ALLELIC VARIATION AND MUTATION AT HUMAN HYPERVARIABLE MINISATELLITE LOCI

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

by

David Langley Neil
Department of Genetics
University of Leicester

September 1994

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

For my family

In memory of Dr. J.D. Neil. BA. PhD.

CONTENTS

PREFACE

I had the great good fortune to arrive in this laboratory just as a very exciting field was opened up by a technical breakthrough. The projects which I have been involved with have generated a large amount of data using only a few of the simplest molecular biology techniques. For this reason the Materials and Methods section does not give detailed protocols for every procedure used, rather it describes the general techniques applied, giving references for those adequately and ably described elsewhere, and where necessary, detailing the assembly of recognised procedures into protocols used in this laboratory. This means that the bulk of the thesis is to be found in the main results sections, each of which each constitute a well defined project in their own right. Much of the work in these sections has been published; each chapter is therefore presented in a similar manner to a scientific paper, which can be read independently of the others. As such, these chapters contain introductions to the work involved, experimental details of specific applications of the general methods already mentioned, a presentation of data obtained and a discussion of its implications. The introductory chapter sets the work described in this thesis in a broader framework; relating it to previous work in the fields of; analysis of genetic variation and genomic flux, tandem repeat biology, and the applications DNA variation to forensic science and individual identification. The concluding discussion sets the work presented here in the context of the most recent advances in these areas coming from this and other laboratories.

ABSTRACT

Human polymorphic tandemly repeated loci have been exploited in linkage analysis and have also had a profound impact on forensic and legal medicine. Most DNA typing systems assay allelic length variation at tandem repetitive loci such as minisatellites. Although these include the most informative loci in the human genome, their forensic application is limited by innaccuracies in allele length measurement, and the need to make population genetic assumptions in statistical analysis. Minisatellite alleles frequently vary not only in repeat copy number, but also in the interspersion pattern of variant repeat units, which can be assayed by minisatellite variant repeat mapping (MVR). This technique uses either restriction analysis, or more efficiently MVR-PCR, to display minisatellite allele internal structures as digital codes that can theoretically distinguish millions of alleles. MVR-PCR profiles from the hypervariable minisatellites MS31A and MS32 generate extraordinarily heterogenous codes, reflecting extreme levels of allelic variability, far in excess of that detectable by allele length analysis. These codes are appropriate for forensic investigations and for analysing allelic diversity and mutation. Comparison of MVR structures showed that most alleles at both loci were different, implying the existence of many thousands of alleles worldwide. Variation between groups of alleles with related MVR-structures was largely confined to one end of the locus, providing evidence for a localised mutation "hotspot". This was confirmed by using MVR-PCR to characterise *de novo* mutation events at both loci. Most mutations involved gains of small numbers of repeats at the ultravariable end of the tandem repeat array, by highly polar intra- and interallelic unequal conversion-like processes. Evidence suggested that this mechanism was male germline-specific, and possibly meiotic. Less frequent intraallelic deletions were seen in male and female germline and in somatic tissues. These observations suggest that the same mechanisms of repeat unit turnover may operate at different hypervariable minisatellites.

ACKNOWLEDGEMENTS

Initially I would like to thank my teachers; Mr. Roger Jeffries, for introducing me to many potential fields of biological study, instilling my interest in molecular biology and for telling me I could do better; Dr. Nick Proudfoot, for a DNA biased slant to a Biochemistry degree, and again much encouragement; Dr. Chris Tyler-Smith, for patiently training me in basic molecular biology techniques; and Dr. John Clegg, for accustoming me to self-organised routine lab work. Without them this thesis would not have been possible. Thanks also to my parents, for supporting and encouraging me throughout my higher education and not telling me to get a "proper job".

I would also like to thank all members of G19, past and present, for making my time here so enjoyable and productive, these include; Mark Gibbs, Ian, Moira, Maureen, Jackie, Jane, Rita, Esther, Shaojie, Celia, Jo, Tara, and especially Ila, who organises the lab so efficiently and ensures that it remains a good working environment. Special thanks go to Annette, for showing me the ropes; to John, Andy, Nicola, Yuri, Duncan, and Mark. J, for helpful discussions; to Darren for being a hard act to follow; and to Max for gossip, moans and being in the same boat. Thanks also to Keiji and Chong-Lek Coh for their enthousiastic and ongoing collaboration on the MS31A work. Further thanks go to the whole Genetics department, for showing that a well run establishment can also be friendly and informal, in particular to Terry, Joan and Jenny for help and advice with orders, finance, etc., and to the secretarial staff, Margaret, for keeping the show on the road and Sarah for prompt delivery of mail and faxes.

For out of lab social activities thanks go to Darren, Annette, Ian, and Max for drinks and parties, Duncan and Catherine for drinks and pinball, and also to other members of the Genetics department; Garry, for drinks and finding me somewhere to live, Tim, Pete and Neale, for drinks and barbeques, the cricket team (champions at last), and most of all to Helen, for making life a lot easier during the writing of this thesis and proofreading beyond the call of duty.

Finally my greatest debt of gratitude is owed to Alec, for taking me on to work in his laboratory, it is ultimately a testament to him that it would be impossible for anyone to arrive in this laboratory during anything other than an exciting period filled with significant discoveries and rapid progress. His relaxed supervision allowed me to get on with my own thing and to learn in the best way possible, by making my own mistakes. I cannot thank him enough for the extreme patience and tolerance he showed during the writing of this thesis, and only hope that I can repay this to some extent during my continued work here.

This work was funded by a grant from the SERC

PUBLICATIONS

Some of the work described in this thesis has been published previously:

Jeffreys, A.J., MacLeod, A., Tamaki, K., Neil, D.L. and Monkton, D.G. (1991). Minisatellite Repeat Coding as a Digital Approach to DNA typing. *Nature* **354**: 204-209 (1991).

Jeffreys, A.J., Monckton, D.G., Tamaki, T., Neil, D.L., Armour, J.A.L., MacLeod, A., Collick, A., Allen, M. and Jobling, M. (1993). Minisatellite variant repeat mapping: Application to DNA typing and mutation analysis. In. *DNA fingerprinting: State of the science.* Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. (eds). Birkhaüser Verlag Basel: 125-139.

Armour, J.A.L., Monckton, D.G., Neil, D.L., Crosier, M., Tamaki, T., MacLeod, A. and Jeffreys, A.J. (1993). Minisatellite mutation and Recombination. In. *Chromosomes Today,* **11**. Sumner, A.T. and Chandley, A.T. (eds). Chapman & Hall, London: 337-349.

Neil, D.L. and Jeffreys, A.J. (1993). Digital DNA typing at a second hypervariable locus by minisatellite variant repeat mapping. *Hum. Mol. Genet.* **2**: 1129-1135.

Jeffreys, A.J., Tamaki, T., MacLeod, A., Monckton, D.G., Neil, D.L. and Armour, J.A.L. (1994). Complex gene conversion events in germline mutation at human minisatellites. *Nature Genet.* **6**: 136-145.

ABBREVIATIONS

A	Adenine
Amp-FLP	Amplified Fragment Length Polymorphism
ApoB	Apolipoprotein B
ATP	Adenosine 5′-TriPhosphate
bp, kb, Mb	BasePair; Kilo-, Mega-
BSA	Bovine Serum Albumin
C	Cytosine
cDNA	Complementary DeoxyriboNucleic Acid
CEPH	Centre d'Etude du Polymorphisme Humain
cM	CentiMorgan
dA	DeoxyAdenosine
dATP	2′-deoxyadenosine 5′-triphosphate
dCTP	2′-deoxycytosine 5′-triphosphate
ddATP	2′3′-dideoxyadenosine 5′-triphosphate
ddCTP	2′3′-dideoxycytosine 5′-triphosphate
ddGTP	2′3′-dideoxyguanosine 5′-triphosphate
ddTTP	2′3′-dideoxythymidine 5′-triphosphate
dGTP	2′-deoxyguanosine 5′-triphosphate
DM	Dystrophia Myotonica (myotonic dystrophy)
DNA	DeoxyriboNucleic Acid
DSB	Double Stranded Break
DSBR	Double Strand Break Repair
dTTP	2′-deoxythymidine 5′-triphosphate
EDTA	EthyleneDiamineTetra-acetic acid
FraX	Fragile X syndrome
G	Guanine
g, mg, µg, ng, pg	Grams; Milli-, Micro-, Nano-, Pico-.
GEM	Gap Expansion Model
HCl	HydroChloric acid
HD	Huntingdon's Disease
HLA	Human Leukocyte Antigens
HNPCC	Hereditary Non-Polyposis Colon Cancer
HVR	Hypervariable Region
l, ml, µl	Litre; Milli, Micro

LINE	Long INterspersed repetitive Element
LTR	Long Terminal Repeat
M, mM, µM, nM	Molar; Milli, Micro, Nano
MER	MEdium Reiteration sequence element
min	Minute
MLP	MultiLocus Probe
MVR	Minisatellite Variant Repeat
MVR-PCR	Minisatellite Variant Repeat mapping using the Polymerase Chain Reaction
NIH	National Institute for Health (USA)
nt	NucleoTide
NTS	Non-Transcribed Spacer region
OD	Optical Density
PCR	Polymerase Chain Reaction
PFGE	Pulsed Field Gel Electrophoresis
RFLP	Restriction Fragment Length Polymorphism
RNA	RiboNucleic Acid
rpm	Revolutions Per Minute
rRNA	Ribosomal RiboNucleic Acid
SBMA	Spino-Bulbar Muscular Atrophy
SCA1	Spino-Cerebellar Ataxia type 1
SDS	Sodium Dodecyl Sulphate
sec	Second
SINE	Short INterspersed repetitive Element
SLP	Single Locus Probe
SP-PCR	Small Pool PCR
SSA	Single Stranded Annealing
SSC	Saline Sodium Citrate
STR	Simple Tandem Repeat
T	Thymine
TBE	Tris-borate EDTA
THE	Transposon-like Human Element
Tris	Tris-(hydroxymethyl)-methylamine[2-amino-(2-hydroxymethyl)-propan-1,3-diol]
U.K.	United Kingdom
U.S.A.	United States of America
USCE	Unequal Sister-Chromatid Exchange
UV	Ultra Violet
VNTR	Variable Number of Tandem Repeats
YAC	Yeast Artificial Chromosome

Chapter 1

INTRODUCTION

Variation within genomes

Historical background. Mendel's observations of the segregation of various phenotypic characteristics of sweet peas in 1865, later encapsulated by Sutton (1903) in the chromosomal theory of heredity, are widely acknowleged as providing the foundations of the science of genetics. The pioneering work of Morgan (1915) on naturally occuring *Drosophila melanogaster* mutants confirmed this theory by defining the chromosomal locations and relative positions of genes coding for a large number of characteristics. Advances in chromosome staining techniques, coupled with the discovery of massively amplified polytene chromosomes in *D. melanogaster* salivary glands, allowed the integration of these genetic linkage maps with physical cytological maps of the banding patterns of Drosophila chromosomes, confirming the linear order of genes along chromosomes. These studies also indicated that the material comprising chromosomes was not homogeneous, although it was impossible to tell whether it was the bands, or the interband regions, that corresponded to the genes. The concept of genetic diversity underlying observed phenotypic diversity was therefore well established long before the elegant experiments of Herschey and Chase (1952) finally demonstrated that DNA was the genetic material. It is now widely accepted that phenotypic variation is largely a reflection of underlying DNA sequence variation; however, analysis of the basis of phenotypic variation in humans has always been limited by the available technology, and is poorly understood except in a few well defined cases.

Human genetic disease. The direction of scientific research is dictated by the level of funding certain areas receive, based on their perceived benefits to society. Therefore it is not surprising that the best characterised phenotypic consequences of sequence variation have been those relating to inherited genetic disease, in the hope that by understanding the molecular basis of a disorder new treatments or cures can be developed. The haemoglobinopathies were an obvious starting point for such investigations. These diseases had the simple inheritance patterns of single gene disorders and occured at high frequency in some human populations, requiring expensive treatment by blood transfusion. At the time DNA sequence analysis was not possible, but haemoglobin was easy to extract from blood and analyse by protein electrophoresis. The definitive demonstration that genotype is linked to phenotype through the expression of proteins came when Ingram (1957) showed that the single-gene trait of sickle cell anaemia is caused by the change of a single amino acid in haemoglobin. This was the first variant human phenotype giving rise to genetic disease to be defined at the molecular level. Other haemoglobinopathies were also characterised in this way, but the inaccessibility of most tissues compared to blood, and the relatively low abundance of most proteins compared to haemoglobin, prevented the widespread extension of this approach to the investigation of other genetic diseases.

Direct analysis of the genetic material only became possible in the 1970s, with the arrival of DNA manipulation and sequencing techniques. These enabled development of "reverse genetics" strategies, by which the genetic lesions responsible for disease phenotypes could be identified without any prior knowlege of their molecular basis. A classic example of the successful application of this approach was the identification of the dystrophin gene by analysis of DNA from individuals with muscular dystrophy (Monaco *et al.*, 1986). Since then direct analyses of human DNA have revealed a wide spectrum of genomic variation in terms of the proportion of the genome involved, the frequency at which particular variants are observed, their phenotypic effects, and the reasons for their presence. The majority of such variation is found in non-coding DNA, and is therefore presumably selectively neutral, but until recently the most intensively investigated variants were the relatively rare mutations which cause genetic disease. These range from single base changes, like that responsible for sickle cell anaemia, through microdeletions or insertions, for example the 3bp deletion responsible for cystic fibrosis (Kerem *et al.*, 1989), to larger chromosomal rearrangements, for example the deletions and duplications between the globin genes that cause several commonly observed thalassaemias (reviewed by Weatherall & Clegg, 1982). Large scale chromosomal abnormalities including cytologically detectable gross chromosomal deletions, duplications and aneuploidies also occur. As would be expected few of these are compatible with viability, and those which are are almost exclusively associated with disease phenotypes of varying degrees of severity.

Characterisation of non-coding sequences regulating gene function and expression are of obvious potential clinical importance. Other much more abundant sequences must contribute to the large-scale architecture of the genome and may therefore influence its function on a wider scale. For example, such sequences may be involved in chromosome movement, and their positioning and conformation near transcribed regions may influence the regulation of gene expression. Our work is concerned with the characterisation of highly variable, tandemly repeated, non-coding genetic loci in humans. Variation in such sequences has recently been found to be responsible for, or associated with, a growing number of inherited human genetic diseases, predominantly neurological disorders (reviewed by Mandel, 1994) and cancers (reviewed by Richards & Sutherland, 1994). Studies of such loci may establish the molecular basis of mutation processes resulting in these diseases as well as providing information on the dynamic processes such as mutation and recombination that give rise to, and maintain, genetic variation in man, thus increasing our understanding of the genome as a whole. In addition comparison of variation between individuals has direct applications in forensic science and individual identification, while population analysis of the non-coding 90% of the human genome is also likely to be important in deducing the events that have shaped its evolution.

The human genome project. Recent improvements in mapping and sequencing technology have made the goal of sequencing the most complex genomes in their entirety a viable proposition (Donis-Keller *et al.*, 1987; Weissenbach *et al.*, 1992; Anderson, 1993; Coghlan, 1994; Reed *et al.*, 1994) and there is a growing list of organisms, besides humans, with their own genome projects (see Oliver, 1994). In this "forward genetic" approach the starting point is a genetic map that can be used to order genomic, or more likely cDNA, sequences. This will be followed by identifying the coding regions they contain and then determining their functions. The human genome project was first proposed as a realistic and desirable scientific endeavour over a decade ago (see Watson, 1990) since which time it has gathered considerable momentum, becoming the first biological "big science". The initial stages of genetic mapping are approaching completion and will soon be followed by the move to physical mapping (see Schmitt & Goodfellow, 1994 and accompanying editorial). Once the physical map is complete, large scale

sequencing can begin; however without further improvements in the technologies for large scale sequencing, it is uncertain whether the long term goal of a complete sequence of the human genome by 2005 can be met (see Collins & Galas, 1993).

Even when this monumental endeavour is complete, it will only be the first step on the pathway to greater understanding of the human genome. The eventual sequence will be the consensus product of many laboratories, libraries and individuals, and is bound to be full of small gaps and sequence errors that will be very difficult to detect. Although, it will undoubtedly provide an an invaluable tool for basic genomic investigation, providing a framework around which more detailed analyses can be constructed, it is only by comparative analysis that the genetic basis of phenotypic differences between individuals, and hence the structure and function of the human genome can be fully understood. For this reason it is important to integrate studies of genetic variation and mutation with global mapping approaches. Furthermore, it is unclear how successful this approach will be in the development of new treatments or even cures for human genetic diseases. There has been encouraging progress towards gene therapy with cystic fibrosis, muscular dystrophy, haematological abnormalities and some metabolic diseases (Morsy, *et al.*, 1993; Kay & Woo, 1994), but for many other disorders molecular characterisation offers presymptomatic diagnosis long before hope of eventual treatment. There is widespread ethical debate in the scientific community concerning uses to which such information may be put and the resources which should be channelled into this effort as opposed to other, perhaps more pressing, healthcare needs of mankind.

Variation between genomes

The C-value paradox. Before direct analysis of DNA became possible, investigators used standard biochemical techniques to investigate the properties of DNA. Early experiments, for example those of Mirsky and Ris (1951), measured the amount of DNA in the haploid genomes of a variety of organisms. This is known as the C-value and can be expressed in basepairs (bp) of DNA. As expected this was found to increase with the complexity of organisms ranging from prokaryotes (~10^6bp) to the higher eukaryotes (~10^8-10^{11}bp). If the complexity of an organism reflects underlying DNA complexity then similar organisms would be expected to have similar C-values, however, this is not the case in the higher eukaryotes, where huge variations in C-values were found between organisms of relatively similar complexity and even between similar species. For example the amphibians have C-values ranging from 10^9-10^{11}bp, making some of their genomes 25 times larger than those of mammals. This unexpected observation is known as the C-value paradox. Subsequent estimates of gene number and average gene size compared to genome size, based on well characterised eukaryotic organisms like *D. melanogaster*, led to the conclusion that only a small proportion of most eukaryotic genomes encodes protein products, with the rest composed of apparently superfluous DNA of unknown function.

