma, 1992). It is possible that some of the enzymes involved in these diseases will also be involved in minisatellite mutation and the analysis of minisatellite mutation in affected individuals may provide an alternative approach to identifying the enzymes involved in minisatellite mutation. As achieved in yeast (see Chapter 7), evidence for specific mechanisms may be obtained by looking for the existence of predicted intermediates such as DSBs and 3’ single stranded overhangs. How this may be achieved in human germline tissue is not obvious, although PCR based strategies may possibly be used to isolate presumably rare and transient intermediates.

An alternative approach to analysing minisatellites in the human genome is to use model systems in other organisms. To this end, a series of 'transminisatellitic' mice have been generated, incorporating the human minisatellite MS32 (Jeffreys *et al.*, personal communication). These mice will provide a tool for the analysis of *de novo* mutation and will allow the analysis of the effects of cis acting sequences, genomic location and environmental factors upon mutation rates. The model eukaryote, yeast, may also provide a suitable system for
studying minisatellite mutation, especially with respect to analysing meiotic events via tetrad analysis. In particular the mechanistic basis for mutation events may prove more tractable in such a simple system, although it should be considered that results obtained in such a system may not be truly reflective of the operation of minisatellites in man.

Minisatellite length changes are not the only mutational processes operating within the genome, although, the processes and rates governing these events are as poorly understood. As with the types of mutation events already described, their study in man is hampered by the relative scarcity of such events. Previously Jeffreys et al. (1990) used PCR based strategies to isolate rare deletion events at a minisatellite from bulk genomic DNA and in this work, attempts to measure the rate of de novo transposition at a defined locus using a similar approach have been described. It is hoped that analogous approaches may be used to detect, measure and characterise further classes of mutation events possibly including deletions, rearrangements, translocations and ultimately base substitutions. Such systems could be used to determine the genetic load in man and ascertain the effects of environmental factors upon it.

Concluding remarks

Minisatellite analysis through the examination of internal variation is providing remarkable new insights into tandem repeat biology and with the development of the simple MVR-PCR procedure its rapid application to a number of fields should prove swift. The potential of MVR-PCR in individual identification is obvious and it is expected that within a few years it will have become another valuable weapon in the forensic scientist's armoury. The ease with which large amounts of highly informative data may be generated has obvious potential in population genetics and human evolution and it is hoped that it will prove a valuable tool in deriving human lineages. Finally, MVR-PCR is allowing us to analyse de novo mutation events in detail and infer from this mutational processes, with for the first time, good evidence for a role for unequal interallelic exchange. The further analysis of such processes and their relevance to general mutational processes should aid enormously in the understanding of the human genome, genetic disease and variation. The mapping and sequencing of the human genome is progressing at an exciting rate, accompanied at the same time as the rapidly expanding and increasingly dynamic field of tandem repeat biology; both should go hand in hand in the understanding of man, his variability and his origins.


cytogenetics of chromosome-specific α-satellite DNA from the centromere of chromosome 2. Genomics 13: 122-128.


References


References
Phylogenetic evidence for multiple Alu source genes. J. Mol. Evol. 35: 7-16.


References


References


References
Minisatellite "Isoallele" Discrimination in Pseudohomozygotes by Single Molecule PCR and Variant Repeat Mapping

DARREN G. MONCKTON AND ALEC J. JEFFREYS

Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom

Received April 1, 1991

The D1S8 hypervariable minisatellite MS32 has a heterozygosity of 97.5% based on detectable differences in allele length using standard Southern blot analysis. It has previously been shown that the basic repeat unit is in itself variable and that this may be used to map the internal structure of an allele. This method has already been used to establish that alleles of the same length may have differing internal structures between nonrelated individuals. We now extend this approach to demonstrate that two apparently homozygous individuals are in fact heterozygotes. For each individual the two comigratory alleles were separated, without cloning, using single molecule dilution (SMD) of genomic DNA and recovery with PCR. Mapping of the variant repeat units revealed highly diverged internal structures and, for one individual, a size difference of one repeat unit (29 bp). SMD and PCR recovery provide an efficient system for separating comigratory alleles without prerequisite for knowledge of sequence differences.


Minisatellite or variable number tandem repeat (VNTR) loci have proven to be of considerable importance in a wide range of genetic applications, especially in the areas of individual identification (Jeffreys et al., 1985) and gene mapping (Nakamura et al., 1987). The primary reason for their widespread use as genetic markers is their often very high level of allele polymorphism and resulting heterozygosity (Wong et al., 1987; Nakamura et al., 1987). Allelic state is usually determined by the measurement of DNA fragment length estimated from Southern blot hybridization of genomic DNA. This system is however limited by the resolving power of agarose gel electrophoresis, not only for distinguishing small differences in allele sizes between individuals, which can lead to spurious departures from Hardy-Weinberg equilibrium (Devlin et al., 1990), but also for discriminating between true and false homozygotes. However, allele length is not the only criterion by which minisatellite loci may be distinguished. DNA sequencing of human minisatellites has revealed repeat unit sequence variability at almost all loci so far investigated (see Jeffreys et al., 1990). This additional level of polymorphism greatly extends the potential resolving power of minisatellite loci allowing alleles of similar, or even identical, size to be distinguished on the basis of internal structure.

The D1S8 minisatellite (MS32) comprises a 29-bp repeat unit and has a reported heterozygosity of 97.5% based on allele length (Wong et al., 1987). Sequencing of D1S8 has revealed an A to G transition in approximately 70% of the repeat units, resulting in the presence or absence of a Haelll restriction site (Wong et al., 1987). In contrast, all repeat units are cut by the restriction enzyme HinII. The location of variant repeat units can be mapped in alleles amplified by PCR. Partial digestion of end-labeled alleles with HinII followed by gel electrophoresis and autoradiography produces a continuous ladder of labeled DNA fragments from which the number of repeat units can be determined. Comparison of Haelll and HinII partial digests enables each repeat unit to be scored as to whether or not it is cleaved by Haelll. Internal maps, or minisatellite variant repeat (MVR) haplotypes, thus generated can be simply encoded as a binary string of repeats cleaved or not cleaved by Haelll (Jeffreys et al., 1990). We have previously shown that alleles of the same length shared by unrelated individuals may have widely differing internal structures, suggesting relatively distant genealogical origins (Jeffreys et al., 1990).