The paradox resolved. Experiments which measured the speed at which sheared DNA fragments reannealed following denaturation indicated that much of this excess DNA is comprised of sequences that are present at more than one copy in the genome (Britten & Kohne, 1968). Such repetitive DNA sequences are ubiquitous in eukaryotes and comprise a considerable proportion of their genomes. For example, it has been estimated that 20-30% of the human genome is comprised of such sequences (see Schmid & Jelinek, 1982). These multi-copy sequences may be dispersed around the genome, or by contrast, found in tandemly repeated arrays. The remainder of

the excess DNA was later accounted for by single-copy DNA sequences. These are found interrupting the coding sequences of eukaryotic genes (Jeffreys & Flavell, 1977), immediately adjacent to genes, often containing transcriptional and translational regulatory signals, and also in the large tracts of DNA between coding sequences. In fact it has been estimated that genes only make up some 10% of the human genome. For a while this provoked some consternation, since it seemed that the number of genes was too small to code for the expected number of protein products. It was only later that these doubts were allayed by the observations that alternative splicing of gene exons can give functionally distinct proteins, and that genes/proteins can be used in different combinations to perform different tasks.

Repeated DNA sequences in the human genome

Dispersed repeats. The high copy number and distinctive characteristics of some of these repeated sequences made them relatively easy to isolate from the rest of the genome and they have therefore been studied in some detail. Reannealing studies showed that large DNA fragments show faster reannealing kinetics than small ones (Schmid & Deininger, 1975) and it was proposed that this was due to the presence in the larger fragments of a short, highly abundant dispersed repeat sequence. In fact there are several short interspersed repeat elements (SINEs) present in the human genome; the most abundant of these is the ~300bp Alu element (Schmid & Jelinek, 1982). Long interspersed repeat elements (LINEs) are also found, the most prominent being the L1 element, which is 6.4kb long in its fully sized version. These two repeat types each comprise some 5% of the genome (Singer, 1982). Both show features characteristic of transposable elements and are presumed to have achieved high copy number by multiple RNA mediated transposition events (Britten *et al.*, 1988; Paulo di Nocera & Sakaki, 1990). Other classes of human dispersed repeats include the related "O" and "THE" elements (Sun *et al.*, 1984; Paulson *et al.*, 1985) and the "MER" sequences (Jurka, 1990). The remainder of the repetitive fraction of the human genome consists of tandemly repeated DNA.

Satellite DNAs. The ability to separate distinct DNA components of the genome was first demonstrated in early biochemical experiments which used density gradient centrifugation to investigate some of the physical characteristics of mouse DNA. This analysis identified a distinct fraction, comprising 8% of the genomic DNA, that migrated to a different position than the majority of genomic DNA, due to a lower than average GC content (Sueoka, 1961; Kit, 1961). Because this fraction appeared as a characteristic peak near the main band on fractionation it was called "satellite" DNA. It was subsequently shown that this fraction contained highly repetitive DNA analagous to the highly repetitive sequences identified by reassociation kinetics. Since then satellite DNAs have been isolated from a number of higher eukaryotic species, including humans, being differentiated by virtue of the presence of high copy number tandemly repeated sequences of distinctive base composition. *In situ* hybridisation shows that these are generally located in the non-expressed constituitive heterochromatin regions of chromosomes (eg. Pardue & Gall, 1970; Jones & Corneo, 1971). More recently, molecular cloning, sequencing and PCR technologies have led to the identification of additional tandemly repeated sequences. As with satellite DNAs, these seem to be ubiquitous and highly abundant in higher eukaryotes, with a number of such sequences comprising a significant proportion of their genomes. Although not all of these can be distinguished by density gradient centrifugation, they have all come to be known as "simple sequence" or satellite DNAs.

A continuum of tandemly repeated sequences. There is a wide variety of tandemly repeated loci with respect both to the length and sequence of the tandem repeat, and to the average size of the total tandem repeat array. Repeat unit lengths range from 1 nucleotide (nt) to >100nt and tandem array sizes can be anything from ~10bp to ~5 megabasepairs (Mb). This spectrum of sequences is somewhat arbitrarily divided into a number of categories which are distinguished on the basis of average total array size, rather than the length of the individual tandemly repeated unit. However, there are no clear boundaries between these size classes, which represent a continuum of tandem repeat array lengths. The smallest loci are defined as microsatellites, followed by the minisatellites, midisatellites through to the large blocks of satellite DNA, in increasing order of size. In addition to the satellite DNAs there are several specialised tandemly repeated loci, for example the telomeres and the ribosomal RNA genes. Some of these categories are sometimes given other labels, for example the microsatellites are often refered to as simple tandem repeats (STRs). Many tandemly repeated loci, most notably the minisatellites, show polymorphism with respect to repeat copy number, and are refered to as variable number tandem repeats (VNTR) loci. Some of the first minisatellites isolated were also called hypervariable regions (HVRs).

Perhaps a more useful criteria by which to assign loci to a particular category is the technology required to resolve individual alleles. Microsatellites are generally resolved on polyacrylamide gels (eg. Litt & Luty, 1989; Tautz, 1989; Weber & May, 1989), although high percentage agarose gels can resolve larger alleles, particularly those with longer repeat units (eg. Gray, 1991a). Minisatellite alleles are resolved by conventional agarose gel electrophoresis (eg. Wong *et al.*, 1987), while midisatellites and satellites are resolved by pulsed-field gel electrophoresis (PFGE) (eg. Mahatani & Willard, 1990, Oakey & Tyler-Smith, 1990). It should be noted that the later two techniques cannot always distinguish alleles differing by only a single repeat unit.

Microsatellites generally have small repeat units of 1nt to ~5nt and array sizes ranging from ~10bp to 1kb. Overlapping these sequences are the minisatellites, often refered to as VNTRs, their repeat units tend to be longer, ranging from ~4nt to 90nt, as do their tandem arrays, which are generally ~0.5kb to ~30kb. Loci with repeat unit sequences of similar length to minisatellites, but with greater average array sizes, 10kb-500kb are referred to as midisatellites. Finally come the very large satellite arrays; although these can cover 200kb to 5Mb of DNA, their repeat unit lengths are not generally much larger than those of the minisatellites, ranging from 4nt to >100nt. It is therefore possible to find tandem repeat units of a given length in all these size classes. For example a microsatellite locus with alleles of ~88-128bp composed of pentamer repeats have been described in human DNA (Edwards, M.C. *et al.*, 1991), while the mouse minisatellite Ms6-Hm has 2-16kb tandem arrays of a pentamer (Kelly *et al.*, 1989) and a basic 5nt repeat is also found in the human satellites 2 and 3, which are the major simple sequence components of the classical satellites II and III/IV (Prosser *et al.*, 1986).

Satellites, gene clusters and telomeres

The major satellites. The first human satellite sequences were identified as a separate series of bands by density gradient centrifugation (Miklos & John, 1979) and at least 5 classes of human satellite DNA have now been defined (see Tyler-Smith & Brown, 1987). Early hybridisation studies indicated that human satellite DNA was preferentially located in the heterochromatic chromosomal regions, in particular at or near the centromeres (Jones & Corneo, 1971). Sequencing indicated that satellites I-IV were made up large tandem arrays of relatively short (5-

25nt) repeat units (Prosser *et al.*, 1986). In contrast the α-satellite, which is by far the most abundant form in primates, consists of tandem arrays of many thousands of 171bp repeat units. This is the predominant sequence in centromeric hererochromatin, spanning 1-5Mb across each centromere (Willard, 1991) and forming ~5% of the genome (reviewed by Willard, 1990). Variant repeat units within these α-satellite arrays give rise to a number of chromosome-specific higher-order repeats that have been proposed to have a functional role in centromeric processes such as homologue recognition, pairing and segregation during meiosis (Willard, 1990). Another human satellite DNA, the β-satellite is has been less well characterised but it contributes a considerable proportion of the short arm of the acrocentric chromosomes and consists of tandem arays of a 68bp repeat unit. These can extend for hundreds of kilobases, again showing higher-order repeat structures (Greig & Willard, 1992).

rDNA and other gene clusters. The short arms of the 5 human acrocentric chromosomes also contain the tandemly repeated arrays of 18S and 28S rRNA genes, which constitute the nucleolar organiser regions. Each chromosome has ~40 repeats of a 44kb monomer which is made up of a 13kb transcribed region, combining the 18S and 28S rRNA transcription units and a non-transcribed spacer (NTS), which itself contains a number of tandemly repeated sequences (see Arnheim *et al.*, 1980). In sharp contrast to the transcribed region the NTS shows extreme interspecies variation in terms of both sequence and length, even between closely related species such as the great apes (Arnheim *et al.*, 1980; Qu *et al.*, 1991). The 5S rRNA genes are also encoded as a tandem repeat, with an array of some 2000 repeat units located on chromosome 1 (see Timofeeva *et al.*, 1993). Other genes, often with related functions, are also arranged as linear clusters. Although these are presumed to have arisen through a series of duplications, followed by rounds of unequal exchange, they are not truly tandemly repeated. Unlike the rRNA genes these are separated by greatly diverged sequences of varying length. Examples of such multigene families gene clusters in man include the histone genes (Heintz *et al.*, 1981) the immunoglobin genes (see Baltimore, 1981) the haptoglobin genes (Maeda & Smithies, 1986) and the α- and β-globin genes (see Weatherall & Clegg, 1982).

Telomeres. The highly conserved sequences defining the termini of most eukaryotic chromosomes are composed of tandem repeat arrays of a specialised hexamer repeat called telomeres (reviewed by Blackburn, 1991). All telomeric DNA consists of tandem repeats of a G-rich sequence, with the G-rich strand orientated 5′-3′ towards the terminus and potruding 12-16bp beyond the complementary strand. This allows binding of the RNA template of a specialised ribonucleoprotein reverse transcriptase, "telomerase" that adds telomere repeats to chromosome ends (Greider & Blackburn, 1989). Without this priming mechanism chromosomes would be progressively shortened during successive rounds of replication in the germline. There is some evidence that this may occur in human somatic tissues which seem to lack telomerase activity, and that telomere shortening leading to chromosome instability might be responsible for cell senescence (Harley *et al.*, 1990; Hastie *et al.*, 1990). On average human telomeres comprise 10kb tandem arrays arrays of a consensus TTAGGG repeat (Moyzis *et al.*, 1988), but array lengths and telomere repeat sequences are highly heterogenous, even within one indivuidual, such that Southern blot length analysis with a telomere specific probe produces a smear rather than a distinct band (see Hastie & Allshire, 1989). Interstitial tracts of telomere like sequence have also been identified in human chromosomes (Hastie & Allshire, 1989). These may result from internalisation of telomeric sequences, possibly due to chromosome rearrangements; for example, interstitial telomeric repeats on chromosome 2 appear to have resulted from the fusion of two primate acrocentric chromosomes (Allshire *et al.*, 1989). However, these could also simply reflect independent expansion of tandemly arrayed sequences with homology to the telomere repeat unit.

Micro-, mini-, and midi-satellites

Microsatellites (STRs)

The $(CA)_n$ dinucleotide repeats were the first microsatellite loci to be described (Miesfeld *et al.*, 1981). Their ubiquity in eukaryotic genomes prompted the proposal of a number of possible functions for these loci. These included; serving as hot-spots for recombination or gene conversion, gene regulation, chromatin folding, telomere formation and X-inactivation (see Gray, 1991b). Since then many $(CA)_n$ loci, and also others with different repeat units, have been isolated using a variety of techniques. They have been located fortuitously during sequencing of genomic regions of interest (eg. Kremer *et al.*, 1991) and by sequencing random clones (Weissenbach *et al.*, 1992). A more direct approach has been detection by hybridisation to tandem repeat oligonucleotide probes (eg. Riggins *et al.*, 1992). A number of STR loci, particularly those with A-rich repeat units, are associated with retroposon tails, usually Alu (Economou *et al.*, 1990; Gray, 1991c; Beckman & Weber, 1992; Armour *et al.*, 1994). STRs are relatively easy to clone and sequence and do not show a tendency to cluster in any particular region of the genome. The development of PCR (Saiki *et al.*, 1988) enabled simple and rapid determination of repeat array length and allelic variability at STR loci (Tautz, 1989; Weber & May, 1989; Litt & Luty, 1989), which have rapidly become the subject of much research. The range of repeat unit sizes and even genomic distribution of STRs suggest that rather than being evolutionarily conserved because they serve a ubiquitous biological function, these sequences are formed frequently and independently by the same universal genomic mechanisms (Tautz & Renz, 1984). The following list is not exhaustive, but gives some illustrative examples of the different types of human STR loci.

Mononucleotides. The best known mononucleotide repeats are the poly(dA) tracts associated with retroposon tails. These can show significant degrees of variability with respect to allele length which can be analysed by PCR amplification, followed by polyacrylamide gel electrophoresis (Economou *et al.*, 1990). A polymorphic $(A)_n$ tract in intron 13 of the amyloid precursor gene on chromosome 21 that can be typed in the same manner has also been reported (Mant *et al.*, 1991).

Dinucleotides. Microsatellites with all nucleotide combinations have been identified, but the most abundant and extensively catalogued short repeats in the human genome are the $(CA)_n$ dinucleotide tandem arrays. The existence of large numbers of these repeat loci dispersed throughout eukaryotic DNA has been known for some time (Miesfeld *et al.*, 1981; Hamada *et al.*, 1982; Sun *et al.*, 1984; Tautz & Renz, 1984) and it has been estimated that there are 50,000-100,000 in the human genome (Miesfeld *et al.*, 1981; Hamada & Kakunaga, 1982) occuring every 30kb in euchromatic DNA (Stallings *et al.*, 1991). PCR analysis of these loci showed that they are generally rather small, typically $(CA)_{10-30}$ and often polymorphic in repeat copy number, with heterozygosities up to 90% (Litt & Luty, 1989; Weber & May, 1989; Tautz, 1989; Weber, 1990).

Trinucleotides. Many trinucleotide repeats have now been isolated, and analysis shows that they are common, frequently polymorphic and seem to be well dispersed throughout the genome (Edwards *et al.*, 1992). There has recently been a flurry of research activity directed to the localisation and characterisation of trinucleotide repeats, following the discovery that expansions of triplet repeat arrays at some of these loci are associated with a rapidly growing list of human inherited genetic diseases (Kremer *et al.*, 1991; Fu *et al.*, 1991; reviewed by Caskey *et al.*,

1992; see also Richards & Sutherland, 1992, 1994; Mandel, 1994). Expanded triplet repeats have been found in the untranslated regions of several genes, for example, 5′ to FMR1 at the fragile X locus (Kremer *et al.*, 1991) and 3′ to the myotonic dystrophy (DM) DM-kinase gene (Brook *et al.*, 1992; Fu *et al.*, 1992). Expansion of triplet repeats can also occur in expressed sequences, coding for polyamino acid regions, for example the polyglutamine tracts of the genes involved in spino-bulbar muscular atrophy (SBMA), Huntingtons's disease (HD), and spino-cerebellar ataxia type 1 (SCA1) (reviewed by Nelson, 1993). The triplet repeat disease loci catalogued to date all have $(CAG)_n$ repeats, except for FRAXA and FRAXE at the fragile X locus, where the repeat is CCG. These loci are polymorphic and can exhibit a high frequency of germline and somatic mutation (reviewed by Richards and Sutherland, 1992; Caskey *et al.*, 1992). The search for additional repeats of these trinucleotides in the genome has used database searches of the coding sequences of genes and screening of cDNA libraries with synthetic repeat unit probes (eg. Riggins *et al.*, 1992; Li *et al.*, 1993). A more generalised approach to isolating triplet repeat loci by hybridisation selection and PCR has been recently described (Armour *et al.*, 1994). Other triplet repeat sequences do exist and may also be polymorphic (Edwards, A. *et al.*, 1991), but have been the subject of less intensive investigation since no disease associations have yet been found with these loci. The most abundant trinucleotides AAT, AAC and AAG are usually associated with retroposon tails (Beckman & Weber, 1992) and are therefore specifically omitted from most triplet repeat screens (eg. Armour *et al.*, 1994).

Tetranucleotides. Among the first reported STRs were GATA/GACA repeats, originally identified and isolated from snake satellite DNA (Epplen *et al.*, 1982). These, and other quadruplet repeats, were later found to be distributed throughout the eukaryotes and several have been characterised. These have generally been detected in a similar manner to triplet repeats. Those isolated by virtue of association with genes of interest include, the $(CTTT)_n$ sequence closely linked to the retinoblastoma gene (Yandell & Dryja, 1989), the $(TGGA)_n$ repeat located 5′ to the myelin basic protein gene (Boylan *et al.*, 1990) and a GATA repeat in intron 40 of the von Willebrand factor gene (Peake *et al.*, 1990). Others have been detected by naturally occuring (Gray, 1991d) and oligonucleotide probes (eg. Melis *et al.*, 1993), and also by hybridisation selection (Armour *et al.*, 1994). Some human tetramer repeats show a high degree of polymorphism for example the AAAG repeat of a β-actin related processed pseudogene (Polymeropoulos *et al.*, 1992) and the DXS981 $(TATC)_n$ locus (Mahtani & Willard, 1993), and recent evidence has suggested that quadruplet repeat loci may be more likely to be highly polymorphic than triplet and dinucleotide repeats (Gray, 1991a; Armour *et al.*, 1994). In general, quadruplet repeat loci appear to be at least as polymorphic as triplets, and distributed at a similar frequency, such that both of these types of locus are interspersed every 300-500kb throughout the human genome (Edwards, A. *et al.*, 1991).

Pentanucleotides. Fewer examples of human polymorphic pentamer repeats have been reported in the literature. Those that have been detected and characterised were associated with genes under investigtion and have A-rich sequences suggestive of derivation from retrotransposon tails. For example an $(AAAAG)_n$ repeat at the CD4 locus (Edwards, M.C. *et al.*, 1991), an $(AAAAT)_n$ locus on chromosome 19P (Chen *et al.*, 1993) and an $(ATAAA)_n$ repeat in the 5′ flanking DNA of the glutathione-S-transferase PI-gene (Harada *et al.*, 1994). There have been no concerted efforts to identify tandem repeats of this type, presumably because they have the potential to be composed of more diverse sequences than smaller STRs, making it more expensive and difficult to design repeat unit oligonucleotides for their detection. There is therefore no particular reason to presume that there are fewer pentamer repeats in the genome than STRs with smaller repeats.

Minisatellites

Serendipitous discovery. Human minisatellites were initially discovered by chance, usually being isolated from essentially random single-copy clones of genomic regions containing a gene of interest. The first was identified as a highly polymorphic RFLP for which most people in the population were heterozygous (Wyman & White, 1980). Other multiallelic loci were soon discovered, including loci 3′ to the α-globin gene (Higgs *et al.*, 1981), in the ζ-globin intron and between the ζ and pseudo-ζ globin genes (Goodbourn *et al.*, 1983), and in non-coding regions of the insulin and H-*ras* genes (Bell *et al.*, 1982; Capon *et al.*, 1983). It was at the insulin locus that it was first demonstrated by sequence analysis that the molecular basis for such polymorphism was variation in the copy number of a short tandemly repeated sequence (Bell *et al.*, 1992). This was subsequently shown to be a common feature of all these loci (Capon *et al.*, 1983; Wyman *et al.*, 1986). The number of minisatellites detected in this manner continues to grow, for example a novel tandemly repeated sequence in an intron of the glucose phosphate isomerase gene has recently been reported (Faik *et al.*, 1994).

DNA fingerprints. A major advance in the systematic isolation of additional highly variable loci came with the discovery that some tandemly repeated probes, derived originally from a tandem repeat sequence in the first intron of the human myoglobin gene, detected a large number of other highly polymorphic loci in human DNA when hybridised at low stringency to Southern blots of genomic DNA (Jeffreys *et al.*, 1985a). It was estimated that the two most informative probes, 33.6 and 33.15, each detected ~1000 loci, 10 and 20 of which had large alleles that could be well resolved on agarose gels. These apparently unlinked, polymorphic loci, were named minisatellites (Jeffreys *et al.*, 1986). The composite profiles generated by simultaneous detection of these loci were shown to be so highly variable as to be individual specific, and they were therefore called "DNA fingerprints" (Jeffreys *et al.*, 1985b). The potential applications of multilocus DNA fingerprinting in the fields of individual identification, kinship testing, forensic medicine, monitoring tumor progression and tissue matching, were obvious and quickly realised (Jeffreys *et al.*, 1985b, 1985c; Gill *et al.*, 1985; Thein *et al.*, 1986, 1987; Thacker *et al.*, 1988). See page 18, "The applications of polymorphic loci in forensic analysis and individual identification", for a more detailed review of human DNA typing technology.

Following this initial work other naturally occuring tandem repeat probes were found to detect multiple variable loci when hybridised at low stringency; for example the α-globin 3′ HVR (Fowler *et al.*, 1988) and a tandem repeat from the phage M13 genome (Vassart *et al.*, 1987). Synthetic tandemly repeated oligonucleitides and polymers of random oligonucleotides have also been used to detect a number of these loci (Ali *et al.*, 1986; Nakamura *et al.*, 1988a, Vergnaud, 1989; Vergnaud *et al.*, 1991; Epplen *et al.*, 1991). However, many of the human minisatellites detected by all of these probes were subsequently shown to be among those originally detected by 33.6 and 33.15 (Armour *et al.*, 1990, 1992a). DNA sequences that detect multiple loci under low stringency hybridisation conditions are collectively refered to as multi-locus probes (MLPs).

MLPs have also been successfully used to detect multiple polymorphic loci in other species, including other primates (Dixson *et al.*, 1992), mice (Jeffreys *et al.*, 1987a), dogs and cats (Jeffreys & Morton, 1987), cattle (Georges *et al.*, 1991), pigs (Signer & Jeffreys, 1993), birds (Burke & Bruford, 1987; Wetton *et al.*, 1987; Signer & Jeffreys, 1993) and even plants (Dallas, 1988; Rogstad *et al.*, 1988). DNA fingerprinting has thus found

applications in the verification of identity and parentage in economically important species, as well as in the establishment of family relationships and the extent of inbreeding in population biological studies and in zoological specimens (reviewed by Burke *et al.*, 1991; see also, Packer *et al.*, 1991; Signer & Jeffreys, 1993).

Deliberate isolation. Several approaches have been used to isolate the individual variable loci detected *en masse* in DNA fingerprints. The first and most direct cloned alleles selected from a human DNA fingerprint (Wong *et al.*, 1986). However, in terms of isolating large numbers of different loci this technique was rather cumbersome and progressively more efficient strategies were later developed. The most sucessful of these have used hybridisation screening with multi-locus DNA fingerprint probes to isolate clones of individual loci from genomic libraries including: human λ libraries screened with 33.15 and 33.6 (Wong *et al.*, 1987), human cosmid libraries (Nakamura *et al.*, 1987a, 1988) and ordered array charomid libraries from both animals and humans (Armour *et al.*, 1990; Hanote *et al.*, 1991; Burke *et al.*, 1991; Signer *et al.*, 1994). The majority of cloned minisatellites detect only their cognate locus when used as probes in high stringency Southern blot hybridisations, and are referred to as single locus probes (SLPs) under these conditions. Like MLPs these have also found applications in individual identification and forensic analysis (see page 18) as well as being used in linkage mapping (see page 16), transplant monitoring (Hutchinson *et al.*, 1989) and the detection of allele loss in tumours (eg. Mathew *et al.*, 1987; Solomon *et al.*, 1987; Vogelstein *et al.*, 1989).

A representative and random sample? The number of human minisatellite loci has been estimated at ~1500 per haploid genome (Braman *et al.*, 1985; Jeffreys, 1987), based on the number of highly polymorphic loci recovered from a large-scale screen of λ clones of human DNA for polymorphism (Schumm *et al.*, 1985). However, minisatellite loci are clearly underrepresented in standard genomic libraries (Wyman *et al.*, 1985; Kelly *et al.*, 1989; Armour *et al.*, 1990), suggesting that this figure may be an underestimate. Furthermore, because monomorphic minisatellites have little to offer for genetic analysis they were not taken into account in these studies, despite the presence of an apparently considerable number in the human genome (Armour *et al.*, 1990). Our view of minisatellites is therefore subject to considerable bias towards the highly polymorphic and the easily isolated loci. Minisatellites can be extremely refractory to cloning using standard vectors in *E. coli* hosts (Wyman *et al.*, 1985, Wong *et al.*, 1986); like other tandemly repeated sequences they are prone to gross rearrangements, usually resulting in deletion of much of the repeat array (Brutlag *et al.*, 1977). This problem has been overcome to a large extent by the use of charomid vectors (Saito & Stark, 1986), which are relatively insensitive to reduction of insert size due to loss of repeats and hence ideally suited to cloning minisatellites (Armour *et al.*, 1990). However, it seems that many of the loci most amenable to cloning have already been found and that, without the development of alternative strategies, future attempts to isolate additional loci by this approach may yield diminishing returns (Armour, 1990a).

Genomic distribution. Although our current impression may be subject to cloning bias, the autosomes all appear to have similar numbers of hypervariable minisatellites. Analysis of the genomic location of these loci has shown that they are not evenly dispersed throughout the genome, but tend to cluster in the subtelomeric regions of human chromosomes (Royle *et al.*, 1988; Nakamura *et al.*, 1988a; Armour *et al.*, 1990; Vergnaud *et al.*, 1991). Where these regions have been characterised in detail, for example 16p (Jarman & Higgs, 1988; Jarman & Wells, 1989) and the pseudoautosomal X-Y pairing region (Cooke *et al.*, 1985; Rouyer *et al.*, 1986; Page *et al.*, 1987),

very high densities of minisatellite loci have been found. In a number of instances this phenomenon has resulted in the fortuitous isolation of two different minisatellite arrays on a single short cloned DNA fragment (Royle *et al.*, 1988; Armour *et al.*, 1989a; Armour & Jeffreys, 1991; Vergnaud *et al.*, 1991).

Locus characteristics. Repeat unit lengths at the human minisatellite loci characterised to date range from a minimum of 9bp at the MS1 locus (Wong *et al.*, 1987) to maximum of 90bp at the MS607A locus (Armour, 1990b). Total minisatellite tandem repeat array sizes are typically in the order of 0.5-30kb. While some minisatellite loci appear to be monomorphic with respect to allele length in the populations studied (Armour *et al.*, 1990), others frequently exhibit very high variability. These hypervariable minisatellites can have allele length heterozygosities >90% making them the most polymorphic loci yet identified in the human genome (Wong *et al.*, 1987, Vergnaud *et al.*, 1991). The differences in array length between different alleles are due to allelic variation in the copy number of their tandem repeat unit. For this reason minisatellites are sometimes refered to as hypervariable regions (HVRs) or variable number tandem repeat loci (VNTRs). This variation can be assayed using a restriction enzyme which cuts outside the tandemly repeated region, followed by Southern blot hybridisation using the minisatellite sequence as a probe at high stringency (eg. Wong *et al.*, 1986). The extreme level of allelic diversity with respect to tandem repeat copy number is caused by high *de novo* germline mutation rates to new length alleles at these loci (Jeffreys *et al.*, 1988a). Variation within the tandemly repeated sequence has been found at all hypervariable human minisatellites characterised to date (Bell *et al.*, 1982; Capon *et al.*, 1983; Owerbach & Aagaard, 1984; Wong *et al.*, 1986, 1987), resulting in internal variation in the interspersion of different repeat unit types between alleles (Owerbach & Aagaard, 1984; Wong *et al.*, 1986, 1987). Assaying this variation is a powerful alternative to distinguishing alleles by differences in length and has proved highly informative in studies of minisatellite allelic diversity and the mechanisms that maintain variability (Jeffreys *et al.*, 1990, 1991a; Armour *et al.*, 1993; Neil & Jeffreys, 1993; Desmarais *et al.*, 1993; Arnot *et al.*, 1993; Buard & Vergnaud, 1994). Analysis of minisatellite internal variation is the basis of all the research described in this thesis and is fully introduced in Chapters 3 and 4.

Structure or function? There is a wide variety of sequence variation between repeat units of different loci, with examples of both AT-rich and GC-rich minisatellites (Wong *et al.*, 1987; Huang & Breslow, 1987; Vergnaud *et al.*, 1991). However, most of the early minisatellites isolated following identification by the 33.6 and 33.15 probes were GC-rich and exibited profound purine-pyrimidine strand asymmetry. The observation that these loci shared an 11-16bp consensus sequence, or "core", with homology to the χ recombination signal of *E. coli*, coupled with their location in highly recombinogenic regions of the genome and their tandemly repeated structure and high mutation rates, let to the proposal that minisatellites were recombination hotspots, with the core sequence functioning as a recombination signal in human DNA (Jeffreys *et al.*, 1985a, 1988). However, more recent evidence, particularly the discovery of AT rich minisatellites (Huang & Breslow, 1987; Vergnaud *et al.*, 1991) and the failure to observe exchange of flanking sequences during minisatellite mutation, (Wolff *et al.*, 1988; 1989) has dampened this speculation. The possible significance of the core is further discussed in Chapter 7. GC-rich minisatellites with little homology to the core were also found; for example, the insulin-HVR, the three α-globin cluster HVRs and the c-Ha-Ras HVR (HRAS1). Interestingly, although these were isolated independently they also show some similarity in repeat unit sequence, suggesting either that certain classes of GC-rich sequence may be predisposed to minisatellite formation, or that there may be functional conservation between these loci (Jeffreys *et al.*, 1987b).

Minisatellites and human genetic disease. Mutations in non-coding minisatellites located near genes have recently been implicated in human genetic disease. These include the HRAS1 minisatellite located 1kb downstream from the proto-oncogene coding sequence (Krontiris *et al.*, 1993) and the minisatellite located immediately 5′ to the insulin gene (Lucassen *et al.*, 1993). The HRAS1 minisatellite has a limited number of common alleles, as would be expected at a minisatellite with a low mutation rate or under the influence of strong selection, and a larger number of rare alleles derived from these by mutation (Kasperczyk *et al.*, 1990). A number of the rare HRAS1 minisatellite alleles have been found to be associated with several common human cancers and there is evidence to suggest that this is due to functional properties of the minisatellite, rather than linkage disequilibrium with pathogenic regions of the HRAS1 locus, or other nearby potential disease loci (Krontiris *et al.*, 1993). This minisatellite has been shown to bind members of the rel/NF-κB family of transcription factors *in vitro* and displays pleiotropic transcriptional regulatory activity that is promoter- and cell-type-specific (Kontiris & Green, 1993). It also possesses position- and orientation-independent enhancer activity which, significantly, is influenced by allele length, for example one rare HRAS1 minisatellite allele has twofold greater enhancer activity than some of its related counterparts (Kontiris & Green, 1993). It has been proposed that mutations interfering with the the binding of regulatory proteins to minisatellite alleles are pathogenic, whether or not the minisatellite has a physiological role in the regulation of HRAS1 (Krontiris *et al.*, 1993). A similar picture is emerging at the insulin locus where alleles of the insulin minisatellite within a certain size range have been associated with insulin dependent diabetes mellitus (Lucassen *et al.*, 1993). This minisatellite lies 365bp from the start of transcription for insulin, immediately adjacent to defined regulatory sequences, prompting speculation that allele length variation may have a direct effect on insulin gene regulation, perhaps by effecting the binding of a regulator of insulin transcription (Lucassen *et al.*, 1993). This was supported by studies which showed that the strand assymetry and base composition of the insulin minisatellite may affect *in vivo* chromatin conformation (Hammond *et al.*, 1992). There are other minisatellites that have binding sites for transcription factors and are associated with genes, for example, in the intron separating diversity and joining segments of the human immunoglobulin heavy-chain gene, and in an intron of the interleukin-1α gene. Collectively these observations suggest that a subset of minisatellite loci may influence neighbouring gene function and make a direct contribution to human genetic disease.

Coding minisatellites. Although the majority of hypervariable minisatellites that have been cloned to date are non-coding (Wong *et al.*, 1987; Nakamura *et al.*, 1987a, 1988; Armour *et al.*, 1990; Buard & Vergnaud, 1994) a few variable coding minisatellites have been identified in humans and other species. For example the *per* gene of *Drosophila melanogaster*, which is involved in the control of pupal-to-adult circadian eclosion rhythms, adult sleep-wake locomotor activity rhythms and aspects of the song cycle (reviewed by Dunlap, 1993), contains repeated sequences coding for a reiterated threonine/glycine motif that shows length polymorphism. The number of repeats increases with latitude, suggesting that this region may be involved in temperature compensation of the *D. melanogaster* biological clock (Costa *et al.*, 1992). The most variable human coding minisatellite known is at the mucin protein (MUC1) locus on chromosome 1. This has an allele length heterozygosity >80% at the DNA level, which is translated into highly variable electrophoretic mobility of its protein product (Swallow *et al.*, 1987). This polymorphism is due to the presence of variable numbers (21-125) of 60bp (20 amino acid) coding repeats (Gendler *et al.*, 1990). Other human coding minisatellites are generally much less variable than their non-coding counterparts, presumably due to selective functional constraints. These include other mucins (Gum *et al.*, 1989), proline rich proteins (Azen *et al.*, 1984) and the involucrin gene (Simon *et al.*, 1991). The collagen gene family

also contain extensive tandemly repeated regions of a 9bp repeat (reviewed by Vuorio & De Crombrugghe, 1990), however no polymorphism in the copy number of these repeats has been reported. These genes code for essential structural proteins found in all eukaryotes and their structures show remarkable conservation, even between distantly related species. Where *de novo* length variants have been identified they are associated with severe dominant disease phenotypes (Wallis *et al.*, 1989; Hawkins *et al.*, 1991). These observations suggest that although collagen genes may be prone to similar mutation mechanisms as other tandemly repeated loci, they are maintained in a non-polymorphic state by intense selective pressure. An interesting coding minisatellite that has recently come to light is the 24bp octapeptide repeat of the prion protein gene. This protein has been implicated as being the causative agent of a bizarre group of neurodegenerative conditions known as the prion diseases, or spongiform encephalopathies (reviewed by Prusiner, 1991). A variety of phenotypes have been described in humans, classically encompassing "kuru", Creutzfeldt-Jakob disease (CJD) and Gerstmann-Straussler Scheinker syndrome (GSS). Changes in octanucleotide repeat number have been implicated as the cause of a number of cases of inherited prion disease (reviewed by Palmer & Collinge, 1993).

Midisatellites

This category of locus fills the gap in the heirarchy of tandem repeats between the extremely large satellite arrays and the much smaller minisatellites. Relatively few midisatellite sequences have been reported, perhaps reflecting the greater difficulties in cloning and characterising larger loci. Two that have been described are the D1Z2 locus, which has 250-500kb tandem arrays of a 40bp repeat and is located near the telomere of chromosome 1 (Nakamura *et al.*, 1987b), and a 61bp tandem repeat in the pseudoautosomal region of the X and Y chromosomes with array sizes of 10-50kb (Page *et al.*, 1987). Both of these are polymorphic with respect to allele length and internal sequence. Two further midisatellites with array lengths of 50-200kb, but a relatively short 9bp repeat unit, have also been isolated from human DNA (Gray, 1991e). Again both of these loci were highly polymorphic in internal structure and total length.

A true reflection of the range of tandemly repeated loci? There are few reports in the literature of large arrays of any of the microsatellite tandem repeat array types listed above. Only one dinucleotide locus with more than 30 repeats has been described, this is located in a subtelomeric region of chromosome 16p and has alleles with 30-450 imperfect CA repeats (Wilkie & Higgs, 1992). This locus is surprisingly uninformative, particularly in light of the observation that variability increases with allele length for shorter CA repeats (Weber, 1990). Two very large, and again imperfect, triplet repeat loci have also been recently identified (Armour, 1994). These are large enough to be defined as minisatellites; one has alleles from 1.8kb-9kb, while the other has alleles from 1.6kb-3.5kb. Both have observed heterozygosities of 80% and are composed largely of blocks of CCT and CCA triplet repeats. All of these loci, together with the large expanded triplet repeats associated with human disease, are extremely difficult to propagate in a bacterial host, and are also refractory to PCR amplification (Wilkie & Higgs, 1992; Fu *et al.*, 1991; Brook *et al.*, 1992; J. Armour, personal communication). It therefore seems probable that the apparently restricted allele length distributions of loci with small tandem repeats may be an artifact of current screening and isolation procedures. If this is a general phenomenon, also explaining the apparent paucity of midisatellites, these observations may well reflect a difficulty in isolating tandem repeat loci with these characteristics, rather than indicating that they are not common in the genome.

The evolution and turnover of tandem repeat loci

Investigating mutation. Tandemly repeated sequences in general are highly variable components of the human genome, often exhibiting a high degree of polymorphism both in terms of array length and internal structure with respect to repeat unit variants. By analysing these characteristics it is possible make some initial inferences as to the processes that may be operating to generate and maintain variation at these loci. Direct analysis of individual mutation events enables even firmer deductions concerning mutational mechanisms to be made. The germline mutation rate at the most polymorphic tandem repeat loci, in particular the hypervariable minisatellites, can be so high that *de novo* germline mutation events can be identified in pedigree analyses, allowing mutation rates to be quantified (Jeffreys *et al.*, 1988a; Vergnaud *et al.*, 1991; Kwiatowski *et al.*, 1992; Wiesenbach *et al.*, 1992; Mahtani & Willard, 1993). Because the size of minisatellite loci made them particularly amenable to cloning in bacteriophage vectors and analysis by Southern blot hybridisation these were the first highly polymorphic loci isolated (Jeffreys *et al.*, 1985a; Wong *et al.*, 1986, 1987; Nakamura *et al.*, 1987a). It is for these reasons that mutation rates and mechanisms have been studied in most detail at the hypervariable minisatellite loci. Since the molecular processes operating at tandem repeat loci are likely to overlap the artificial boundaries we have defined in this continuum of sequences, the observations made at minisatellites may serve as a useful paradigm for studies of the generation of polymorphism at other tandemly repeated loci.