Screening of D1S8 across large panels of unrelated people revealed two individuals (AS89 and MACH, of Pakistani and Chinese descent, respectively) who by

1 To whom correspondence should be addressed.
A B

\[ \begin{align*}
\text{kb} & \quad \lambda \quad 0 \quad M \quad A \\
4.361 & \\
2.322 & \\
2.027 & \\
\end{align*} \]

\[ \begin{align*}
\text{kb} & \quad 2.45 & \\
\end{align*} \]

\[ \begin{align*}
\text{AS89} & \\
\end{align*} \]

\[ \begin{align*}
\text{MACH} & \\
\end{align*} \]

\[ \begin{align*}
\text{kb} & \quad 2.45 & \\
\end{align*} \]

FIG. 1. Separation of minisatellite alleles in pseudohomozygotes by single molecule dilution and PCR. (A) D1S8 alleles amplified from total genomic DNA. For each individual 20 ng blood DNA was PCR amplified in a 10-\(\mu\)l reaction using the nested primers Cl and D corresponding to the flanking region of the D1S8 minisatellite (Ref. (3)). Amplified alleles were electrophoresed through a 1% agarose gel and visualized by ethidium bromide staining. Lanes: \(\lambda\) HindIII marker DNA; 0, zero DNA negative control; M, amplified alleles from individual MACH; A, amplified alleles from individual AS89. (B) Allele separation by single molecule dilution. For each individual, genomic DNA was diluted in 5 mM Tris-HCl (pH 7.5) in the presence of 0.1 nM PCR primers. Ten 10-\(\mu\)l PCR reactions containing either 6, 3, or 0 pg DNA were amplified for 28 cycles using the D1S8 minisatellite flanking primers A plus B. Products were detected by Southern blot hybridization. Primer sequences are given in Ref. (3). Derivatives C1 and D1 incorporate a 5' extension containing an EcoRI restriction site that was used to generate end labeled PCR products.

Sau3AI restriction digestion and Southern blot analysis appeared to be homozygous for an approximately 2.5-kb allele at D1S8 (data not shown). PCR amplification of D1S8 from total genomic DNA revealed a single band on an ethidium bromide-stained gel for each individual (Fig. 1A). To determine if these individuals were indeed true homozygotes we attempted to map the internal structures of these alleles. Internal mapping from total genomic DNA produced an ambiguous autoradiograph with widely differing band intensities, presumably reflecting a composite internal map derived from two comigratory but discrete alleles (Fig. 2, lanes marked T). Since the alleles were inseparable by standard agarose gel electrophoresis, the individual alleles were separated by single molecule dilution (SMD) of genomic DNA (Jeffreys et al., 1990; Ruano et al., 1990) and recovery with PCR (Saiki et al., 1988) using PCR primer pair A plus B, which correspond to the flanking sequence of the D1S8 minisatellite and allow amplification of the entire minisatellite allele (Jeffreys et al., 1990). After 28 cycles of amplification a 5-\(\mu\)l aliquot was removed, electrophoresed, and Southern blot hybridized (Fig. 1B). Internal nested primers C plus D were used to reamplify the alleles from each of the presumptive single molecule-positive reactions up to a level visible on an ethidium-stained gel. All four alleles were completely mapped from both ends using DNA derived from at least three separate single molecule reactions for each allele (Fig. 2).

Individual AS89 was found to have two alleles of identical size, containing 71 repeat units, but of widely differing internal structures. One allele shares a 5' MVR haplotype common to many individuals (unpublished data). The other allele belongs to another previously characterized set of homogeneous alleles composed almost entirely of repeats cleaved by HaeIII. This individual is therefore a true heterozygote at this locus. Analysis of the second individual MACH, of Chinese origin, revealed this person to be a compound heterozygote at this locus, each allele having widely diverged internal structures and a length difference of 1 repeat unit (allele lengths 72 and 73 repeat units). Both alleles describe new 5' haplotypes previously unseen in our studies on Caucasian individuals (Jeffreys et al., 1990).

Our method for internal mapping does not involve cloning of single amplified molecules (minisatellites are frequently unstable on cloning in *Escherichia coli*), but samples the average properties of the entire amplified pool of PCR products. For each allele at least three separate single molecule amplifications were performed and in no case was any MVR map discrepancy found between them. As noted in previous experiments (Jeffreys et al., 1990), it appears that Taq polymerase-induced misincorporation errors are not a problem with single molecule minisatellite mapping. Finally, in each case the superimposition of the two separated alleles gave rise to the same composite map as derived from total genomic DNA.
SHORT COMMUNICATION

AS89 alleles

MACH alleles

FIG. 2. Internal mapping of PCR-amplified D1S8 alleles. Internal mapping of all alleles was performed as previously described (3). (A) End-labeled products were partially digested with HindIII (F) or HaeIII (H), electrophoresed on a 1.2% agarose gel, dried, and autoradiographed. Lanes marked A and B are single molecule-derived alleles; lanes marked T are internal maps derived from total genomic DNA. (B) Encoded maps of all four single molecule-derived alleles, where A = repeat unit cleaved by HaeIII and T = repeat unit not cleaved by HaeIII.

REFERENCES

Minisatellite variant repeat (MVR) mapping: analysis of ‘null’ repeat units at D1S8

Keiji Tamaki, Darren G. Monckton, Annette MacLeod, David L. Neil, Maxine Allen and Alec J. Jeffreys*

Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK

Received June 4, 1992; Revised and Accepted July 27, 1992

ABSTRACT

Minisatellite variant repeat mapping by PCR (MVR-PCR) is a new approach to studying variation in human DNA which analyses interspersion patterns of variant repeats within minisatellite arrays. MVR-PCR has been applied to the hypervariable human minisatellite D1S8 which contains two major classes of variant 29bp repeat units designated a-type and t-type. The MVR-PCR assay uses a- or t-type specific primers, together with an amplimer at a fixed site in the DNA flanking the minisatellite, to reveal the interspersion patterns of variant repeats along an allele. Extreme levels of variation are seen both in the internal structures of individual alleles and in the digital code generated from the two superimposed alleles in total genomic DNA. However, occasional repeat units fail to amplify in MVR-PCR, signifying the existence of further repeat sequence variants termed ‘null’ or O-type repeats. Although not significant in individual identification, correct genotyping of null repeats is important when using MVR digital codes in parentage analysis. We have therefore characterised these null repeats and show that most null repeats share a common variant repeat sequence. We discuss the possible origins of null repeats and their application to paternity testing and the analysis of minisatellite evolution.

INTRODUCTION

Minisatellites or variable number tandem repeat (VNTR) loci sometimes show very high levels of allele length polymorphism and heterozygosity (1), making them particularly useful for individual identification (2), paternity testing (3) and linkage mapping (4). In addition to length variability, most previously sequenced human minisatellites also differ in the interspersion pattern of variant repeat units along alleles (see 5).

One such minisatellite, D1S8 (MS32), consists of a 29bp repeat unit showing two classes of minisatellite variant repeat (MVR) which differ by a single base substitution, resulting in the presence or absence of a HaeIII restriction site (1), [designated a-type, cut by HaeIII and t-type, not cut by HaeIII, (5)]. Very high levels of variation in the interspersion patterns of a- and t-type repeats within MS32 alleles have been revealed by HaeIII digestion of PCR-amplified alleles (5). Recently we have developed a technically much simpler PCR-based mapping system (MVR-PCR) to assay the same sites of internal variation by using an MVR primer specific to one or other type of variant repeat unit plus a primer at a fixed site in the DNA flanking the minisatellite (6). PCR amplification from a single allele generates two complementary ladders of amplified products, each band or rung corresponding to the position of either an a-type or t-type repeat unit within the minisatellite repeat array. Progressive shortening of resultant PCR products by internal priming of the MVR specific primer is prevented by the use of a ‘tagged’ amplification system which uncouples MVR detection from subsequent amplification (see reference 6 for details). This method can be applied not only to separated alleles, giving binary codes for the interspersion patterns of a-type and t-type repeats along alleles, but also to total genomic DNA to display the diploid digital code derived from the superimposed MVR maps of both alleles. The resulting codes are highly suitable for computer analysis and for the creation of DNA profile databases, with no error-prone size estimations involved. This technique reveals enormous levels of individual variation, arising from extreme variability in allelic MVR maps, and has clear potential for individual identification.