Possible mechanisms. A variety of mutational mechanisms that may operate in the germline to alter the number of repeats in tandemly repeated blocks of DNA have been proposed. These include interallelic processes, like unequal recombination and gene conversion (Dover, 1982), and intraallelic mechanisms, such as replication slippage (Tautz *et al.*, 1986; Levinson & Gutman, 1987), unequal sister chromatid exchange (USCE) (Smith, 1976) and deletion by intramolecular recombination. It now seems that the spectrum of tandemly repeated loci is a reflection of a range of mutation processes that vary between loci, with relative contributions determined by the precise nature and environment of each locus. The mutation of minisatellites and other tandemly repeated sequences is discussed in detail in Chapters 6 and 7. However, I will briefly summarise the processes that have been proposed to operate at tandemly repeated loci, again making broad divisions between categories of these sequences, while bearing in mind the caveat that this may not reflect true biological distinctions.

Microsatellites. Higher order STR repeats are frequently polymorphic in humans (Weber, 1990; Caskey *et al*, 1992; Edwards *et al.*, 1992) and have been shown to undergo spontaneous germline mutation to new length alleles (Kwiatkowski *et al.*, 1992; Weissenbach *et al.*, 1992; Mahtani & Willard, 1993; Weber & Wong 1993; Banchs *et al.*, 1994). The recent association of microsatellite instability with a growing number of human genetic diseases has prompted considerable speculation as to the mutational mechanisms operating at these loci, with replication slippage currently the most favoured (reviewed by Wells & Sinden 1993; Kunkel, 1993). This intraallelic process can operate at the level of single or multiple nucleotide repeats, resulting in the deletion or duplication of repeat units depending on the direction of slippage. Several features of human STRs are consistent with those predicted for mutation by this mechanism. For example, it has been noted that polymorphism in dinucleotide repeats, a reflection of underlying instability, is proportional to perfect repeat copy number (Weber, 1990). Polymorphism also appears to increase with increased numbers of perfect repeats (Kunst & Warren, 1994), and/or array size and repeat unit length in general at other human STR loci (Gray, 1991a; Schlotterer & Tautz, 1992; Armour *et al.*,

1994), although it is not clear whether these differences reflect intrinsic differences in mutability between STR loci, or are due to other factors, such as chromosomal location. There is also growing experimental evidence that human microsatellites mutate predominantly by slippage, however, USCE cannot be ruled out (reviewed by Lustig & Petes, 1993; Richards & Sutherland, 1994; 1992; see also Strand *et al.*, 1993 and Chapter 7).

Minisatellites. Initial analyses of the general properties of minisatellite structure and mutation gave conflicting indications as to the nature of the mutation mechanisms that may be involved. At first observations provided circumstantial evidence of a role for unequal homologous recombination and led to speculation that minisatellites may be recombination hotspots in the human genome (Jeffreys *et al.*, 1985a, 1988). However, subsequent analyses of individual minisatellite mutation events failed to show a single example of exchange of flanking markers, as would be expected in simple homologous recombination (Wolff *et al.*, 1988, 1989). Early surveys also indicated that gains and losses of repeat units occurred with approximately equal frequency (Jeffreys *et al.*, 1988a), ruling out a major role for intramolecular recombination, which can only result in the loss of repeat units, in minisatellite mutation. Unfortunately, the techniques used in these studies could not be used for a widespread and detailed survey of minisatellite mutation, and the resolution they afforded was generally too poor to provide information concerning mutation processes at the repeat unit level. This meant that they provided no information concerning the involvement of other potential mutation mechanisms, for example, replication slippage, USCE and gene conversion. The recent development of a novel technique for analysis of minisatellite internal structure, and hence the patterns of repeat unit turnover accompanying minisatellite mutation, has enabled us to examine these possibilities directly (Jeffreys *et al.*, 1990, 1991a, 1994). Initial results provide evidence for the involvement of all three aforementioned mechanisms in the turnover of repeat units at minisatellite loci, often resulting in a complex mutational profile. This study has been one of the major thrusts of research in this laboratory over the last four years and has dramatically advanced our understanding of minisatellite mutation processes. It also comprises a major part of this thesis and is discussed in greater detail in Chapter 6.

α-Satellite. The size and type of α-satellite blocks at homologous loci can often show extreme levels of individual variation (eg. Oakey & Tyler-Smith, 1990; Ge *et al.*, 1992; Haaf & Willard, 1992). As with minisatellites, studies of internal variation within satellite blocks have been used to infer the turnover processes operating at these loci (reviewed by Willard & Waye 1987). Features such as the clustering of the RFLP variants that define chromosome-specific higher-order repeats appear to suggest that the homogeneity of satellite blocks is maintained by short range intrachromosomal processes, probably USCE. However, the ubiquity of α-satellite on all human chromosomes and the existence of shared subfamilies between non-homologous chromosomes implies that interchromosomal exchange has also played an evolutionary role (Dover, 1982).

rDNA. The large size of the repeat unit and difficulties in cloning these loci in YACs (Labella & Schlessinger, 1989) have made it difficult to find higher order repeat structures at these loci. However, species-specific variants have been identified and found to have become fixed at different loci within a genome, suggesting a relatively high rate of exchange between nonhomologous chromosomes over a short evolutionary period (Arnheim *et al.*, 1980). rDNA evolution has been proposed to involve unequal exchange between sister-chromatids and also homologous and non-homologous chromosomes, giving rise to the spread of variant repeats and generation variation in the length of the NTS (see Arnheim *et al.*, 1980; Dover, 1982).

The applications of polymorphic loci in genetic analysis

Linkage analysis. Polymorphism in DNA structure provides the basis of genetic analysis. The ability to distinguish between two or more different alleles at a particular locus defines it as an informative genetic marker. Because of the linear arrangement of DNA in eukaryotic chromosomes, two genetic markers on the same chromosome can be said to be physically linked to one another (syntenic). Such markers are vital for the construction of genetic linkage maps, since the analysis of several such loci can be used to determine their linear order along a chromosome.

Mapping the human genome

Human linkage maps are constructed by analysing the segregation of informative markers in established pedigrees. Where a variant is manifested as an inherited disease phenotype, linkage studies can be instrumental to identifying the mutant gene responsible, by using the reverse genetics, or positional cloning, approach. This involves the identification of particular alleles of a polymorphic marker that co-segregate with the disease trait in each affected pedigree and assesment of the degree of linkage between the marker and disease loci. Sequential chromosome walking through progressively more closely linked informative loci followed by physical mapping can then allow identification of the gene itself. This approach has had notable successes in identifying the genes responsible for some of the more common human genetic diseases, for example, cystic fibrosis (Rommens *et al.*, 1989). The human genome project has taken on the considerable undertaking of locating and identifying the many other genes in the human genome; an endeavour that will require initial mapping and eventual sequencing of the entire human genome. Since the power of a genetic map increases with the quality and density of its markers, the crucial first stage of this task is the generation of a detailed human linkage map, with a high density of sufficiently informative polymorphic markers dispersed evenly throughout the genome.

Protein polymorphism. The first human biochemical polymorphism to be studied was the ABO group system, which was identified serologically (Landsteiner, 1900). Several other variants derived from expressed sequences were subsequently detected, usually by protein electrophoresis, protein sequencing, or immunological assays. Many of these were of considerable intrinsic interest because of their association with human disease or histocompatibility. For example, sickle cell anaemia and many other haemoglobinopathies, as well as several commonly polymorphic blood group antigens, were identified and characterised by analysis of protein and amino acid variation (Pauling *et al.*, 1949; Ingram, 1957; Weatherall & Clegg, 1976; Race & Sanger, 1975). Such polymorphisms provided the first generation of markers for human genetic analysis, but unfortunately the majority only have a limited number of alleles and are therefore relatively uninformative. A notable exception are the human leucocyte antigens (HLA) which are encoded by a cluster of loci on chromosome 6. Each of these has a large number of alleles, many of which have low population frequency, resulting in extreme variability of HLA haplotypes and providing a highly informative genetic system (reviewed by Bodmer, 1981).

DNA sequence polymorphism. The advent of recombinant DNA technology allowed the investigation of variation in the human genome at the DNA sequence level. DNA sequence variants as small as a single base change can be detected if they create or destroy sites for restriction endonucleases, since this alters the migration patterns of

restriction fragments seen on Southern blot hybridisation. This technique was first used to indirectly diagnose sickle cell anaemia, using an *Hpa*I site linked to the β-globin locus (Kan & Dozy, 1978). Initial studies at the β-globin cluster suggested that such restriction fragment length polymorphisms (RFLPs) could be common and widely dispersed throughout the genome (Jeffreys, 1979), an observation which is supported by more recent sequence analyses (Nickerson *et al.*, 1992). RFLPs have since been used to generate reasonably detailed maps of human chromosomes (Donis-Keller *et al.*, 1987). New PCR-based techniques have broadened the range of single base polymorphic markers by enabling the analysis of single base substitutions that do not result in RFLPs. A single base mismatch corresponding to a polymorphic base substitution at the extreme 3′ end of a PCR primer can prevent priming from one allele, allowing allele-specific amplification from the other allele in heterozygotes (Newton *et al.*, 1989). PCR amplified DNA fragments that differ at a single base position can also be distinguished, since the two strands may adopt different conformations and therefore migrate differently following denaturation and non-denaturing gel electrophoresis (Orita *et al.*, 1989). However, despite these advances, single base polymorphisms do not make ideal markers for linkage analysis. Their utility is severely limited because most are relatively uninformative being diallelic and therefore having only three possible genotypes. The maximum heterozygosity of such a system in a population at Hardy-Weinberg equilibrium can never exceed 50% and is usually found to be considerably lower.

Tandem repeat polymorphism. Many of the VNTR loci are highly polymorphic with large numbers of alleles and consequently can have heterozygosities considerably higher than 50%, making them highly suitable for linkage analysis. The discovery and isolation of minisatellite loci with heterozygosities in excess of 90% (Wong *et al.*, 1986, 1987; Armour *et al.*, 1990; Vergnaud *et al.*, 1991) raised considerable expectations, since it seemed that these would offer extremely informative marker loci. While this initial promise has, to a large extent, been realised, particularly in the linkage mapping of chromosome ends (eg. Nakamura *et al.*, 1987a) and disease loci (eg. Reeders *et al.*, 1985; Malcolm *et al.*, 1991), the non-random distribution of minisatellite loci (Royle *et al.*, 1988; Nakamura *et al.*, 1988a; Armour *et al.*, 1990; Vergnaud *et al.*, 1991) has precluded their use in global linkage analysis. Similarly, the predominantly centromeric location of major satellite DNAs means that polymorphisms here are also of limited use as linkage markers. Furthermore, it is considerably more difficult to analyse variation at these loci than at the minisatellites. Conventional α-satellite assays require the analysis of high molecular weight DNA by pulsed field gel electrophoresis and the resulting banding patterns are complex and often require the interpretation of intensity differences (Ge *et al.*, 1992; Haaf & Willard, 1992). A few PCR assays that can be used to haplotype variant repeats within centromeric α-satellite have been developed (Warburton *et al.*, 1991) and although these are more easily reproducible, they are still technically involved. Despite these difficulties, analysis of satellite markers has a specific and useful application in "anchoring" linkage maps at the centromeres of human chromosomes (Willard *et al.*, 1986). Work is currently underway in this laboratory to find telomeric markers which can be used to define the ends of chromosome linkage maps (D. Baird, personal communication).

By contrast the STR loci, and in particular CA repeats, have proved to be perfect for genome-wide linkage mapping (Wiessenbach *et al.*, 1992; Dietrich *et al.*, 1992). Although they are generally considerably less informative than the most variable minisatellite loci, this deficiency is compensated by their abundance and apparently even dispersion throughout the genome (Luty *et al.*, 1990; Stallings *et al.*, 1991). Furthermore analysis of these loci can be rapidly performed by PCR and is amenable to a large degree of automation, in a factory-like approach (see

Coghlan, 1994), it is for these reasons that this system has been adopted as one of the universal methods of choice for the construction of detailed eukaryotic linkage maps (Wiessenbach *et al.*, 1992; Dietrich *et al.*, 1992). In humans this work is well underway, with detailed microsatellite linkage maps of several chromosomes already available (eg. Wilkie *et al.*, 1992; Bowcock *et al.*, 1993; Wang *et al.*, 1993; Weber *et al.*, 1993) and the recent publication of a human genomic linkage map with an average marker spacing of 2.9cM (Gyapay *et al.*, 1994). These maps of anonymous loci are gradually being integrated with maps of known gene loci, for example that of the NIH/CEPH Collaborative Mapping group (1992), to build a fully integrated human genetic linkage map (Matise *et al.*, 1994; Buetow *et al.*, 1994).

The applications of polymorphic loci in forensic analysis and individual identification

All individuals, with the exception of identical twins, are genetically unique; therefore it is theoretically possible to unambigously identify any individual by DNA analysis. Interestingly one of the only ways of distinguishing identical twins is by traditional fingerprinting, emphasising the important role non-genetic influences have in determining phenotypic characteristics. It is obviously impossible to completely define all the distinctive characteristics of an individual genome, but the next best approach, analysis of the most polymorphic loci, provides a good starting point. The rationale for the development of such typing systems is obvious, the many potential applications include the matching of samples in forensic science and the establishment of relatedness in pedigree testing.

Analysis of protein variants. Classically such analyses have used enzyme polymorphisms and cell surface antigen markers, themselves an expression of underlying genetic variation, such as the blood group antigens and the HLA system. These are still widely used in much forensic work, sometimes being favoured over the more recently developed and hence less well established DNA typing systems, particularly in the United States, where the powerful American Association of Blood Banks has a virtual monopoly on paternity testing and the utility of DNA typing is viewed with scepticism in some quarters. However, these systems suffer from several serious drawbacks. Firstly they show only modest levels of variation and are therefore incapable of providing a unique biological identifier. Secondly most of these markers are based on blood group substances which are not present in other body tissues and can therefore only be used to type blood. Thirdly proteinaceous markers are frequently unstable in forensic specimens. Fourthly, in paternity testing sequential analysis of many markers can exclude the majority of falsely-accused non-fathers, but cannot provide proof positive of paternity (Jeffreys, 1991).

DNA typing systems

Multilocus DNA fingerprints. It is now 10 years since the generation of the first DNA fingerprint and the realisation of the considerable potential of this technique to address some of the difficulties outlined above, by typing DNA variation directly (Jeffreys *et al.*, 1985b). The implementation of this technology has been extremely rapid, and to a large extent sucessful, with DNA typing systems in place in public and commercial forensic laboratories in at least 25 different countries and many others actively considering DNA analysis in forensic and legal medicine (reviewed by Jeffreys & Pena, 1993). There is now a wealth of genetic and population data

confirming the accuracy of multilocus human DNA fingerprint evidence, which has been accepted in civil and criminal court cases in both the U.K. and the U.S.A. (Jeffreys & Pena, 1993). By far the most commonly used and universally established of the many potential applications of multilocus DNA fingerprinting has been the determination of disputed family relationships in either paternity (Jeffreys *et al.*, 1991b) or immigration cases (Jeffreys *et al.*, 1985c). Initial experiments indicated that multilocus DNA fingerprints were highly individual specific (Jeffreys *et al.*, 1985b) and could be applied to forensic casework, for example the DNA typing of, even relatively old, semen and blood stains and the comparison of fingerprints from semen DNA with those of suspected rapists (Gill *et al.*, 1985). A notable early success was the first acceptance of DNA typing evidence in a U.K. criminal court which led to the elimination of a false suspect from a murder enquiry (Gill & Werrett, 1987). Since then it has been shown that under ideal conditions multilocus DNA fingerprints can be used to match forensic samples from a single test with near certainty (Jeffreys *et al.*, 1991b). However, there are several limitations associated with this technique that have prevented its widespread application in forensic science. Multilocus probes need reasonably large quantities (≥500ng) of good quality DNA per test. Although such a quantity is readily available from fresh blood in relationship testing, this requirement renders MLPs too insensitive for fingerprinting of many forensic samples. Furthermore, forensic samples frequently contain partially degraded DNA, or DNA from more than one individual, resulting in complex fingerprints that are difficult to interpret (Jeffreys & Pena, 1993). Additional drawbacks are caused by the overall limitations of agarose gel electrophoresis. Poor resolving power compared to fingerprint detail, gel distortions and the requirement of DNA size markers all result in an inability to precisely determine band size. This difficulty is compounded by the need for inter-blot comparisons in some forensic analyses, for example where DNA fingerprints are run on different gels, or in different laboratories. Again this is in direct contrast to relationship testing where all samples for comparison are run on the same gel. Finally the information contained in the complex patterns generated by multilocus probes is difficult, although not impossible, to store in the computer databases of the kind that are becoming increasingly used in forensic science.

Single locus DNA profiles. One way of overcoming the lack of sensitivity of multilocus DNA fingerprinting is to use a panel of single locus probes (SLPs) to genotype some of the individual highly variable loci that contribute to the phenotype of a DNA fingerprint. To be appropriate for forensic analysis the loci detected should ideally have high (>90%) heterozygosity and be unlinked and anonymous. Each SLP should be locus-specific and detect one band per allele, to give a two band pattern in heterozygotes on Southern blot analysis of restriction digested DNA. The original panel of minisatellites selected for single locus analysis was made up of five highly variable minisatellites conforming to these criteria. Since none of them contained a site for *Hinf*I in their repeat unit they could all be typed simultaneously or sequentially on the same Southern blot of *Hinf*I digested DNA (Wong *et al.*, 1987). The multiple single locus profiles generated by pooling probes can be obtained from as little as 50ng DNA (Wong *et al.*, 1987) , corresponding to 2µl blood, <1µl semen or 20µl saliva. Such profiles are technically easier to produce and easier to interpret, particularly in the case of mixed or degraded DNA samples, and are therefore well suited to forensic analysis (Wong *et al.*, 1987). Improvements in the technique have reduced the lower limit of detection to ~10ng DNA (Jeffreys & Pena, 1993), a level usually sufficient to obtain a profile from a single hair root.