However, in MVR-PCR, minisatellite repeat units on the coding ladder occasionally fail to be amplified by either a-type or t-type MVR-specific primers, indicating the existence of unamplifiable ‘null’ repeats. Runs of apparent ‘null’ repeats can arise in the coding ladder beyond the end of short alleles; such non-existent ‘null’ repeats can be reliably identified by MVR-PCR both of separated alleles and in total genomic DNA (6). Null or O-type repeats can also arise from additional repeat unit sequence variants within alleles which differ enough to prevent priming by either a-type or t-type repeat primers. 1.6% of repeat units within Caucasian alleles are O-type repeats (6), and can be accurately identified in separated alleles by the absence of specific internal rungs on the MVR ladder. However, their detection in total genomic DNA requires correct interpretation of MVR-PCR band intensities (dosage). While the correct discrimination of heterogeneous null positions (for example, homozygous a/a versus heterozygous a/O) has very little effect on the power of digital codes in individual identification (6), correct identification of heterogeneous null codes is important when using diploid codes for parentage analysis. For example, suppose that, at a given repeat position, the mother is t/t, the father a/O and the child t/O; mis-scoring of the father as a/a or the child as t/t would lead to a false parental exclusion, exactly analogous to the problems created by null alleles at classical marker systems.

We now describe the sequence analysis of MS32 null repeats and their detection using new specific MVR primers which define

* To whom correspondence should be addressed
the major proportion of null repeats. We also analyse the
distribution of null repeats within and between alleles and discuss
their application in paternity testing and analysing minisatellite
evolution.

RESULTS

Sequence analysis of MS32 null repeat units

We have previously described how haplotypic MVR maps from
individual alleles can be determined, either from
electrophoretically separated alleles or by pedigree analysis of
digital codes generated by MVR-PCR from total genomic DNA
(6). From this survey, three individuals were chosen, each of
whom had an MS32 allele containing one or more null or O-
type repeat(s) within the first 20 repeat units. Separated alleles
or total genomic DNA were amplified by MVR-PCR to the point
where PCR products could be visualised directly on agarose gels
by staining with ethidium bromide; up to 20 repeat rungs on the
MVR ladder could be generated (data not shown). For separated
alleles, the rung two repeat units above the null repeat was excised
from the gel, re-amplified and sequenced. For total genomic
DNA, a suitable band specific to the relevant allele was identified
at an a/t heterozygous rung position above the position of the
O-type repeat, followed by purification and sequencing.

The sequences of the three null repeats characterised are shown
in Table 1. All three shared the same deletion of an A 3bp 3'
to the G/A polymorphic site which distinguishes a- and t-type
repeat units. The null repeat unit sequences were otherwise
normal and contained either G or A at the major polymorphic
site. This single base deletion is therefore sufficient to block
priming by the MVR-PCR primers 32-TAG-A and 32-TAG-T.

Null repeats containing this variant are referred to as N-type
repeats.

MVR-PCR of N-type repeats

To determine the frequency of N-type repeats in MS32 alleles,
a new MVR-PCR primer, 32-TAG-N, was designed to prime
specifically from these repeats. This primer incorporates the TAG
sequence as previously described (6) and can be used in MVR-
PCR as a replacement for the a- or t-type specific primer (Table
1 , Figure 1 ). The majority of individuals previously identified
as containing alleles with O-type repeats (6) were remapped u sing
32-TAG-N (Figure 1). Most O-type repeats were positively
identified by primer 32-TAG-N at the position previously
identified from intensity differences in the A and T lanes
(32-TAG-A, 32-TAG-T) as being heterozygous or homozygous
for an O-type repeat. A minority of O-type repeat units failed
to amplify with 32-TAG-N (Figure 1, individuals 4 and 5),
indicating the presence of additional repeat unit variant(s) which
could not be detected by primers 32-TAG-A, -T or -N.

In a survey of the first 50 repeat units in 391 different Caucasian
and Japanese alleles (18,790 repeat units in total), 285 repeats
were O-type (1.5%), of which 241 were detected by 32-TAG-N
(Table1). Thus 84.5% of O-type repeats share the A deletion,
the possibility of additional variation between these N-type repeats
which does not block priming by 32-TAG-N cannot however be
excluded. The incidence of N-type repeats is very similar in
Caucasians and Japanese (1.39% and 1.26% of all repeats,
respectively).

Sequencing of one of the minor null repeats

In an attempt to further characterise the remaining O-type repeat
units not detected by 32-TAG-N, a single repeat unit of this type
was sequenced from a Japanese allele. This J-type repeat
contained a C—T transition immediately 3' to the major G/A
polymorphic site in an otherwise normal repeat unit sequence
(Table 1). A new PCR primer (32-TAG-J) designed to assay this
sequence variant was tested on all DNA samples that contained
O-type repeat units not detected by 32-TAG-N. Only 3 repeat
units in 2 different Japanese alleles were detected with this primer
(data not shown). The remaining O-type repeat units not detected
by 32-TAG-N or 32-TAG-J are referred to as U-type
(undetectable) repeats and contain as yet uncharacterised repeat
variant(s). The frequency of U-type repeats varies substantially
between Caucasian and Japanese alleles (0.18% vs. 0.49% of
all repeat units, respectively).

Table 1. Sequence of MS32 variant repeat units and their incidence in the first 50 repeat units of Caucasian and Japanese alleles

<table>
<thead>
<tr>
<th>Repeat type</th>
<th>Repeat unit sequence</th>
<th>Caucasian (%) (n = 15922)</th>
<th>Japanese (%) (n = 2868)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>5'-GGCCAGGGGTGACTCAGAATGGAGCAGGY-3'</td>
<td>73.40</td>
<td>75.10</td>
</tr>
<tr>
<td>t</td>
<td>5'-GACCAGGGGTGACTCAGAATGGAGCAGGY-3'</td>
<td>25.17</td>
<td>22.94</td>
</tr>
<tr>
<td>N</td>
<td>5'-GRCCGGGGGTGACTCAGAATGGAGCAGGY-3'</td>
<td>1.26</td>
<td>1.39</td>
</tr>
<tr>
<td>J</td>
<td>5'-GGTCAGGGGTGACTCAGAATGGAGCAGGY-3'</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td>U</td>
<td>unknown</td>
<td>0.18</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Y=C or T, R=G or A, a-, t-, N-, and J-type repeat units were detected by the following MVR-specific primers: 32-TAG-A, 5'-tcatgcgtccatggtccggaCATTCTGAGTCACCCCCGGC-3'; 32-TAG-T, 5'-tcatgcgtccatggtccggaCATTCTGAGTCACCCCCGGT-3'; 32-TAG-N, 5'-tcatgcgtccatggtccggaCCATTCTGAGTCACCCCCGG-3'; 32-TAG-J, 5'-tcatgcgtccatggtccggaCCATTCTGAGTCACCCCCGGA-3'. The 3' sequence of each primer (uppercase) is complementary to each repeat unit
variant and is preceded by a common TAG sequence (lowercase) used to drive subsequent amplification. U-type repeat units are not amplified by any of these
MVR-specific primers. a, number of repeat units scored in 324 different Caucasian alleles and 59 different Japanese alleles.