Although these profiles are are not as discriminatory as a DNA fingerprint, sequential application of single locus probes gives a level of individual specificity between unrelated people comparable to that achievable using one multilocus probe, because all the alleles at hypervariable loci have low population frequency (Wong *et al.*, 1987). However, SLPs are relatively poor at discriminating between close relatives. For example, a single probe has, at best, only a 75% chance of distinguishing two siblings (<99.8% for 5 probes). It is for this reason that SLP analysis is referred to as DNA profiling rather than DNA fingerprinting. It is easier to database this information and compare profiles from different tests than with DNA fingerprints, even if these are from different gels/laboratories. For each locus the expected match frequency is calculated from an allele frequency database under the assumption that the population is at Hardy-Wienberg equilibrium. The match frequencies are then multiplied together, assuming linkage equilibrium, to obtain the overall match probability.

The limitations of MLP and SLP analysis

Minisatellite mutation. The power of MLP and SLP typing derives from the very high levels of allelic variability of the loci used in these analyses. However, although this makes minisatellites well suited to the generation of individual specific information, the associated high mutation rate to new length alleles potentially compromises their application to relationship testing. Mutations at a locus detected in MLP or SLP analysis will generate one or more bands in the offspring that are not attributable to either genuine parent, providing evidence which could be interpreted as an exclusion. For example, 27% of offspring show one mutant band in 33.6 and 33.15 MLP fingerprints (Jeffreys & Pena, 1993). However, analysis of paternity cases showed that this is not a significant problem, since the proportion of non-maternal bands in a child which cannot be attributed to the alleged father is consistently higher in falsely accused non-fathers than in true fathers (Jeffreys *et al.*, 1991c). In practice pedigree analyses showing mutant bands are further tested using SLPs. Although the mutation rates at individual minisatellite loci can be high, mutation rates of upto ~10^{-2} per gamete do not significantly interfere with the use of SLPs in paternity analysis, provided that they can be quantified and incorporated into stastistical likelihood ratio analyses of paternity against non-paternity (Jeffreys & Pena, 1993). However, less variable loci with unknown mutation rates where this likelihood cannot be determined, except by estimation of mutation rates from locus heterozygosity (Jeffreys *et al.*, 1988a), will occasionally generate inconclusive results. For this reason only the most variable SLP probes should be used in paternity testing.

Minisatellite mutation is not restricted to the germline, but also occurs somatically as shown by the analysis of clonal tumor cell populations, lymphoblastoid cell lines (Armour *et al.*, 1989b) and single molecule PCR analysis of normal somatic DNA (Jeffreys *et al.*, 1990). Although the proportion of cells harbouring new mutant alleles can be significant, the heterogeneity in *de novo* mutant allele length will prevent their detection by Southern blot hybridisation. This will only pose a potential problem if a mutation occurs in a very early stem cell lineage, which will create a tissue, either mosaic for original non-mutant alleles plus cells descended from the same mutant progenitor cell (creating a tissue with three alleles), or a tissue homogenously composed of mutant cells. Such a process could result in the divergence of single locus profiles between different tissues of the same individual, for example blood and sperm, of obvious concern in forensic analysis. Fortunately this appears to be an extremely rare occurence at human minisatellite loci; the only known example discovered to date is described in Chapter 6.

Quantifying allelic diversity. The stastistical evaluation of DNA profile evidence requires knowledge of allele frequencies and assessement of the validity of the assumption of Hardy-Wienberg equilibrium in human populations. The single locus probes chosen for forensic casework all show extraordinary levels of allelic variability. The most informative locus is MS1 (D1S7); Southern blot length analysis indicates that it has alleles ranging from 1-23kb long, heterozygosity >99% (Wong *et al.*, 1987) and a mutation rate of 5% per gamete (Jeffreys *et al.*, 1988a). The MS1 repeat unit is 9bp long yielding in principle 2400 different length alleles. Determination of human MS1 allele length frequency distributions show that there are no common alleles at this locus in Caucasians (Smith *et al.*, 1990) and theoretical considerations suggest that most, if not all, of the possible allelic states exist in human populations (Jeffreys & Pena, 1993). Estimation of allele frequencies at such hypervariable minisatellite loci is limited by the resolving power of agarose gel electrophoresis. Below a certain threshold, allele length differences cannot be resolved by Southern blot hybridisation, making allele length distributions at these loci quasicontinuous. It has been estimated that a maximum of at most ~50 alleles of a hypervariable minisatellite may be distinguished using this technique (Wong *et al.*, 1987). In contrast, loci with lower variability (<96% heterozygosity) tend to show a more limited number of distinct alleles, with real and measurable population frequencies (Wong *et al.*, 1987; Smith *et al.*, 1990). Allele length distributions at such loci tend to be discontinouous or "spiky" with the result that small errors in allele sizing can result in large errors in allele frequency estimates. These less variable loci will also be more vulnerable to genetic drift and inbreeding effects, again indicating that only the most hypervariable loci should be used in forensic analysis.

The inability to resolve closely spaced alleles at the most variable loci will result in an apparent excess of homozygotes compared to Hardy-Weinberg predictions. It is also possible that the aberant migration, or "bandshift", of a DNA fragment, due to contamination with co-purifying impurities (see Thompson & Ford, 1991) may result in the false declaration of an exclusion between two profiles that actually match. In order to overcome these limitations it is necessary either to pool, or "bin", alleles into groups covering a small size range, or to define an allele with qualifying error margins (Budowle *et al.*, 1991a). It is important that bins or error margins are conservative relative to the criteria used to declare a forensic match, so that either of these approaches will result in the decreased statistical significance of a match, and therefore bias the weight of evidence in favour of the defendant. This approach has come to be widely accepted among the forensic science community, but not without considerable scientific debate. Much of the initial hostility highlighted the limitations of defining match criteria and the problems of bandshifts discussed above (eg. Lander, 1989), although to some extent this was based on examples of bad laboratory practice and the over-enthusiastic interpretation of data. These doubts over aspects of procedure and profficiency have now been largely alleviated by the instigation of rigorous quality controls and interlab standardisation (eg. Gill *et al.*, 1992).

The DNA profiling controversy. Scientific arguments about the population assumptions used in the calculation of expected match frequencies and statistical assessment of profile matches have been much more bitter and protracted (eg. Cohen, 1990; Lewontin & Hartl, 1991; Chakraborty & Kidd, 1991; Brookfield, 1992). This has resulted in considerable confusion over the interpretation of DNA evidence in both the U.S.A. (eg. Lander, 1991) and the U.K. (eg. MacIlwain & Dickson, 1994). The major concern of the antagonists is that significant population substructuring may exist. They argue that differences in allele length frequency distributions between ethnic groups will alter expected match probabilities and that localised inbred sub-groups within a population can lead to

homozygote excess, and hence deviation from Hardy-Weinberg equilibrium. Differences in ethnic allele frequency distributions are increasingly seen as probe variability is reduced, but for the most variable loci similar allele length distributions are found even amongst radically different ethnic groups (reviewed by Jeffreys *et al.*, 1991c). The first objection can therefore be overcome by establishing databases from different ethnic groups and by using only the most variable loci, such as MS1, where a high mutation rate counteracts the effects of genetic drift.

The question of departures from Hardy-Weinberg equilibrium is more fraught (reviewed by Monckton & Jeffreys, 1993). Apparent homozygote excess at some loci has been noted in some populations (Lander, 1989; Budowle, 1991a) and proposed to be the result of population substructuring (Lander, 1989; Cohen, 1990). Inbreeding studies have shown that although there was some evidence for enhanced allele-sharing in some ethnic groups (Bellamy *et al.*, 1991), these effects are modest and unlikely to significantly affect the statistical evaluation of single-locus profiles (Jeffreys *et al.*, 1991b, 1991c). Statistical reappraisal of the data, using only information from heterozygous individuals, has shown that there is no departure from Hardy-Weinberg equilibrium, making it seem likely that this phenomenon results from the inability to resolve closely spaced bands and is largely caused by differences in the criteria used to bin alleles of similar size and define homozygotes (Devlin *et al.*, 1990, Devlin & Risch, 1992). Furthermore, theoretical calculations have shown that the degree of substructuring required to account for the observed departures from Hardy-Weinberg equibrium is too large to be consistent with known demographic data (Chakraborty & Jin, 1992). However, these questions are not simply answered and deliberation on this point continues (eg. Balazs, 1993; Budowle, 1993; Chakraborty & Jin, 1993). An additional source of apparent homozygotes is the presence of genuine "null" alleles, which do not exist, and of apparently "null" alleles which are so small that they migrate off the bottom of the gel, or contain too few repeats to be efficiently detected by Southern blot hybridisation. Examples of all these types of "null" alleles have been reported at different variable loci (Wong *et al.*, 1990; Armour *et al.*, 1990, 1992b).

In an attempt to reduce some of the scepticism surrounding the forensic applications of DNA profiling, and to find a compromise approach that could reconcile the opposing factions, the National Research Council of the US National Academy of Sciences commissioned an investigating committee to evaluate the technology and propose guidelines for its use. Their report concluded that DNA typing technology was scientifically sound and endorsed the use of DNA profiling in forensic casework. To allay the arguments about population substructure, in the absence of further investigation, the use of a "ceiling" approach to estimating match significance was proposed. This uses the maximum allele frequency (ceiling frequency) identified in any subpopulation, or 5%, whichever is the greater, as the basis for estimating genotype frequencies and hence match probabilities. Although this conservative estimate results in a loss of informativeness, this can be compensated by using an increased number of tests. This report provoked immediate and sometimes hostile reaction (see Monckton & Jeffreys, 1993), mainly from the population genetic community who criticised the ceiling principle, and had little effect in stemming the acrimonious scientific debate (eg. Balding & Donnelly, 1994). In light of data published since the original report a new study has been proposed, largely at the instigation of the Federal Bureau of Investigations. Again this has raised a storm of controversy among lawyers and scientists, many of whom fear that it could undermine the recommendations of the first report and tilt the balance of interpretation of DNA evidence in favour of prosecutors and the FBI (see MacIlwain & Dickson, 1994).

PCR based DNA typing systems.

Many forensic specimens do not yield DNA of sufficient quantity and/or quality for standard SLP analysis, notwithstanding the improvements in sensitivity offered by this technique. For this reason there has been considerable interest in applying PCR, which can in principle genotype DNA at the single molecule level, to the problems of forensic biology. At this stage the most enthusiastic proponents of a variety of PCR based approaches are the research laboratories where they have been developed and are being continually improved, rather than the forensic community, which is evaluating the technology and remains an interested, but cautious, potential end user.

Minisatellites. Soon after the isolation of the first minisatellite loci it was demonstrated that it was possible to amplify their alleles by PCR, using primer sites in the DNA flanking these loci (Jeffreys *et al.*, 1988b, 1990; Boerwinkle *et al.*, 1989; Horn *et al.*, 1989). Initial experiments indicated that it was possible to type subnanogram quantities of DNA, and in some cases single molecules of DNA (Jeffreys *et al.*, 1988b, 1990). By simultaneously amplifying alleles from several different loci, relatively informative PCR fingerprints were generated from as little as 1ng DNA (Jeffreys *et al.*, 1988b). However, there were difficulties associated with this approach to forensic typing that prevented it from living up to its original promise. The major problem was the inability to amplify alleles larger than ~10kb using PCR, and the fact that the smaller of a pair of alleles amplified much more efficiently than the larger, making large alleles harder to detect. Furthermore, at the high cycle numbers required to amplify some of the larger alleles to detectable levels, PCR products "collapse" into a heterodisperse smear of spurious minisatellite products (Jeffreys *et al.*, 1988b). It is presumed that this occurs due to out-of-register annealing of single-stranded repeats during the extension phase, although spurious products could also arise from mispriming elsewhere in the genome (Jeffreys *et al.*, 1988b). A solution to these problems is to use minisatellites with small, efficiently amplified, alleles. Three minisatellites with these characteristics have been identified and successfully used in forensic casework. These so called amplified fragment length polymorphisms, or "Amp-FLPs", are ApoB (Boerwinkle *et al.*, 1989), D17S5 (also known as D17S30, Horn *et al.*, 1989) and D1S80 (Budowle *et al.*, 1991b). Because most alleles of these loci are <1kb these minisatellites are very easy to type by PCR and can be detected following polyacrylamide gel electrophoresis by ethidium bromide or silver staining. An added advantage is that mutations at these loci are very rare, although as a consequence they are considerably less variable, and therefore less informative, than the minisatellites used in SLP analysis.

Microsatellites. The short alleles of polymorphic microsatellite loci are also well suited to analysis by PCR, but like the Amp-FLPs STRs are relatively uninformative, with the exception of some higher order STR repeats (eg. Polymeropoulos *et al.*, 1992; Mahtani & Willard, 1993). However, these loci, in particular CA repeats, are also highly abundant, making it possible to obtain highly discriminatory information by using a large battery of moderately informative STRs, or alternatively typing two closely linked loci, thus obtaining a large number of haplotypes (Pena *et al.*, 1994).

The analysis of multiple, polymorphic STR loci has a particularly useful application in the typing of samples from which only small DNA fragments can be extracted. There has therefore been considerable forensic and anthropological interest in the possibility of typing DNA recovered from skeletal remains or other forensic specimens, since this is frequently too degraded to allow analysis of loci any larger than a few hundred base pairs.

The ability to extract small fragments of human DNA from increasingly ancient sources and PCR amplify regions of the mitochondrial genome has been widely reported (eg. Paabo *et al.*, 1988; Hagelberg *et al.*, 1989). Several groups have recently extended this analysis to STR markers, albeit in younger skeletal remains exhumed several years after burial. Although the human DNA component usually comprised <1% of the DNA recovered, was severely degraded and contained PCR inhibitors, STR typing still proved possible. This technique has not yet enjoyed widespread forensic application, but it has been successfully used in a few well publicised cases. These include: the identification of the skeletal remains of a murder victim (Hagelberg *et al.*, 1991), the identification of the of the skeletal remains of the Nazi war criminal, Josef Mengele (Jeffreys *et al.*, 1992) and the identification of the remains of the Romanov family (Gill *et al.*, 1994).

HLA-DQα. The HLA system contains a large number of single base substitutions that together comprise a set of informative genetic markers which can be assayed at both the protein and DNA levels. A PCR based DNA typing system has been developed for HLA-DQα, which is one of the most informative of these loci (Saiki *et al.*, 1986; Higuchi & Blake, 1989). This involves PCR amplification of a short (<250bp) genomic fragment, followed by allele classification by dot-blot hybridisation with a range of allele-specific oligonucleotide probes. At present this system can distinguish 6 alleles and thus 21 different genotypes. However, the locus has an average hererozygosity of only ~80% with several common alleles (Helmuth *et al.*, 1990; Tamaki *et al.*, 1991) making it insufficiently informative to be used for individual identification, and complicating interpretation of data from of mixed samples. Furthermore, analysis is dependent on signal intensity differences which are prone to fluctuations resulting from rare variants. For these reasons this system is mainly used for exclusion in forensic casework (Blake *et al.*, 1992; Comey *et al.*, 1993). Similar assays have been developed at other HLA loci, and this approach has also been applied to haplotype other small clusters of base substitutional polymorphisms (eg. Nickerson *et al.*, 1992). The use of DNA typing systems based on coding loci have raised some ethical concerns, because of the potential for deriving phenotypic information from these genotypes. For example HLA haplotypes have been linked with a number of genetic diseases (eg. Yanagawa *et al.*, 1993). For this reason DNA typing has been restricted to non-coding loci in German courts, although paradoxically, serological HLA typing is still admissable as evidence.

Mitochondrial D-loop sequencing. There are ~1000 mitochondria in most human cells, each with their own small genome. This vastly improves the chances of obtaining fragments of mitochondrial DNA, rather than nuclear DNA, from specimens in which the DNA is badly degraded. This is reflected in the ability to PCR amplify mitochondrial, but not nuclear, DNA sequences from very ancient samples. Although mitochondrial DNA is strictly maternally inherited and can give no information in paternity analyses, it is extremely useful for the identification of human remains by comparative studies with matrilineal relatives (eg. Gill *et al.*, 1994). The highly variable control region of mitochondrial DNA, otherwise known as the D-loop, contains many positions of base substitutional polymorphism (Greenberg *et al.*, 1983). PCR amplification and sequencing of fragments from this region can therefore provide useful forensic information (Sullivan *et al.*, 1992; Ginther *et al.*, 1992). DNA sequencing is labour intensive compared to allele length analyses; however, recent advances in automating this process may increase the efficiency of this forensic application (Hopgood *et al.*, 1992).

The advantages and limitations of PCR based typing systems. The plethora of PCR based typing systems continues to grow and the approaches outlined above are continually being improved. As the research laboratories set out their stalls it becomes increasingly difficult for the window-shopping forensic community to pick out a method of choice, which can be quality controlled, standardised and universally implemented. Each of these techniques has particular advantages and disadvantages, meaning that their application will probably be piecemeal, with combinations of techniques appropriate to the circumstances of each case being selected.

The biggest advantage of these PCR based systems is the sensitivity they confer, increasing the potential range of forensic application to hair root, saliva, urine and skeletal remains (Jeffreys *et al.*, 1988). However, this is accompanied by ever attendant and formidable problem of sample contamination, particularly when very small quantities of DNA are being analysed. This can occur either by carry-over of extant PCR products or inadvertant contamination of evidentiary material with extraneous human cells. This is of great concern in a forensic context and necessitates the adoption of particularly rigorous quality control procedures in the analysis of forensic samples. Another possible concern is that PCR analyses operating at, or close to the single molecule level, will detect somatic mutants of the kind identified at low levels in blood and sperm DNA (Jeffreys *et al.*, 1990). However, there is evidence for a correspondance between variability and somatic mutation rate (Armour *et al.*, 1989b), and fortunately most of the PCR typed loci are of much lower variability than minisatellites, making it likely that they also have lower somatic mutation rates.

Many aspects of PCR based typing systems, which are already quick and easy to use, are amenable to automation, for example by using automated in-gel, or microcapillary, electrophoretic analysis of fluorescence-tagged PCR products (Sullivan *et al.*, 1991, 1993). This will be important in reducing the cost and improving the efficiency of any system adopted for forensic use. At first it seemed that the PCR amplification of STR or small minisatellite loci may be able to overcome the problems of allele length measurement. HLA-DQα assays the presence or absence of a particular allele, and microsatellite alleles and mitochondrial sequences are resolved on polyacrylamide gels, which should enable absolute length determination. However, a PCR "stuttering" effect is frequently observed during polyacrylamide electrophoresis of dinucleotide repeats, resulting in the appearance of spurious shadow bands in the microsatellite profile (Weber & May, 1989) which can introduce uncertainty in evidentiary use. This difficulty may be overcome by analysing highly polymorphic tetranucleotide repeats on agarose or polyacrylamide gels (Gray, 1991a). Amplifications of Amp-FLPs are cleaner than microsatellites, and it was originally hoped that alleles at these loci would fall neatly onto the steps of an integral allele length ladder. It now appears that this was somewhat premature, as repeat unit length variants have been detected at these loci (eg. Berg *et al.*, 1993). This means that match criteria are not absolute, once again requiring the adoption of a binning procedure, which results both in a loss of statistical power, and in court arguments about what constitutes a match .