Figure 2. Distribution of O-type repeats in MS32 alleles. A: the incidence of N-type and U-type null repeat units at each position over the first 50 repeat units
of 391 different MS32 alleles (331 Caucasian and 60 Japanese). The data does not show a significant deviation from a null hypothesis of random distribution of
total O-type repeats along alleles, ANOVA I Analysis F 4/4 = 0.97, p > 0.05. B: variation in the number of null repeat units within the first 50 repeats of 391
different MS32 alleles. Distributions are shown for N-type, U-type and total O-type repeats. The data (total O-type) shows a significant deviation from a null hypothesis
of random distribution of null repeat units between alleles assuming a poisson distribution (λ = 235, 3 d.f., p < 0.001). Nonexistent 'null' repeats beyond the end
of short (<50 repeat units) alleles (2.2% of all alleles) are ignored in this analysis. C: examples of alleles containing unusual arrangements of null repeat units.
The second allele was derived from the diploid codes of a single child-mother-father trio and contains an ambiguous position (marked ?).
Distribution of null repeats in MS32 alleles

23% of alleles (91 alleles out of 391 different Caucasian and Japanese alleles typed) contained one or more null repeats within the first 50 repeat units. Null repeats appear to occur with equal likelihood at any position within the mapped region of these alleles (Figure 2A). Analysis of the number of null repeats in different alleles (Figure 2B) showed clear evidence of clustering of nulls, particularly N-type repeats, within a limited number of alleles. In one extreme case, 12 N-type repeats were present within the first 50 repeats, and in another bizarre case, an allele contained a succession of 8 U-type repeats followed by NaN embedded in an allele otherwise fixed for a-type repeats (Figure 2C).

Although the vast majority of MS32 alleles so far typed have different MVR maps, different alleles can nevertheless show internal regions of significant map similarity suggesting recent common ancestry of these allele segments [Figure 3; (5, 6)]. These shared haplotypic segments occur much more frequently at one end of MS32 alleles, distal to the unstable proximal region mapped by MVR-PCR which contains a localised mutation hotspot (6). 59% of the 391 different MS32 alleles so far mapped can be aligned into 40 different groups of related alleles. 77% of the 91 alleles that contain O-type repeat units fall within these aligned groupings. In every case where two or more alleles shared a null repeat at equivalent positions within the shared haplotype, MVR-PCR showed that the null repeats were of identical types (almost always N-type repeats) (Figure 3, groups A, B). Additional N-type repeats restricted to just one of the alignable alleles almost always lay outside the shared haplotypic region. In contrast, U-type repeats tend to occur sporadically within otherwise preserved haplotypes shared by related alleles and are usually confined to only one of the aligned alleles (Figure 3, group C).

Effect of null repeats in paternity testing

To use digital MVR codes from total genomic DNA for paternity analysis (6), it is necessary to identify correctly coded positions heterozygous for null repeats (a/O, t/O). However, since null repeats are scarce, the presence of a null-containing paternal allele in a child will add substantially to the ability of MVR-PCR to exclude non-fathers of such a child. To estimate the overall effect of null repeat units in paternity testing, the MVR codes of 141 different Caucasian mother-child duos were each compared with 302 different unrelated Caucasians over the first 50 repeat units (42,582 different mother-child-nonfather trios in total). On average, 9.6 exclusions were obtained per comparison, of which 4.6 were paternal-specific and the remainder directionally ambiguous (Figure 4A); 97.5% of non-fathers showed at least one paternal-specific exclusion, and 98.6% showed one or more exclusions in total (paternal-specific plus directionally ambiguous). Since maternity is seldom an issue in paternity cases (8), the first 50 repeats contain enough information to exclude on average 99.86% of unrelated non-fathers. To determine the contribution of null repeats to this efficiency, the simulated paternity cases were re-evaluated after elimination of all code positions in each child heterozygous for a null repeat (a/O, t/O), including both authentic null repeats and non-existent 'null' repeats beyond the end of short alleles (Figure 4B). As expected, the mean number of exclusions fell significantly, causing a drop in the proportion of non-fathers showing exclusions in total from 99.86% to 99.50% (97.5% to 95.9% for paternal-specific exclusions only).

These estimates for the efficiency of non-paternal exclusion are a mean over all mother-child duos. Variation between duos in levels of exclusion was therefore investigated (Figure 4C). The proportion of the 302 'non-fathers' excluded varied substantially.

**Figure 3.** Examples of groups of aligned alleles containing null repeats. Groups of alignable alleles were identified as previously described, allowing for misalignments between the beginnings of different alleles (6). Common haplotypic segments shared by different alleles are shown in uppercase, and the positions of null repeat units are indicated by "+" for N-type repeats and "*" for U-type repeats. High-order repetitive structures within alleles are arrowed. Some allele maps were derived from the diploid codes of single child-mother-father trios and thus contain ambiguous positions (marked '?').

**Distribution of null repeats in MS32 alleles**

23% of alleles (91 alleles out of 391 different Caucasian and Japanese alleles typed) contained one or more null repeats within the first 50 repeat units. Null repeats appear to occur with equal likelihood at any position within the mapped region of these alleles (Figure 2A). Analysis of the number of null repeats in different alleles (Figure 2B) showed clear evidence of clustering of nulls, particularly N-type repeats, within a limited number of alleles. In one extreme case, 12 N-type repeats were present within the first 50 repeats, and in another bizarre case, an allele contained a succession of 8 U-type repeats followed by NaN embedded in an allele otherwise fixed for a-type repeats (Figure 2C).

Although the vast majority of MS32 alleles so far typed have different MVR maps, different alleles can nevertheless show internal regions of significant map similarity suggesting recent common ancestry of these allele segments [Figure 3; (5, 6)]. These shared haplotypic segments occur much more frequently at one end of MS32 alleles, distal to the unstable proximal region mapped by MVR-PCR which contains a localised mutation hotspot (6). 59% of the 391 different MS32 alleles so far mapped can be aligned into 40 different groups of related alleles. 77% of the 91 alleles that contain O-type repeat units fall within these aligned groupings. In every case where two or more alleles shared a null repeat at equivalent positions within the shared haplotype, MVR-PCR showed that the null repeats were of identical types (almost always N-type repeats) (Figure 3, groups A, B). Additional N-type repeats restricted to just one of the alignable alleles almost always lay outside the shared haplotypic region. In contrast, U-type repeats tend to occur sporadically within otherwise preserved haplotypes shared by related alleles and are usually confined to only one of the aligned alleles (Figure 3, group C).

**Effect of null repeats in paternity testing**

To use digital MVR codes from total genomic DNA for paternity analysis (6), it is necessary to identify correctly coded positions heterozygous for null repeats (a/O, t/O). However, since null repeats are scarce, the presence of a null-containing paternal allele in a child will add substantially to the ability of MVR-PCR to exclude non-fathers of such a child. To estimate the overall effect of null repeat units in paternity testing, the MVR codes of 141 different Caucasian mother-child duos were each compared with 302 different unrelated Caucasians over the first 50 repeat units (42,582 different mother-child-nonfather trios in total). On average, 9.6 exclusions were obtained per comparison, of which 4.6 were paternal-specific and the remainder directionally ambiguous (Figure 4A); 97.5% of non-fathers showed at least one paternal-specific exclusion, and 98.6% showed one or more exclusions in total (paternal-specific plus directionally ambiguous). Since maternity is seldom an issue in paternity cases (8), the first 50 repeats contain enough information to exclude on average 99.86% of unrelated non-fathers. To determine the contribution of null repeats to this efficiency, the simulated paternity cases were re-evaluated after elimination of all code positions in each child heterozygous for a null repeat (a/O, t/O), including both authentic null repeats and non-existent 'null' repeats beyond the end of short alleles (Figure 4B). As expected, the mean number of exclusions fell significantly, causing a drop in the proportion of non-fathers showing exclusions in total from 99.86% to 99.50% (97.5% to 95.9% for paternal-specific exclusions only).

These estimates for the efficiency of non-paternal exclusion are a mean over all mother-child duos. Variation between duos in levels of exclusion was therefore investigated (Figure 4C). The proportion of the 302 'non-fathers' excluded varied substantially.

**Figure 3.** Examples of groups of aligned alleles containing null repeats. Groups of alignable alleles were identified as previously described, allowing for misalignments between the beginnings of different alleles (6). Common haplotypic segments shared by different alleles are shown in uppercase, and the positions of null repeat units are indicated by "+" for N-type repeats and "*" for U-type repeats. High-order repetitive structures within alleles are arrowed. Some allele maps were derived from the diploid codes of single child-mother-father trios and thus contain ambiguous positions (marked '?').
Figure 4. Efficiency of MS32 diploid codes and the effects of null repeats in paternity testing. Diploid codes extending for at least 50 repeat units were obtained from 141 Caucasian mother-father-child trios. For each trio, the father was removed and replaced sequentially by each of 302 different unrelated Caucasian individuals ('non-fathers'). The MVR codes of each mother-child plus non-father trio were analysed over the first 50 repeat units to determine the total number of repeat unit positions which gave an exclusion, plus the number of paternal-specific exclusions and the number of exclusions which were directionally ambiguous (e.g. mother and non-father both a/a, child a/t). A: frequency distribution of paternal-specific exclusions (filled bars), ambiguous exclusions (shaded bars), and total exclusions (open bars) for each of the 42,582 combinations of mother-child and non-father. The mean number of exclusions was 4.56, 5.09 and 9.65 per child, respectively. The overall proportion of non-fathers showing no exclusions, or no paternal-specific exclusions, was 0.14% and 2.5% respectively. B: Comparison between the frequency distribution of total exclusions obtained before (open bars) and after (shaded bars) elimination of all child code positions heterozygous for an O-type repeat (a/O, t/O), including nonexistent 'null' repeats beyond the end of short alleles. In the latter analysis, on average 3.93 paternal-specific, 4.30 ambiguous and 8.23 total exclusions were obtained per trio. 0.50% of non-fathers showed no exclusions, and 4.1% showed no paternal-specific exclusions. C: variation in the proportion of non-fathers excluded for each of the 141 mother-child combinations. The proportions are binned into 1% intervals. Filled bar, non-fathers eliminated by paternal-specific exclusions. Hatched bar, men showing paternal-specific exclusions after elimination of all heterozygous O-type repeat positions in the child's code. Shaded bar, non-fathers showing any exclusion (paternal-specific plus ambiguous). Open bar, non-fathers with any exclusion after elimination of O-type repeats from the child.
from duo to duo, depending on the precise nature of the MVR codes in the mother and child. In the worst case, only 97% of non-fathers could be excluded (75% if only paternal-specific exclusions are used). As expected, these estimates are worsened if null repeat positions are eliminated from the analysis.

**DISCUSSION**

Internal mapping of variant repeat units within minisatellites provides an important new approach both to DNA typing and to the analysis of allelic variability and minisatellite mutation processes. Work to date on minisatellite MS32 has concentrated on a G/A base substitutional polymorphism originally defined by the presence/absence of a \textit{Hae}III cleavage site within repeat units (5, 6). A second common polymorphic site 2 bp from the variable G/A site has been found from sequence analysis of cloned MS32 [(1); see Table 1] but has yet to be used for internal mapping. MVR-PCR has now revealed additional rare variants defined operationally as 'null' repeats which cannot serve as priming sites for the MVR-PCR primers 32-TAG-A or -T. These variants have presumably arisen by repeat unit sequence mutation, and their incidence governed by a balance between mutation and fixation/extinction of variant repeats within and between repeat arrays by processes such as unequal exchange and replication slippage. The relative scarcity of null repeats makes them particularly useful for identifying related alleles and confirming the authenticity of allele alignments.

87% of null repeats in Caucasian alleles share a common variant, the N-type repeat, which can now be detected reliably by MVR-PCR. The widespread occurrence of N-type repeats in both Caucasian and Japanese alleles and their presence in many groups of aligned alleles suggest that this variant arose fairly early in the evolution of MS32 alleles. Several different groups of aligned alleles contain N-type repeats within a '\textit{NataNata}' motif (Figure 3), suggesting a 'supergroup' of alleles sharing homologous patches of tandem repeats within alleles which are otherwise not obviously alignable.

The remaining null repeats include the rare J-type repeat and the as yet unsequenced U-type repeats. 8% of alleles contain U-type repeats, and the majority of these alleles (26/33) have only a single U variant over the region mapped, suggesting recent mutation without subsequent diffusion into neighbouring repeats. This is supported by U-containing alleles which can be aligned with other alleles; in each of the five cases where the U-type repeat lay within a haplotypic segment shared by several alleles, other alleles contained an a- or t-type repeat at the corresponding position (see Figure 3, group C). This suggests either very recent repeat unit mutation from a or t to U, and thus sequence heterogeneity amongst different U-type repeats, or possibly that the U repeat is ancestral within a group of aligned alleles and has recently been replaced by an a- or t-type repeat by a process such as microconversion which does not affect repeat unit copy number or the flanking MVR map. Two probable instances of U-type repeat diffusion subsequent to mutation have been found. In one Japanese allele with two U repeats, the variants are contained within a perfect high-order tandem repeat of a 24 repeat unit segment commencing 3 repeat units from the beginning of the allele, and presumably contain the same variant (not shown). In the second case, an English allele contains a block of 8 U-type repeats (Figure 2C) which again have presumably expanded from a single mutant repeat.

The existence of variant repeats with abnormal repeat length, for example the N-type repeat 28bp rather than 29bp long, could create problems in digital coding from genomic DNA, by moving the MVR ladders of each allele out of register. In practice however, aberrant length repeats do not appear to present a significant problem; in the worst individual so far found with 12 N-type repeats in one allele and none in the other, the digital code could be unambiguously read for more than 50 repeat unit positions, although the normally perfect spacing of rungs on the ladder was slightly perturbed by the progressive misalignment of the two allele ladders (maximum misalignment of 12bp for the 50 repeat unit PCR products 1714bp long) (data not shown). MVR-PCR can also be used for paternity testing, provided that heterozygous null positions (a/O, t/O) in diploid codes (3.0% of all positions) can be reliably identified. Experience to date suggests that these positions can be identified with >90% reliability solely from band intensity information using primers 32-TAG-A and -T alone. The ability to detect definitively the substantial majority of null repeats using primers 32-TAG-N and -J substantially increases the reliability, to provide a single locus which is remarkably effective at excluding unrelated non-fathers, though limited by the high \textit{de novo} mutation rate creating new MVR haplotypes at MS32 (6).