Although some of these methods address the problem of accurate allele length analysis, the population genetic concerns raised over single locus analysis of hypervariable minisatellite loci apply even more to the less variable PCR based systems. The numbers of alleles at these loci are much smaller, because they have lower mutation rates, and are therefore more vulnerable to the effects of population bottlenecks and genetic drift. A further potential problem with PCR based analysis is that the presence of sequence variants in primer annealing sites will prevent amplification from some alleles. Such variants may be rare in the population from which the primer sequence was

derived but common in other populations, in either case the result may be allele loss and apparent homozygosity. Depending on the position of such variants in the primer sequence allele-dropout may be susceptible to minor fluctuations in PCR conditions (D. Monckton, personal communication) and therefore non-reproducible. The only solution in such cases would be to retype alleles with alternative primers, considerably increasing the workload.

MVR-PCR: an ideal DNA typing system? Consideration of the merits of the various DNA typing systems available can be used to compose a wish-list of characteristics for the ideal DNA typing locus. Such a marker should have limited allele size, perhaps 100-500bp, so that all alleles can be amplified by PCR, even in degraded DNA. It should also be possible to resolve all allelic states, enabling precise allelic classification. There should also be a high level of allelic variation with a large number of alleles, generated by a quantifiable and sufficiently high mutation rate (perhaps 10^{-2} to 10^{-3} per gamete) to counteract genetic drift effects. Needless to say, no such marker has yet been identified, nor may one exist in the human genome.

Most DNA typing systems used in forensic and legal medicine assay allelic length variation at tandem repetitive DNA regions such as minisatellites. A novel alternative approach to DNA typing that fits many of the criteria listed above has recently been developed in this laboratory. This uses PCR to assay the interspersion of repeat unit sequence variants within minisatellite alleles (MVR-PCR). The development and application of this method was the basis of the work described in this thesis and it is therefore introduced in detail in Chapter 3. Briefly, MVR-PCR produces an unambigous, and highly individual-specific digital code from genomic DNA and can distinguish between most, if not all allelic states (Jeffreys *et al.*, 1991a; Chapter 3). The codes are resolved on agarose gels but do not require band size measurements. Furthermore they are simple to score and highly amenable to computer databasing and analysis, allowing them to be transferred between laboratories. Match criteria are unambigouous and the use of diploid code phenotypes avoids the use of Hardy-Weinberg assumptions in calculating match frequencies. MVR-PCR is also applicable to mixed DNA samples, degraded DNA and kinship testing. This technology has been patented and is being commercially developed by Cellmark Diagnostics and evaluated by a number of forensic laboratories worldwide. However, it remains to be seen whether MVR-PCR will become the standard and universal DNA typing tool long sought by the forensic community.

Chapter 2

MATERIALS AND METHODS

The data presented in this thesis was collected using a few very well established molecular biology techniques which are adequately described elsewhere in standard laboratory manuals (eg. Sambrook *et al.*, 1989; Ausubel *et al.*, 1994). This chapter provides an overview of the general methods used, gives references for more detailed descriptions and also describes protocols assembled from existing techniques to develop some of the procedures used in this laboratory. The exact experimental conditions and techniques used are given in the relevant results chapters.

Materials

Hardware and consumables. All chemicals, reagents and plasticware used were standard and purchased from recognised suppliers of molecular biology reagents (Applied Biotechnologies Limited, Boehringer Mannheim, Fisons, FMC bioproducts, Gibco-BRL, New England Biolabs, Serva, Sigma, Pharmacia and Perkin Elmer Cetus) according to cost availability and applicability. Custom built gel tanks were constructed in house.

Oligonucleotides. Hexadeoxyribonucleotides for random oligonucleotide priming were supplied by Pharmacia. Oligonucleotides for polymerase chain reaction amplification of DNA were synthesised by J. Keyte (Department of Biochemistry, University of Nottingham) and by D. Langton (Department of Biochemistry, University of Leicester). They were ethanol precipitated and dissolved in PCR clean water prior to use.

DNA and blood samples. DNAs from lymphoblastoid cell lines derived from 40 large Caucasian families were supplied by Professors H. Cann and J. Dausset of the Centre d' Etude du Polymorphisme Humain (CEPH, Paris, France). Blood and DNAs from other families were provided by Dr. M. Webb (Cellmark Diagnostics, Abingdon, UK.) Dr. C. Mathew (Guy's Hospital, London, UK.) and Dr. L Henke (Institut für Blutgrupenforschung, Dusseldorf, Germany). Blood and DNA samples from unrelated Japanese individuals were kindly donated by Professor Y. Katsumata, (Nagoya University, Japan.), Malaysian DNAs were from Dr. C.L. Koh (Department of Genetics and Cellular Biology, University of Malaya, Malaysia) and Caucasian DNA was obtained from members of the Department of Genetics, (Leicester University) who donated venous blood samples which were taken by Dr. J.A.L. Armour.

Methods

DNA preparation. DNA was extracted from venous bloods under PCR clean conditions using phenol/chloroform extraction followed by ethanol precipitation, as described previously (Jeffreys *et al.*, 1990).

DNA concentration. A 1µl aliquot of DNA was run electrophoresed on an agarose gel alongside a known amount of λ/*Hind*III DNA. DNA concentration was estimated by ethidium bromide staining of the gel followed by visual comparison of the staining intensity of the DNA sample with λ/*Hind*III bands containing known quantities of DNA. The concentrations of ethanol precipitated oligonucleotide PCR primers were estimated more precisely by measuring the OD_{260} of 3 different dilutions of the oligonucleotide (1/500, 1/200, 1/100) in a Cecil Instruments CE 202 Ultraviolet Spectrophotometer. The average of these readings was used to calculate the concentration, based on the approximation that 1 OD unit is equivalent to 33µg/ml oligonucleotide. The dilution factor needed to make 10µM primer stocks used the approximation that the mass of a single deoxyribonucleotide is 330kd.

DNA manipulation. DNA modifying enzymes were used according to manufacturer's instructions in the buffer systems provided. Gel electrophoresis, ethidium bromide staining, Southern blotting *etc.* were performed as described previously (Sambrook *et al.*, 1989).

Preparative gel electrophoresis. Size fractionated DNA was prepared by digesting total genomic DNA with an appropriate restriction enzyme, followed by electrophoresis through an agarose gel. The size-fraction required was identified by comparison with DNA molecules of known length, stained with ethidium bromide and viewed under UV illumination, and was then excised from the gel in an agarose block. DNA was recovered by electroelution onto dialysis membrane followed by ethanol precipitation. To do this a small peice of dialysis tubing ~$1cm^2$ is cut along both edges, boiled in 10mM Tris-HCl (pH 7.5), 1mM EDTA for 3 minutes and the two halves of the membrane are separated. A gel of the same agarose concentration as the agarose block, but thicker than the gel from which it came, is prepared and a slot slightly larger than the dimensions of the agarose block is cut into it, such that the width of the slot is perpendicular to the eventual flow of current. The agarose block is then inserted into this slot and a peice of dialysis membrane is then inserted in front of it, to trap DNA that migrates out of the block by electrophoresis when current is applied. DNA is electrophoresed at 10-15 volts/cm until it is loaded onto the dialysis membrane (1-10 minutes depending on the size fraction required), as judged by UV illumination (if the DNA can be visualised by ethidium bromide staining), or by comparison with the migration of a coloured marker dye of known mobility. With the voltage still applied the membrane, along with a small amount of the DNA-containing buffer, is deftly removed into a 1.5ml microfuge tube using a pair of tweezers. The corner of the membrane is trapped in the lid of this tube so that the DNA solution can be separated at the bottom of the tube by centrifugation (15000 rpm for 3min) followed by removal of the membrane. The DNA is then purified by standard ethanol precipitation (Sambrook *et al.*, 1989).

DNA amplification. Large quantities of specific regions of genomic DNA were prepared by amplification using the polymerase chain reaction (PCR) (Saiki *et al.*, 1988). Precautions were taken to ensure that all tools and reagents used for PCR were free from contaminating DNA and all reactions were performed with the appropriate zero DNA controls.

PCR amplifications used approximately 100ng input template DNA and were performed in a 7µl reaction solution, (unless otherwise stated). This contained 45mM Tris-HCl (pH8.8), 11mM $(NH_4)_2SO_4$, 4.5mM $MgCl_2$, 6.7mM β-mercaptoethanol, 4.4µM EDTA (pH 8.0), 1mMdATP, 1mMdCTP, 1mMdGTP, 1mMdTTP, 113µg/ml BSA, 1µM each primer (unless otherwise stated) and 0.025 units/µl of thermostable DNA polymerase. The reactions were thermocycled in a Geneamp 9600™ (Perkin Elmer Cetus) with denaturation at 94°C for 30 seconds, annealing at the stated temperture for 30 seconds and extension at 70°C for the stated time; this was followed by a single-cycle "chase" of 30 seconds at the annealing temperature used and 70°C for 10 minutes. Cycle numbers and primer sequences are given in the relevant chapters.

MVR-PCR. MVR-PCR uses a combination of an MVR-specific primer and a primer at a fixed site in the DNA flanking the minisatellite to generate a set of PCR products extending to each MVR along minisatellite alleles. MVR detection and amplification are uncoupled by providing a 5′ extension ("TAG") to the MVR-specific primer, which is used at low concentration, and driving amplification with a high concentration of the flanking primer and the TAG sequence itself. 50-100ng of genomic DNA, or the equivalent quantity of a single allele of known size isolated from genomic DNA by preparative gel electrophoresis of the appropriate size fraction of an *Mbo*I digest, was used as the template for MVR-PCR. Reactions were carried out in a volume of 7µl containing 0.025 units/µl thermostable DNA polymerase, the standard buffer system, (see above) and the flanking primers indicated in the figure legends. Two state MS31A mapping used a flanking primer and TAG, both at a concentration of 1µM, and either 50nM 31-TAG-A to reveal the position of a-type repeats or 25nM 31-TAG-G to map t-type repeats. MS32 mapping used the same concentrations of flanking primer and TAG together with either 10nM 32-TAG-A (a-type repeats) or 20nM 32-TAG-T (t-type repeats). Amplification from both loci was carried out in the Geneamp 9600™ (Perkin Elmer Cetus), with denaturing at 94° for 30 sec, annealing at 68° for 30 sec and extension at 70° for 2.5 min for the first 10 cycles, after which the extension time was incremented by 20 sec per cycle for a further 10 cycles. Cycling was followed by a chase of 68° for 1 min and 70° for 10 min. PCR products were resolved by electrophoresis through a 35cm 1.2% agarose (Sigma Type 1) gel in 44.5mM Tris-borate (pH 8.3), 1mM EDTA, 0.5 µg/ ml ethidium bromide (0.5 x TBE), until the lowest rung of the MVR-PCR ladder was estimated to be close to the bottom of the gel by comparison with 1µg ϕX/HaeIII size markers.

DNA detection. The DNA was depurinated, alkali-denatured and transferred from both MVR-PCR and standard agarose gels to Hybond-N FP (Amersham) hybridisation membrane by Southern blotting for 2 hours using using 20 x SSC (see Sambrook *et al.*, 1989). Membranes were dried and the DNA crosslinked to them by exposure to UV radiation from a transilluminator for 40secs. Confirmation of the presence of DNA fragments from a particular locus was achieved by hybridisation to a specific radioisotopically labelled probe sequence followed by autoradiography. MS31A was detected by the 5.7kb *Sau*3AI minisatellite insert isolated from the plasmid pMS31; MS32 was detected by the 5kb *Dra*I fragment from plasmid pMS32. 10ng probe DNA was labelled by the hexamer priming method (Feinberg and Vogelstein, 1984) incorporating α-^{32}P-dCTP. After labelling, 70µl of 'stop solution' (20mM NaCl, 20mM Tris-HCl (pH 7.5), 2mM EDTA, 0.25% SDS) was added to the reaction, followed by 100µg of high molecular weight herring sperm DNA, which acted as a carrier. The probe was recovered by ethanol precipitation, washed in 80% ethanol to remove unincorporated α-^{32}P-dCTP, and redissolved in 0.5ml distilled water. Probes were boiled for 3 minutes immediately prior to use.

Filters were pre-hybridised at 65° for at least 15 minutes in 10ml modified Church and Gilbert (1984) phosphate/SDS hybridisation solution: 0.5M sodium phosphate (pH 7.2), 7% SDS, 1mM EDTA (Wong *et al.*, 1987) contained in a bottle rotating in a hybridisation oven (Hybaid), and hybridised at 65° overnight in 10ml of the same solution containing ^{32}P-oligolabelled probe. Filters were washed at high stringency in a total of 1l 0.1 x SSC, 0.01% SDS for 1 hour at 65°C, with changes of washing solution every 10 mins. Visualisation was carried out by autoradiography. Filters were placed in autoradiographic cassettes with a sheet of Fuji RX100 X-ray film. Exposures were either at -70°C, with an intensifying screen, or at room temperature without a screen, for 1 hour to 14 days, depending on estimated signal strength and the band intensity required.

DNA sequencing. Double stranded template was prepared for direct sequencing by PCR amplification from 100ng genomic DNA, or reamplification of fragments obtained by preparative gel electrophoresis using the method described above. PCR reactions were then made up to 50µl with water and the PCR products were recovered by precipitation with 60µl isopropanol and 10µl 10M ammonium acetate, (isopropanol precipitation removes unincorporated primers more efficiently than ethanol precipitation), and resuspended in 8µl water. 1µl of this solution was used in each of four, 7.5µl, Taq cycle sequencing termination PCRs, each containing a different dideoxynucleotide. These reactions contained: 66nM PCR primer, end-labelled by T4 kinase with either, γ-^{32}P-dATP, or γ-^{33}P-dATP, (fragments were usually sequenced from both ends with the primers used in the original PCR), 45mM Tris-HCl (pH 8.8), 11mM $(NH_4)_2SO_4$, 4.5mM $MgCl_2$, 6.7mM β-mercaptoethanol, 4.4µM EDTA, 113µg/ml BSA, 8µM each of, dATP, dCTP, dGTP and dTTP and one of the ddNTPs with the following concentrations; 80mM ddGTP, 250µM ddATP, 330µM ddTTP or 160µM ddCTP. After 10 cycles of PCR using the same parameters as originally used to amplify the fragments, 4µl sequencing dye (95% deionised formamide, 20mM EDTA (pH 7.5), 0.05% w/v xylene cyanol and bromophenol blue) was added. PCR products were denatured by heating to 85°C and immediately loaded onto a 5% denaturing polyacrylamide sequencing gel for resolution and detection. Polyacrylamide gel electrophoresis was performed as described previously (Sambrook *et al.*, 1989). Visualisation was by autoradiography at room temperature.

Photography. DNA in ethidium bromide stained gels was visualized by UV fluorescence on a transilluminator (Chomato-vue C-63, UV Products Inc.) and photographed with a Polaroid MP-4 camera using Kodak negative film (T-max Professional 4052). Films were processed with Kodak LX24 developer, FX40 fixer and HX40 hardener. Prints of these negatives, autoradiographs and other photographic work was carried out by Ian Ridell (Dept of Genetics) and the Central Photographic Unit (Leicester University). Colour figures and laser photocopies were produced with the kind assistance of the Central Reprographics Unit (Leicester University).

Computing. DNA sequences were analysed using the Genetics Computer Group Sequence Analysis Software Package version 6.2 developed at the University of Wisconsin (Devereux *et al.*, 1984); this was run on both a VAX 8650 Mainframe computer operating on VMS 5.4-2 and also using a Silicon Graphics Inc. 4D/480S system running IRIX. MVR codes and allele frequency distributions were analysed with software written by A.J. Jeffreys (Jeffreys *et al.*, 1991a & unpublished). Data were either stored as ASCII files and analysed using software written in VAX BASIC V3.4 and run on a VAX 8650 computer operating on VMS 5.3-1, or stored in Microsoft word™ files and analysed using programs written in Microsoft QuickBasic™, running on an Apple Macintosh personal computer.

Chapter 3

VARIANT REPEAT UNIT MAPPING AT THE HUMAN HYPERVARIABLE MINISATELLITE MS32

Summary

Minisatellite alleles frequently vary not only in repeat copy number but also in the interspersion pattern of variant repeat units along alleles. Previous analysis of the hypervariable locus D1S8 (MS32) showed two classes of repeat unit which differ by a single base substitution that creates or destroys a *Hae*III restriction site. Interspersion patterns of *Hae*III$^+$ and *Hae*III$^-$ repeat units were assayed by PCR amplification of an entire allele, followed by partial digestion with *Hae*III. This approach, although cumbersome and limited to alleles small enough (<6kb) to amplify by PCR, provides an unambiguous binary code defining the structure of an allele (minisatellite variant repeat unit map; MVR map), and has revealed extraordinarily high levels of variation at MS32, particularly at one end of these alleles. A PCR based method (MVR-PCR) which can provide the same information from alleles of any length has vastly increased the efficiency of this analysis, enabling the structural characterisation of large numbers of single-alleles. MVR-PCR of individual MS32 alleles confirms the observations of allelic hypervariability made using the original mapping system, revealing the huge amount of information hidden in the internal structure of minisatellite alleles and showing that the maximum population frequency of even the most common allele is very low. It has also allowed the characterisation of mutation events at this locus, showing them to be biased towards the gain of small numbers of repeats at the end of the tandem array seen to exhibit the greatest allelic variability. At loci such as MS32, which show little variation in repeat unit length, the application of this modified mapping technique to genomic DNA produces a profile comprised of the superimposed maps of both alleles. The extreme allelic variability at this locus makes these MS32 diploid codes highly individual-specific, indicating that diploid MVR-PCR has important potential applications in forensic DNA typing. Besides providing enormous exclusionary power, MVR-PCR is possible on mixed and degraded DNA samples of the type often encountered in forensic work and can also be used in parentage testing. Furthermore, the results can be encoded in digital form, thereby simplifying database construction and sample comparison. As such diploid MVR-PCR provides a new DNA typing system much more widely applicable and sensitive than any seen to date.