**MATERIALS AND METHODS**

Preparative MVR-PCR and sequencing of null repeat units
100ng samples of total genomic DNA, or equivalent amounts of individual alleles separated from Mbol digests of total genomic DNA by preparative gel electrophoresis, were amplified in 30ul reactions in the presence of 1.5 units AmpliTag (Perkin-Elmer-Cetus), using the primers and PCR buffer system described previously (6). Reactions were cycled for 1.3 min at 96°C, 1 min at 68°C, and 5 min at 70°C for 30 cycles on a DNA Thermal Cycler (Perkin-Elmer-Cetus), followed by a 2 cycle chase of 1 min at 68°C, 10 min at 70°C. Amplified products were electrophoresed through a 1.1% agarose gel and visualized by ethidium bromide staining. An appropriate 'rung' in the MVR-PCR ladder required for sequencing was excised from the gel and purified by electroelution onto dialysis membrane. The purified PCR product was reamplified with PCR primers 32-O and TAG using the same cycling conditions as before for a further 18 cycles. The double-stranded PCR product was re-purified by electrophoresis and electroelution and sequenced directly (7).

**ACKNOWLEDGEMENTS**

We thank Debra Langton for oligonucleotide synthesis, John Armour, Nicola Royle, Ian Gray, Mark Gibbs, Andrew Collick and Yuri Dubrova for helpful discussions and H.Cann, J.Dausset, Y.Katsuramata and P.Debenhin for providing human DNA samples. This work was supported by grants to AJJ from the Medical Research Council, the Wolfson Foundation and the Royal Society. This work is the subject of Patent Applications.

**REFERENCES**

DIGITAL DNA TYPING

New views of the Kenya rift
Minisatellite repeat coding as a digital approach to DNA typing

Alec J. Jeffreys, Annette MacLeod, Keiji Tamaki, David L. Neil & Darren G. Monckton

Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK

Most DNA typing systems used in forensic and legal medicine assay allelic length variation at tandem repetitive DNA regions such as minisatellites. A simple alternative approach that displays patterns of variant repeat units along minisatellite alleles is described here. This produces DNA profiles as extraordinarily variable digital sequences appropriate for forensic investigations, including computer databasing, and for analysing allele diversity and the role of recombination in minisatellite instability.

HYPERVERSIBLE human DNA markers have had a profound impact on forensic and legal medicine. Most DNA typing systems use tandem-repetitive minisatellites or VNTR (variable number tandem repeat) loci which can show extreme levels of allelic variability in repeat copy number and therefore DNA fragment length. Polymerase chain reaction (PCR) amplification of hypervariable loci, including very short tandem-repetitive 'microsatellites', has greatly increased the sensitivity of DNA typing systems and the ability to type degraded human DNA. Despite the power of current DNA typing systems, technical problems have prevented their full potential from being realized. Minisatellite allele lengths can vary in a quasicontinuous fashion, making unequivocal allele identification impossible. Error-prone allele length estimates and electrophoretic 'band-shifts' can occasionally lead to apparent exclusions between 'matching' DNA profiles, and more seriously greatly weaken the potential statistical power of population and investigative DNA profile databases. Although microsatellites and other simple tandem repeat loci generate PCR-amplifiable alleles that can in principle be sized with precision on DNA sequencing gels, restricted allelic variability and the frequent occurrence of PCR-artifact bands limit their usefulness.

There has also been considerable debate over the evaluation of the statistical weight of DNA profile evidence (see for example refs 18–20), in particular over the estimation of appropriate 'allele' frequencies from limited population databases and the assumption, required for genotype frequency estimation, that alleles associate at random in human populations.

We show here that many of these limitations of current DNA typing systems can, in principle, be overcome by assaying sequence variation in minisatellite alleles, rather than length differences.

Minisatellite variant repeat mapping

Minisatellite alleles frequently vary not only in repeat copy number but also in the interspersion pattern of variant repeat units along alleles (Fig. 1a). Previous analysis of the hypervariable locus DIS8 (probe MS32) showed two classes of repeat unit (designated a-type, t-type) that differ by a single base substitution which creates or destroys a HaeIII restriction site. Interspersion patterns of HaeIII and HaeIII repeat units along minisatellite alleles is described here. This produces DNA profiles as extraordinarily variable digital sequences appropriate for forensic investigations, including computer databasing, and for analysing allele diversity and the role of recombination in minisatellite instability.
FIG. 2 Examples of minisatellite allele repeat coding by MVR-PCR on separated alleles (lanes 1–3) and total genomic DNA (lanes 4–9). DNA samples were amplified using 32-TAG-A (A) or 32-TAG-T (T) in a high concentration of primers 32D and TAG. PCR products were separated by agarose gel electrophoresis and detected by Southern blot hybridization. The first repeat unit (asterisk) is weakly detected and cannot be scored reliably. Diploid scoring on total genomic DNA therefore commences at the second repeat unit (code position 1). This start position for reading the code was confirmed by scoring on total genomic DNA therefore commences at the second repeat unit (asterisk) is weakly detected and cannot be scored reliably. Diploid

FIG. 3 Individual variation in diploid MVR codes. Codes extending for at least 50 repeat units were assayed for PCR amplification of an entire allele followed by partial digestion with HaeIII. This approach, although cumbersome and limited to alleles small enough (<5 kilobases (kb)) to amplify by PCR, provides an unambiguous binary code for an allele (Fig. 1a), and has revealed high levels of allelic variation in MS32 minisatellite variant repeat (MVR) maps, particularly at one end of these alleles.  

MVR mapping by PCR

We have now developed a much simpler MVR mapping system applicable to MS32 alleles of any length. The approach outlined in Fig. 1b,c uses two different MVR-specific primers which prime off either a-type or t-type repeat units. Amplification using one or other primer together with amplimer 32D from a fixed site in the minisatellite flanking DNA will generate two complementary sets of products from the ultraviable end of any

METHODS. Samples (100 ng) of genomic DNA, or equivalent amounts of DNA were digested with Mbol and with MS32 (data not shown).  

METHODS. All MVR codes were determined as described in Fig. 2. 62±6 repeat units were scored per individual. If there was any doubt about the coding state at any given repeat position in an individual, the position was marked with a question mark. Only 0.3% of code positions (59 in 20,720 scored) were marked ? and were ignored in database searches. Data were stored as ASCII files and analysed using software written in VAX BASIC V.3 and run on a VAX 8650 computer operating on VMS 5.3–1.
MS32 allele, from which the MVR map can be deduced. To prevent progressive shortening at each PCR cycle because of MVR-specific primers priming internally in PCR products, MVR detection and subsequent amplification were uncoupled by providing each MVR-specific primer with a 20-nucleotide (nt) 5' extension ‘TAG’ and carrying out amplifications with a low concentration of one or other tagged primer and high concentrations of 32P and the TAG sequence itself.

Application of MVR-PCR at limited cycle number to MS32 alleles separated from genomic DNA generated continuous complementary ladders of PCR products detectable by Southern blot hybridization and extending >3 kb (100 repeat units) into each allele, from which allele binary codes could be read (Fig. 2). These were consistent with codes determined by partial digestion with HaeIII (data not shown). With additional PCR cycles, binary codes could be determined from PCR products directly visualized on ethidium bromide-stained gels, although overamplification and collapse of minisatellite PCR products' limiting coding to roughly the first 15 repeat units (data not shown).

Occasionally, a ‘run’ on the MVR coding ladder failed to be amplified by either MVR-specific primer (Fig. 2), indicating the presence of ‘null’ repeats containing additional sequence variant(s) 3' to the HaeIII site which block priming by either primer. It was found that 1.6% of repeat units scored from 32 separated Caucasian alleles were null or O-type repeats, compared with 72.9% a-type repeats and 25.5% t-type repeats. O-type repeats tend to cluster in a limited number of alleles (see for example Figs 4 and 5), and can correspond both to HaeIII-cleavable and HaeIII-resistant repeat units.

MVR-PCR on genomic DNA

MVR-PCR on total genomic DNA produces a profile of both alleles superimposed, to generate a ternary code for two-variant alleles (Fig. 1a), where each rung in the ladder can be coded as if both alleles a-type at that position, a, 2(both t-type, t) or 3(heterozygous, at). The presence of O-type repeats creates three additional coding states, namely 4(aO), 5(tO) and 6(OO). The last will appear as a gap on the ladder. Coding states 4-6 will also be generated beyond the end of the shorter allele, as the code will be derived from only one allele. No PCR products will appear beyond the end of the longer allele, generating a 66666... code.

Examples of MVR-PCR on total genomic DNA are shown in Fig. 2. In each case, diploid codes could be read at least 50 repeat units into the minisatellite. The two tracks generating the code contain considerable informational redundancy; in almost all cases, an intense band in the A track was matched by no band in the T track (code 1, aa), a faint A band by a faint T band (code 3, at) and no A band by an intense T band (code 2, tt). This dosage phenomenon provides a detailed check on the authenticity of the code generated, and also makes it possible to identify with good reliability individuals with short alleles and rung positions which are heterozygous for a null or O-type repeat (code 4, aO; code 5, tO) (Fig. 2).

Variation in diploid codes

MVR-PCR typing of 334 unrelated individuals showed that there were on average 30 code mismatches per pair of individuals over the first 50 repeat units (Fig. 3a). No two individuals shared the same MVR code, and all individuals could be distinguished using only the first 17 repeat positions. Individual specificity remained when band intensity information was removed by converting all code 4(aO) and code 5(tO) positions to codes 1(aa) and 2(tt), respectively, to generate quaternary codes (1, 2, 3, 6) corresponding to bands present only in the A track, only in the T track, in both tracks and in neither track, respectively. The two most similar individuals had MVR codes dominated by code 1(aa) (Fig. 3b), indicating that all four alleles in these two individuals were composed largely of a-type repeats; such homogenous alleles have been noted previously\(^6\). The most dissimilar pairs of individuals arose where one contained a short allele, creating a diploid code dominated by the rare codes 4, 5 and 6. Short (<50 repeats) alleles with allele lengths ranging from 19 to 44 repeat units were found in 7.8% of individuals. Short alleles do not occur with equal frequency in all populations; thus 5.6% of Caucasians typed contained short alleles, compared with 23% of Japanese (P < 0.001).

Heterozygosity levels at MS32

Diploid codes provide an objective method for identifying homozygotes, in contrast to allele length analysis of minisatel-
Presumptive homozygotes will show diploid MVR codes restricted to code 1(aa), 2(tt) and 6(00), with no heterozygous repeat positions. Three individuals (one French, two Japanese) out of 334 surveyed showed homozygosity by this criterion, suggesting a mean heterozygosity level of 99.1%. These apparent homozygotes were not heterozygous for a second unamplifiable allele resulting from a 32D primer mismatch in the flanking DNA (Fig. 1c), as they showed, as predicted, a single band on Southern blot hybridization of genomic DNA (data not shown). Conversely, the majority (8/10) of apparently single-band individuals initially detected by hybridization with MS32 were in fact heterozygous for similar or identical length alleles as shown by diploid coding.

**Allelic variability**

Diploid code variability is governed by the number and frequencies of different MS32 alleles in human populations. Caucasian alleles (337) were mapped, either using electrophoretically separated alleles (Fig. 2) or more simply from pedigree data by extracting the MVR haplotypes of all four parental alleles from the diploid MVR codes of a mother, father and child(ren) (Fig. 4). Haplotype comparisons revealed 326 different alleles, 316 detected only once in the alleles surveyed, together with two alleles sampled twice and one allele detected three times. The maximum frequency of any allele at this locus is therefore very low (3/337 = 0.009). If all alleles were equally rare, Poisson analysis indicates that about 3,500 different MS32 alleles must exist in Caucasians to give this sampling frequency distribution. Given the high mutation rate of MS32 (see below), the true level of allelic diversity in humans is likely to be enormous, with >10^5 different and distinguishable alleles in the current world population of 5x10^9 individuals. Note that MVR mapping could in theory distinguish 3^30 (7 x 10^25) different allelic states using information from the first 50 repeat units.

Different MS32 alleles can have related MVR haplotypes. All 326 different alleles were therefore compared to identify groups of alleles showing significant similarities in repeat maps (Fig. 5a). This heuristic alignment approach showed that 47% of alleles could be classified into 32 different groups each containing 2-22 significantly related alleles; each of the remaining 174 alleles showed no detectable matches with any other alleles. Examples of groups of related alleles are shown in Fig. 5b. Most significantly, the majority of interallelic differences in repeat copy number and variant repeat interspersion pattern cluster at the extreme beginning of the tandem array, over the region previously identified as showing greatest allelic variability.

**Mutation rates and processes**

The extraordinary levels of allelic variability must be maintained by a high de novo mutation rate altering the MVR map of MS32 alleles. To quantify mutation rates, diploid MVR codes were analysed in offspring from the CEPH collection of large families where parentage is beyond dispute (Fig. 6). Seven offspring had MVR codes showing multiple parental exclusions, indicating the presence of a mutant allele. In each case, code positions specifically excluding only one parent were detected (Fig. 6a), defining the parental origin of the mutant allele. Using non-mutant children, it was possible to deduce the MVR map of each mutant allele (Fig. 6a).

The overall mutation rate in MS32 MVR maps is about 0.012 per gamete, with paternal and maternal mutations arising with similar frequency. Curiously, all seven mutation events detected were associated with a gain in repeat copy number, in most cases of a very small number (1-3 repeats), suggesting a directional bias in the mutation process. Although MS32 alleles average 200 repeat units long, mutation events are extremely clustered, in most cases in the first 10 repeat units, the region with maximum allelic variability (Fig. 6c), confirming the presence of a localized mutational hot-spot.