Introduction

The MS32 locus (D1S8). The single locus human minisatellite probe MS32 was originally isolated by virtue of cross-hybridisation to the multilocus fingerprinting probe 33.15 when this was used to screen a λ library of size fractionated (5-15kb) human *Sau*3AI inserts (Wong *et al.*, 1987). MS32 detects a hypervariable locus with alleles ranging in size from 2-30kb exhibiting length heterozygosity of 97.5%, as estimated from Southern blot analysis of *Alu*I digested DNA (Wong *et al.*, 1987). The locus was localised to an interstitial position on the long arm of chromosome 1 (1q42-43) by *in situ* hybridisation and linkage mapping, and assigned the locus name D1S8 (Royle *et al.*, 1988). For the sake of clarity I will refer to this hypervariable minisatellite as MS32 throughout this thesis.

Characterising variation at MS32. Sequence analysis of fragments cloned from MS32 alleles revealed that this minisatellite comprises of a tandem array of a G/C rich, 29bp sequence that has expanded from within a region of DNA that shows homology to a retroviral long terminal repeat (LTR) (Wong *et al.*, 1987; Armour *et al.*, 1989a) (Fig. 3.1A.) The very high length heterozygosity observed at this, and other hypervariable minisatellite loci, is due to the existence of alleles with many different numbers of repeat units, generated by a high germline mutation rate to new length alleles. Allelic repeat copy number at MS32 varies from 12 to 800 or more units (mode, 180 repeats) (Wong *et al.*, 1987; Royle *et al.*, 1988; Armour *et al.*, 1989a) and a germline mutation rate to resolvable new length alleles of approximately 1% per gamete at this locus has been established by pedigree analysis (Jeffreys *et al.*, 1991a). In addition to variation in the number of tandem repeats in an allele, most minisatellite alleles characterised to date also show subtle variation in the sequence of the repeated units (Bell *et al.*, 1982; Capon *et al.*, 1983; Owerbach & Aagaard, 1984; Wong *et al.*, 1986, 1987). Alleles at such loci are composed of an interspersed mixture of the different repeat unit types (Owerbach & Aagaard, 1984; Jeffreys *et al.*, 1985a; Jarman *et al.*, 1986; Wong *et al.*, 1986, 1987; Nakamura *et al.*, 1987a; Page *et al.*, 1987; Gray & Jeffreys, 1991). The positions of the minisatellite variant repeats (MVRs) along the tandemly repeated array define a map (MVR map) which represents the internal structure of an allele with respect to these repeat unit types (Fig. 3.1A). MS32 is a typical example of this phenomenon; there are two common positions of internal sequence variation between MS32 repeat units. These base substitutional polymorphisms are an A/G transition separated by 1bp from a C/T transition (Wong *et al.*, 1987) (Fig. 3.1B).

Enzymatic mapping of allele internal structure at MS32. In order to characterise detailed allelic structures at MS32, a method of visualising the MVR map of two of the repeat unit types was developed. The technique exploits the observations that MS32 repeat units almost always contain a site for *Hin*fI, but those repeat units with the G variant at the A/G polymorphic site also contain the recognition sequence for *Hae*III (Jeffreys *et al.*, 1990). This *Hae*III RFLP provides the basis for a restriction mapping based strategy (MVR mapping) used to define G variant repeats and reveal their internal positioning along MS32 alleles.

Single MS32 alleles small enough (<6kb) to be PCR amplified in their entirety are isolated by amplification from genomic DNA to levels where they are detectable as a discrete band on an ethidium bromide stained gel. Alleles are resolved by electrophoresis and recovered by removal of the ethidium bromide stained band, electroelution of PCR products onto dialysis membrane and ethanol precipitation (see Materials and Methods). Amplification between either flanking primer C or D and a modified version of the the opposite flanking primer, D1 or C1 containing an

*Eco*RI site in a 5′ extension of primers C and D (Fig. 3.1C, 1; see Table 3.1 for primer sequences), enables PCR products to be fill-in end-labeled with γ-^{32}P-dCTP following cleavage with *Eco*RI. Aliquots containing an end-labelled MS32 allele are then divided for partial digestion, either with *Hinf*I, to cut every MS32 repeat unit, or with *Hae*III, to cut only those repeats with the G variant. *Hinf*I and *Hae*III partial digests from each allele are electrophoresed in adjacent lanes on an agarose gel, which is dried down and autoradiographed (Fig. 3.1D, i). This produces a map where the products of the *Hinf*I partial digest give a ladder of bands derived from every repeat unit along the minisatellite array, while the track from the *Hae*III digest only has bands at those positions where there is a repeat unit with the "G" sequence variant in that allele. Thus the *Hinf*I lane indicates the number of repeat units along an allele, and the *Hae*III lane whether these repeats have "A" or "G" variants. Repeat units cut by *Hae*III are called a-type and those not cut by *Hae*III, t-type, enabling the banding pattern on the autoradiograph to be encrypted as a binary code which can be analysed using standard DNA sequencing software. This code describes the interspersion pattern, or MVR map, of these variant repeats along an MS32 allele (Fig. 3.1A & D, i). PCR amplified MS32 alleles can be MVR mapped from either end, depending on which of the flanking primers incorporates the *Eco*RI site allowing alleles of up to ~5kb to be mapped completely.

This technique has enabled the first large scale survey of minisatellite allelic structures at any locus to be made, revealing extraordinary levels of allelic variation at MS32, much greater than had previously been estimated from Southern blot allele length analysis (Jeffreys *et al.*, 1990), which can at best distinguish ~50 alleles (Wong *et al.*, 1987). Even alleles with the same number of repeat units have been found to have very different internal structures, implying widely diverged ancestral origins (Monckton & Jeffreys, 1991). Systematic pairwise comparison of these MVR maps by dot matrix analysis revealed a marked gradient of variability along MS32 alleles. Several alleles were found to share blocks of identical or similar repeat unit haplotypes at one end. These alleles were assumed to be related and were grouped accordingly, however, alleles within such groups had no haplotypes in common at the other end of the tandem repeat array. This observation suggested for the first time the existence of a mutational hotspot which creates polarity in allelic variation by turning over repeat units faster at one end of MS32 alleles than along the rest of their length (Jeffreys *et al.*, 1990). The ultravariable end of MS32 alleles was originally defined as the 3′ end (Jeffreys *et al.*, 1990), however, this designation has since been changed and it is refered to as the 5′ end in all subsequent papers and in this thesis.

While very informative, this mapping approach has certian limitations. There is a size constraint (~6kb) on alleles that can be PCR amplified to levels detectable by ethidium bromide staining of agarose gels (Jeffreys *et al.*, 1990). Such alleles are in the minority at this locus and therefore comprise a biased dataset. In addition, the technical difficulty of the mapping procedure, which is extremely laborious, tedious and time consuming, calls for a more efficient approach to the aquisition of MVR data.

MS32 variant repeat mapping using PCR (MVR-PCR). To overcome the difficulties outlined above an alternative MVR mapping system was developed. This distinguishes between the two repeat unit types directly, by using two different repeat-unit-specific PCR primers to recognise and prime from different repeat unit types along a minisatellite allele. The primers have 20nt complementary to the MS32 repeat unit, a non-complementary 20nt 5′ extension ("TAG"), and are identical except for the 3′-most base, which corresponds to either the "A" or the "G" form of the polymorphic *Hae*III site, (Fig. 3.1B). Over a certain annealing temperature window these primers should

Figure 3.1. MVR mapping at the human minisatellite locus MS32.

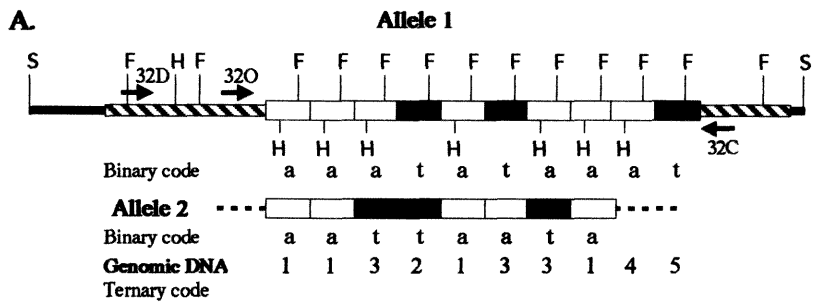
A. Schematic of the MS32 (D1S8) locus with examples of binary and ternary coding. Key: Flanking DNA (plain line), retroviral LTR (diagonal stripes); repeat units cut by *Hae*III (white boxes), repeat units not cut by *Hae*III (red boxes); restriction enzyme sites S (*Sau*3AI), F (*Hinf*I), H (*Hae*III); PCR primers (arrows, see Table 3.1 for sequences). The internal structure of minisatellite alleles with at least two variant repeat types can be described by encoding these as a binary string extending from the first repeat unit. At MS32 repeats cut by *Hae*III are called **a**-type and those not cut by *Hae*III are called **t**-type (shown below each allele). If all repeat units are the same size the repeat unit interspersion pattern of both alleles from an individual can be described simultaneously, by a ternary code, as shown. At each repeat unit position alleles may have both **a**-type repeats (code **1**), both **t**-type repeats (code **2**), or be heterozygous with one **a**-type and one **t**-type repeat (code **3**). Beyond the end the shorter allele the hemizygous code from the longer is described by codes **4**, (**aO**) and **5** (**tO**).

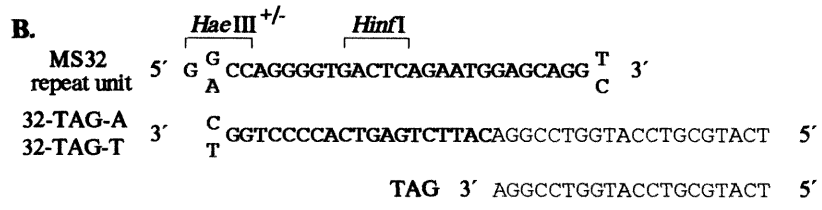
B. The consensus 29bp repeat unit of MS32, showing the two major sites of polymorphic base substitution, the constant *Hinf*I site and the variable *Hae*III site. Repeat units with the "G" variant are cut by *Hae*III (**a**-type), those with an "A" are not cut (**t**-type). 32-TAG-A and 32-TAG-T are variant repeat unit-specific oligonucleotide primers terminating at this polymorphic site, they detect **a**-type and **t**-type repeats respectively. Each primer has 20nt complementary to the MS32 repeat unit sequence (bold) preceded by a 20nt 5′ non-minisatellite extension with no known homologue in the human genome. A primer, "TAG", identical to this sequence, is also shown.

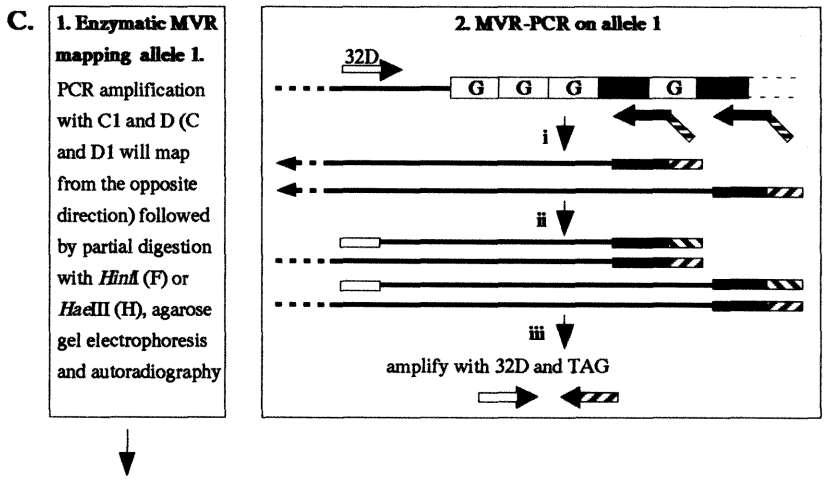
C. Enzymatic and MVR-PCR internal mapping of allele 1 (see Table 3.1 for all primer sequences). **1.** C1 and D1 are derivatives of primers C and D with a 5′ extension incorporating an *EcoR*I site. Amplification of alleles using C1 and D, or C and D1, gives PCR products with an unique *EcoR*I site, which can be cut and then end-labelled. MVR maps are generated by partial digestion of end-labelled PCR products with either *Hinf*I (F), or *Hae*III (H), followed by agarose gel electrophoresis and autoradiography. Repeats cut with *Hae*III are scored as **a**-type, while uncut repeats are scored as **t**-type (Fig 3.1D, i). **2.** The principle of MVR-PCR, illustrated for allele 1 using MVR-specific primer 32-TAG-T. **i.** During the first PCR cycle 32-TAG-T (red), at low concentration, anneals to approximately one **t**-type repeat unit per molecule and extends into the flanking DNA. **ii.** During the next cycle DNA synthesis from primer 32D creates a sequence (diagonal red stripes) complementary to TAG when the 20nt extension of the repeat unit primer is copied. **iii.** Products of the second round of PCR, which terminate with 32D and the TAG complement are now amplified efficiently by a high concentration of 32D and TAG, creating a stable set of PCR products extending from the 32D site to each **t**-type repeat unit. Occasional internal priming from PCR products by 32-TAG-T will generate authentic, but shorter PCR products. The use of primer 32-TAG-A at the first stage will create a complementary set of products terminating at each **a**-type repeat.

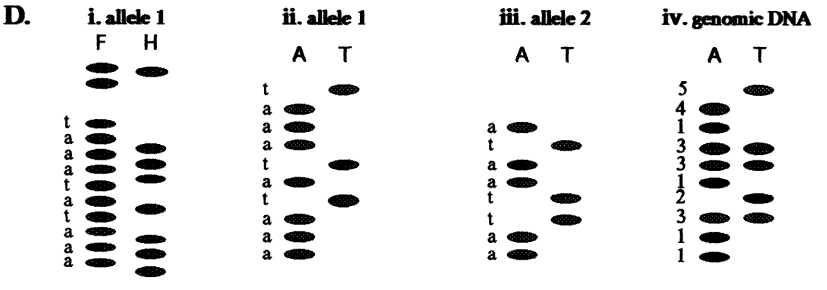
D. The profiles expected from enzymatic mapping of allele 1 (**i**), and Southern blot hybridisation of MVR-PCR products generated from MVR-PCR mapping of allele 1 (**ii**). MVR-PCR mapping of allele 2 (**iii**). MVR-PCR on genomic DNA from an individual with alleles 1 and 2 (**iv**). The binary single-allele codes and ternary code generated from genomic DNA are shown to the left of the respective maps.

This figure was adapted from Jeffreys *et al.*, (1990, 1991) and Monckton, (1993)









Chapter 3 Figure 1

Table 3.1. MS32 MVR mapping primers.

Primer uses	Primer names	Primer sequences[3] 5′-3′	Final conc. μM
Enzymatic mapping [1] primers	32C	CTTCCTCGTTCTCCTCAGCCCTAG	1.00
	32C1	TCACCGGTGAATTCCTTCCTCGTTCTCCTCAGCCCTAG	1.00
	32D	CGACTCGCAGATGGAGCAATGGCC	1.00
	32D1	TCACCGGTGAATTCCGACTCGCAGATGGAGCAATGGCC	1.00
Flanking primers	32D	CGACTCGCAGATGGAGCAATGGCC	1.00
	32O	GAGTAGTTTGGTGGGAAGGGTGGT	1.00
MVR-PCR [2] primers	TAG	TCATGCGTCCATGGTCCGGA	1.00
	32-TAG-A	tcatgcgtccatggtccggaCATTCTGAGTCACCCCTGGC	0.01
	32-TAG-C	tcatgcgtccatggtccggaCATTCTGAGTCACCCCTGGT	0.02

1. See Jeffreys *et al.*, (1990) for detailed enzymatic mapping protocols.

2. Amplifications were carried out in a Geneamp 9600 thermal cycler (Perkin Elmer Cetus), with denaturing at 94° for 30 sec, annealing at 68° for 30 sec and extension at 70° for 2.5 min for the first 10 cycles, after which the extension time was incremented by 20 sec per cycle for a further 10 cycles. Cycling was followed by a chase of 68° for 1 min and 70° for 10 min. See Materials and Methods for detailed MVR-PCR mapping protocol.

3. Uppercase denotes flanking primer sequences, TAG primer sequence and the region of MVR-specific primers complementary to the MS32 repeat unit. Lowercase denotes the TAG sequence of MVR-specific primers.

only allow extension from repeat units which they match perfectly, priming off either one or the other of the repeat units previously designated as a-type (32-TAG-A) or t-type (32-TAG-T). The TAG sequence was included to enable the prevention of progressive shortening, or "collapse" (Jeffreys *et al.*, 1988b), of amplified products down to the first few repeat units, due to MVR-specific primers priming internally within extant PCR products at each PCR cycle. This was achieved by uncoupling MVR detection and subsequent amplification by using different primers for each process. PCR amplifications are carried out with a very low concentration of either 32-TAG-A or 32-TAG-T detector primer, plus high concentrations of the two driver primers, 32D and the TAG sequence itself (see Materials and Methods for PCR conditions; Table 3.1 for primer sequences). At each PCR cycle an MVR-specific primer will prime from different cognate repeat units along the minisatellite input molecules and extend into the flanking DNA past the 32D priming site (Fig. 3.1C, 2, i). At the next cycle, 32D will prime synthesis of a complementary strand to these products back into the minisatellite, terminating at the TAG sequence and creating a sequence complementary to TAG, from which the TAG primer can now prime (Fig. 3.1C, 2, ii). The high concentration of the flanking primer and TAG now allow efficient amplification of this second PCR product during subsequent PCR cycles (Fig. 3.1C, 2, iii). Any occasional internal priming off PCR products by 32-TAG-A or 32-TAG-T will create authentic, but relatively short, PCR products in each reaction. Although the amplification of shorter molecules will be favoured in later rounds of PCR, this will be compensated by increased hybridisation of radioactive probe to longer products (Jeffreys *et al.*, 1988b). Separate amplifications under these conditions between one or other MVR-specific primer and a primer complementary to a fixed site in the minisatellite flanking DNA (32D) will generate two complementary sets of PCR products, extending from the flanking primer site into the ultravariable end of any MS32 allele, and terminating at each a-type or t-type repeat unit (Fig 3.1C, ii) (except those with additional variant positions that prevent binding by/extension from, the MVR-specific primers; these are called "null" repeats) . The two sets of PCR products can then be resolved side-by-side by electrophoresis through an agarose gel and detected by Southern blot hybridisation and autoradiography, to reveal the MVR map (Fig 3.1D, ii, iii; Jeffreys *et al.*, 1991a).

This MVR mapping system is much simpler and more efficient than its predecessor and generates MVR data equivalent to that obtained by enzymatic mapping, which can therefore be encoded in the same way. It is applicable to MS32 alleles of any length and can also be used on genomic DNA to display the superimposed MVR maps of both alleles, thereby generating a ternary, rather than binary, code (Fig. 3.1A, 3.1D, iv).