Possible mutation mechanisms were investigated by comparing mutant and progenitor alleles. For mutant d (Fig. 6b, e), the site of the repeat unit addition in the mutant allele is...
preceded and followed by MVR code derived from the same maternal allele. The mutation event is therefore probably intra-allelic, and could have arisen for example by unequal sister chromatid exchange or replication slippage. Mutants a, b and g also seem to have arisen by an intra-allelic event, but the presumptive exchange point lies too close to the beginning of the allele to be certain that the mutant allele does not contain a recombinant haplotype. By contrast, mutant e (Fig. 6a-c) provides clear evidence for interallelic unequal exchange between the two paternal alleles, the mutant allele commencing with one paternal haplotype, then switching after two a-type repeats of unknown origin to the beginning of the other paternal allele. Mutant f also seems to have arisen by interallelic unequal exchange. DNA markers flanking MS32 are insufficiently close and informative to test whether these two mutations have been accompanied by exchange of distal flanking markers.

A recombination hot-spot?

Previous studies of minisatellite mutation have shown that allele length change events are largely, if not completely, restricted to single alleles. Our data provide the first direct evidence that interallelic recombination or gene conversion can have a major role in minisatellite instability. If the two apparently interallelic mutation events have arisen by a conventional (if unequal) interallelic recombination process, this implies a recombination frequency of 2.572 = 0.3 centiMorgans (cM) over the first ~400 base pairs (bp) of the minisatellite, a 700-fold enhancement over the mean frequency of 1cM per 10^6 bp in the human genome. If correct, this would represent a dramatic example of a human recombination hot-spot, and revitalizes earlier speculation that minisatellites may be actively involved in chromosomal processes such as homologue recognition, synopsis and meiotic recombination.

Forensic considerations

MVR-PCR is simple, and preliminary investigations suggest that it should be applicable to forensic analysis. It is sensitive, generating normal profiles down to 10 ng human DNA. Below this level, apparently random fluctuations in band intensity can arise, owing to stochastic loss of PCR products from the small number of input molecules; reliable consensus diploid codes can nevertheless be obtained by comparing replicate MVR profiles from subnanogram amounts of genome DNA. MVR-PCR can also be applied to partially degraded DNA, as it recovers information from any DNA fragments long enough to include the flanking 32D priming site plus at least some 29 bp repeat units. Additional information can be recovered by substitution of the flanking primer 32D for 320 (sequence GAGTTTGTGTTGGAGGTTGGT), which primes immediately adjacent to the start of the tandem-repeat array (Fig. 1c). Although degraded DNA yields a truncated diploid code, it is still compatible with database searches though with reduced discriminating power.

MVR-PCR can also be applied to mixed DNA samples (for example victim plus rapist DNA recovered from semen-bearing vaginal swabs), particularly if pure DNA from one of the two individuals (for example victim) is available. DNA mixing experiments have shown that 10% admixture can be detected, and that comparison of ‘victim’ and mixed DNA samples can yield an ambiguous diploid code of the ‘rapist’ (Fig. 7). For example, if the victim is code 1 (aa) and the mixture contains an additional band in the T track, then the rapist must be code 2 (tt), 3 (at) or 5 (tO). The efficiency of identification using mixed DNA information was assessed by creating 2 x 10^5 different combinations of ‘victim’, ‘rapist’ and ‘false suspect’ from the database of MVR codes. For each case, the ambiguous rapist code deducible from a victim-rapist mixture was checked against the suspect for exclusions. On average, 14 exclusions per case were detected over the first 50 repeat units, and only 14 out of 2 x 10^5 ‘false suspect’s failed to show any exclusion (99.9993% mean exclusion rate).

MVR-PCR could also be used in parentage testing because the diploid code of a nonparent will frequently show exclusionary mismatches with the child (see for example Fig. 6a). To determine the effectiveness of MVR-PCR in excluding nonfathers, 28,635 Caucasian mother-child-nonfather trios were created from the MVR code database and analysed for paternal exclusions. On average, 9.9 exclusions were obtained over the first 50 repeats, of which 4.7 were paternal-specific and the
FIG. 7 Incomplete MVR code information recoverable from mixed DNA samples. Two individuals (X, Y) were chosen at random from a collection of 450 typed people, and their identities were concealed from the analyst. Samples (100 ng of genomic DNA from X and Y, or from X plus Y mixed in the indicated proportions, were amplified by MVR-PCR for 18 cycles and PCR products detected by Southern blot hybridization. S, Standard individual included on all gels. By comparing the MVR-PCR profile of X with the profile of the mixed DNA samples, possible genotypes of Y can be deduced, as indicated, at all repeat positions where X is not code 3(tO), by checking for A- or T-Track specific bands present in the mixture but not X. For example, if the mixture but not X contain a band in the T-track at a given position, then the possible codes of Y at that position are 3(tO) or 5(tO). The MVR code of X and the incomplete and ambiguous MVR code of Y deduced from the mixed DNA samples were screened across the database of 450 individuals to reveal, correctly and uniquely, the identities of X and Y.

between a forensic sample and a criminal suspect, and for creating very large communal population and investigative DNA databases. As there are far in excess of 3,500 different alleles (×10^6 diploid codes), it is likely that very large databases can be constructed before any significant saturation of MVR code types occurs. (Note that this does not hold for pairs of siblings who will have a roughly 1/4 chance of sharing the same parental alleles and therefore MVR code.) Very large databases can provide a simple method for determining the statistical significance of a match between a forensic sample and a suspect, by counting the frequency (probably zero) of the particular MVR code in the appropriate database. This approach uses observed phenotype frequencies, rather than genotype frequencies deduced from limited population databases under assumptions of Hardy-Weinberg equilibrium.

MVR-PCR can be used on any minisatellite locus showing internal variation, provided that variant repeats with abnormal repeat length do not exist, as seems to be the case with MS32 but not with most other minisatelites. Abnormal length repeats will throw the MVR ladders of the two constituent alleles out of register, making at least part of the diploid coding ladder uninterpretable. We have identified a few minisatelites which may be suitable for MVR-PCR and ultimately for multiplex MVR-PCR where several loci are amplified simultaneously to improve specificity, particularly in parentage analysis.

Discussion

MVR-PCR provides a novel and simple method for generating unambiguous and highly discriminatory digital information from human DNA. Many of the problems associated with other DNA typing approaches, such as the definition of DNA profile matching and measurement errors in allele sizing, are obviated. MVR-PCR has also provided a clear view of the extraordinary levels of allelic variability that can exist in human minisatellites, and revealed that recombination is involved in the generated of ultravariability.

remained directionally ambiguous. At least one paternal-specific exclusion was shown by 98.9% of nonfathers, and 99.8% showed at least one exclusion in total (paternal-specific plus ambiguous). The efficiency of maternity testing is, however, limited by the high de novo mutation rate at MS32, and by the need to correctly identify heterozygous null positions in the child (code 4, a0; 5, t0) from interpretation of band intensities.

Diploid MVR coding potentially offers many advantages over currently used DNA typing systems that involve allele length measurements. MVR profiles contain considerable informational redundancy enabling code authenticity to be checked. Code generation does not require standardization of electrophoretic systems, is immune to gel distortions and band shifts, does not involve error-prone DNA fragment length measurement, and does not require side-by-side comparisons of DNA samples on the same gel. MVR-PCR generates digital DNA typing information ideal for determining objectively a match