This work. In this chapter I describe the application of MVR-PCR to both single MS32 alleles, and pairs of alleles simultaneously mapped directly from genomic DNA. The information obtained is considered from the biological standpoints of minisatellite allelic diversity, mutation and evolution and also in a forensic context, in terms of its potential uses for individual identification and parentage testing. The experiments described comprise a major and ongoing part of the research carried out in this laboratory and several individuals contributed to the data collected. These are: A.J. Jeffreys, A. MacLeod, K. Tamaki, D.G. Monckton, M. Allen and myself. This work has been published (Jeffreys *et al.*, 1991a, 1994). Figures contributed by these colleagues are acknowledged in the relevant figure legends.

Results

1. MVR-PCR

Single-allele MVR-PCR. The MS32 allele sizes of a large number of unrelated Caucasians had already been estimated by Southern blot length analysis of *Alu*I digested genomic DNA (Wong *et al.*, 1987). Single MS32 alleles of known size were isolated from the genomic DNA of some of these individuals by preparative gel electrophoresis of the appropriate size fraction of an *Mbo*I digest. Following MVR-PCR at limited cycle number (18-20), **a**-type and **t**-type PCR products from each allele were resolved side-by-side by agarose gel electrophoresis, detected by Southern blot hybridisation with radiolabelled MS32 probe and visualised by autoradiography. The interspersion pattern of the two repeat unit types was revealed by complementary ladders of PCR products extending ~3kb (100 repeat units) into each allele (Fig. 3.2A). These single-allele MVR maps could be readily scored and presented as binary codes which were fully consistent with those determined by partial digestion with *Hae*III (data not shown). Occasionally, a rung on the MVR coding ladder failed to be amplified by either MVR-specific primer (arrowed positions in individual 3, Fig. 3.2A), indicating the presence of null repeats, presumably containing additional sequence variant(s) 3′ to the A/G site, which block priming by either primer. 1.6% of repeat units scored from 32 separated Caucasian alleles were null, or **O**-type repeats, compared to 72.9% **a**-type repeats and 25.5% **t**-type repeats. **O**-type repeats tended to cluster within a limited number of alleles (see Fig. 3.6) and corresponded to both *Hae*III-cleavable and *Hae*III-resistant repeat units (Jeffreys *et al.*, 1991a). With additional PCR cycles, binary codes could be determined from PCR products directly visualised on ethidium bromide stained gels, though over-amplification and collapse of minisatellite PCR products (Jeffreys *et al.*, 1988b) limited coding to ~25 repeat units (data not shown, see eg. Jeffreys *et al.*, 1993).

Diploid MVR-PCR. Performing MVR-PCR on total genomic DNA produced a diploid profile comprised of the superimposed maps of the two individual alleles. These profiles could be reliably scored at least 50 repeat units into the minisatellite array (Fig. 3.2B). For individuals with alleles containing both **a**-type and **t**-type repeats, the diploid map can be described by a ternary code in which each rung on the ladder can be coded as; **1** (both alleles **a**-type at that position, **aa**), **2** (both **t**-type, **tt**) or **3** (heterozygous, **at**). The two tracks generating the ternary 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 rung positions which are heterozygous **O**-type repeats. The presence of **O**-type repeats creates three additional coding states, namely **4** (**aO**), **5** (**tO**) and **6** (**OO**). Bands corresponding to states **4** and **5** are half the intensity of those for states **1** and **2**, while code **6** positions, which are rare, appear as a gap on the ladder (eg. individual 6, Fig. 3.2B). Coding states **4**, **5** and **6** will also be generated beyond the end of the shorter of a pair of alleles, since the code above this position will be derived from only one allele (eg. individual 4, Fig. 3.2B). No PCR products will appear beyond the end of the longer allele, generating a **66666**.... code (eg. individual 7, who is homozygous for an 84 repeat allele, Fig. 3.2B). Positions which can not be scored reliably, for example those corresponding to PCR products too small to hybridise efficiently with the probe, are designated as ambigous by a "**?**". The diploid codes of individuals 4 and 5 (Fig. 3.2B) had several null repeats at identical positions, suggesting that they may share a common allele. Assuming this to be the case it was possible to deduce single-allele codes for these individuals at all positions except those where both were heterozygous (Fig. 3.2C).

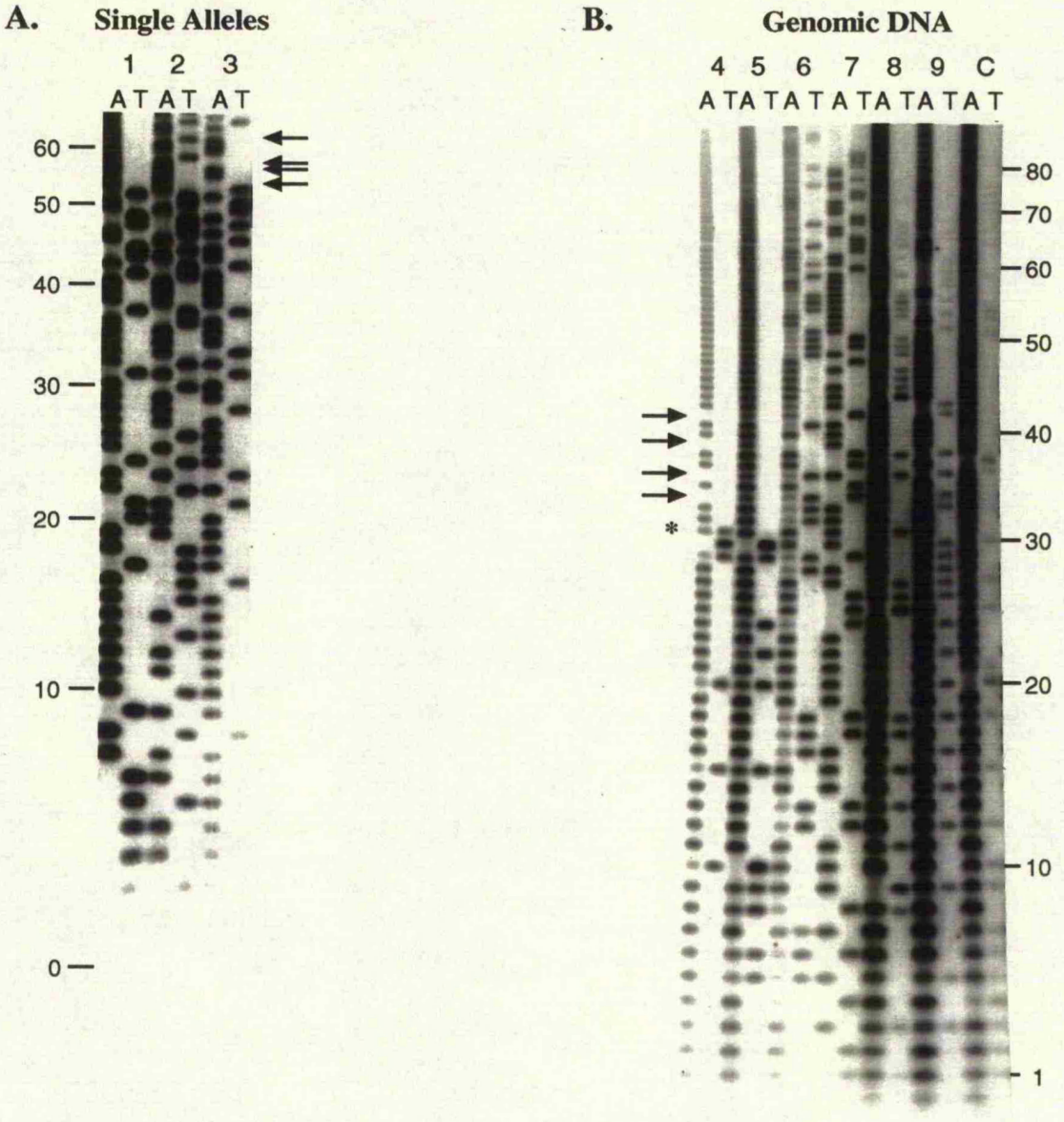
Figure 3.2. Examples of MS32 repeat coding by MVR-PCR.

MVR-PCR was performed on 100ng samples of genomic DNA (4-9 & C), or equivalent amounts of 5-19 kb long single alleles, isolated from genomic DNA by preparative gel electrophoresis of size fractions from *Mbo*I digests (1-3). DNA samples were amplified for 18 cycles using 32-TAG-A (A) or 32-TAG-T (T) in the presence of high concentration of primers 32D and TAG. PCR products were separated by agarose gel electrophoresis and detected by Southern blot hybridisation. (see Materials and Methods for detailed description of MVR-PCR and Table 3.1 for primer sequences.)

A. MVR-PCR on single alleles. Code positions are defined with reference to a standard of known code (not shown). The first few repeat positions in single-allele coding are weakly detected by Southern blot hybridisation and are not visible. Null or O-type repeat units in allele 3, which do not amplify with either MVR-specific primer, are arrowed. (Figure kindly provided by A.J.Jeffreys)

B. MVR-PCR on genomic DNA. Diploid scoring on total genomic DNA commences at the second repeat (code position 1), since the first repeat unit is weakly detected and cannot be scored reliably. This start position for reading the code was confirmed by running a standard individual of known code (individual C) on all gels. Ternary MVR codes are scored as **1** (**aa**) **2** (**tt**) and **3** (**at**). Individuals 4, 5, 6 and 7 also show O-type repeat in one of their alleles, generating code **4** (**aO**) positions (arrowed in individual 5), detected as a relatively faint **A** track band with no band in the **T** track and code **6** positions (same positions arrowed in individual 4) where there is no band in either track. Individual 4 has a short allele of 31 repeat units, as shown by loss of codes **1**, **2** and in particular **3** above the asterisked position and the presence of only code **4** (**aO**, faint **A**) code **5** (**tO**, faint **T**) and code **6** (**OO**, no product) repeats, similar to the separated allele profiles. The presence of this short allele was confirmed by conventional Southern blot hybridisation analysis of genomic DNA, probed with MS32 (data not shown). Individual 7 is homozygous (or hemizygous, with a deletion of the whole locus from one chromosome), having code states 1 and 2 only. This was confirmed by Southern blot hybridisation, which revealed a single band as predicted, and by MVR-PCR using an alternative flanking primer, which also gave a homozygous code (data not shown). The allele in this individual is 84 repeat units long and the diploid code above this position can be seen to terminate in a string of code "**6**s" (ie. non-existent repeats).

C. Deduction of single-allele codes from the ternary codes of individuals sharing a common allele. It is suggested by the identical positions of null repeat units in the diploid codes of individuals 4 and 5 that they share a common allele. Assuming this to be the case it is possible to deduce allele codes for the other alleles in these individuals at all positions except those where both individuals are code 3, these ambigous positions are denoted by a "?".

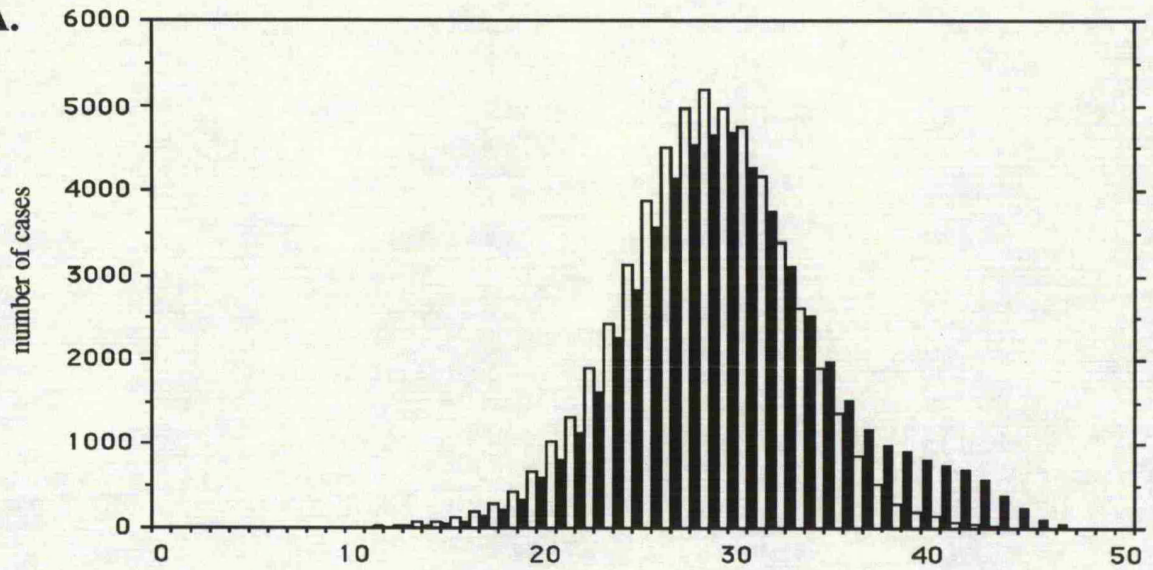


C. Deduction of single allele codes from diploid codes of individuals 4 & 5.

	5	10	15	20	25	30	35	40	45	50
4. alleles 1+2	11111	11113	11113	11113	11111	11132	34464	64464	46444	44444
5. alleles 1+3	11113	31332	11113	11113	13131	11132	11141	41141	14111	11111
allele 1	aaaaa	aaaat	aaaa?	aaaa?	aaaaa	aaa?t	aaa0a	0aa0a	a0aaa	aaaaa
allele 2	aaaaa	aaaaa	aaaa?	aaaa?	aaaaa	aaa?t	t<<<			
allele 3	aaaat	tattt	aaaa?	aaaa?	atata	aaa?t	aaaaa	aaaaa	aaaaa	aaaaa

Chapter 3 Figure 2

A.



B. Most similar:

```
                  10        20        30        40        50        60
                  .         .         .         .         .         .
English  413233322311331111111131111131111111111111111111111111113111111133
         :   x        xx       x                                        x       xx
Mormon   113223322311323111111111111131111111111111111111111111111111111111
```

Most different:

```
                  10        20        30        40        50        60
                  .         .         .         .         .         .
English  331111323334131413335333413153332141133331314323131 13
         xxxxxxxxxxx:xxx:xxxxxxxxxxx:xxxxx:x:xxxxx:x:xxxxxx::x
Japanese 223333131211223132446454645464446454645464445544644444455444454544
```

Figure 3.3. Individual variation in MS32 diploid MVR codes.

Codes extending for at least 50 repeat units were determined for 251 unrelated Caucasians (177 British, 20 French, 48 Utah Mormons, 2 Amish, 4 Venezuelan) and 83 Japanese. If there was any doubt about the coding state at any given repeat position in an individual, the position was scored as "?". Only 0.3% of code positions (59 in 20,702 scored) were scored as "?" and were ignored in database searches.

A. Filled bars show the number of ternary code differences seen over the first 50 repeat units, determined for every pairwise comparison of individuals (55,611 pairwise comparisons in total). Open bars indicate the number of differences after removal of band intensity information, scoring codes **1** (**aa**) and **4** (**aO**), and codes **2** (**tt**) and **5** (**tO**) as indistinguishable. Mean numbers of differences per pair of individuals were 30.1 and 27.9 respectively.

B. Examples of the most similar and most dissimilar pairs of individuals over the first 50 repeat units, with discordances marked with "x", or ":" for discordances which rely solely on intensity differences. The complete code scored for each individual is shown, together with additional differences beyond repeat unit 50.

This figure was adapted from Jeffreys *et al.*, (1991).

Ternary code variability. Initial diploid MVR-PCR typing of 334 unrelated Caucasians, followed by ternary code comparison, showed that there were on average 30 code mismatches per pair of individuals over the first 50 repeat units (Fig. 3.3A). 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 MVR codes were dominated by code **1** (**aa**) positions (Fig. 3.3B), indicating that all four alleles in these two individuals were composed largely of **a**-type repeats; such homogeneous alleles have been noted previously (Jeffreys *et al.*, 1990). The most dissimilar pair of ternary codes arose where one individual had a short allele, creating a diploid code dominated by the rare codes **4**, **5** and **6** (Fig. 3.3B). In total 7.8% of individuals contained short (<50 repeats) alleles, which were identified by dosage with good reliability, these had lengths ranging from 19 to 44 repeat units. Short alleles do not occur with equal frequency in all populations; for example, 5.6% of Caucasians typed had short alleles, compared with 23% of Japanese.

2. Applications of ternary code information

Allele analysis through diploid coding. The extreme variability in diploid codes at MS32 is a reflection of underlying allelic diversity, itself a product of high mutation rate to new length alleles. It is of primary importance to be able to efficiently and accurately define allelic structures, in order to investigate this allelic variability and the mutation mechanisms that give rise to it. Although the structures of individual size-separated alleles can be determined by MVR-PCR, this approach is not very efficient requiring a large quantity of starting DNA (≥5μg) as well as previously obtained information on allele size. Furthermore, it is extremely difficult to apply to individuals with two similarly sized alleles and impossible in the rare case of individuals with different alleles of the same size. In such cases alleles can be obtained by single molecule dilution followed by reamplification (Monckton & Jeffreys, 1991). Single alleles can also be recovered from genomic DNA using differential PCR amplification of the whole minisatellite array from individuals heterozgous for a small amplifiable and a larger non-amplifiable allele (Jeffreys *et al.*, 1988b). However, both these strategies are limited to those relatively scarce MS32 alleles small enough to be amplified in their entirety.

An alternative means of accessing allele structures uses logic to deduce single-allele codes from the diploid codes of parents and their offspring. For a family with a single child, the binary codes of all alleles segregating within that family are extracted sequentially along each position of the diploid code. This can be performed manually, but in the interests of efficiency and accuracy was done by a computer program that was part of the MS32 diploid code database analysis software (written in Microsoft QuickBasic™ by A.J. Jeffreys). For each position, the codes of the father, mother and child are checked in a look-up table to determine whether exclusions exist, and if not, to determine which repeat unit types were transmitted from each parent to the child. For example, if the paternal and maternal alleles transmitted to the child are called 1 and 3 respectively, and the non-transmitted alleles 2 and 4, and the father is **1** (**aa**), mother **3** (**at**) and child **3** (**at**), then no exclusions exist and the repeat unit at that position on each allele is given by allele 1, **a**; allele 2, **a**; allele 3, **t**; allele 4, **a**. In contrast, codes **1** (**aa**), **2** (**tt**) and **2** (**tt**) from father, child and mother respectively would indicate a paternal mutation/exclusion. In this way the repeat unit type at each position can be unambigously defined, except at those positions where mother, father and child are all heterozygous,

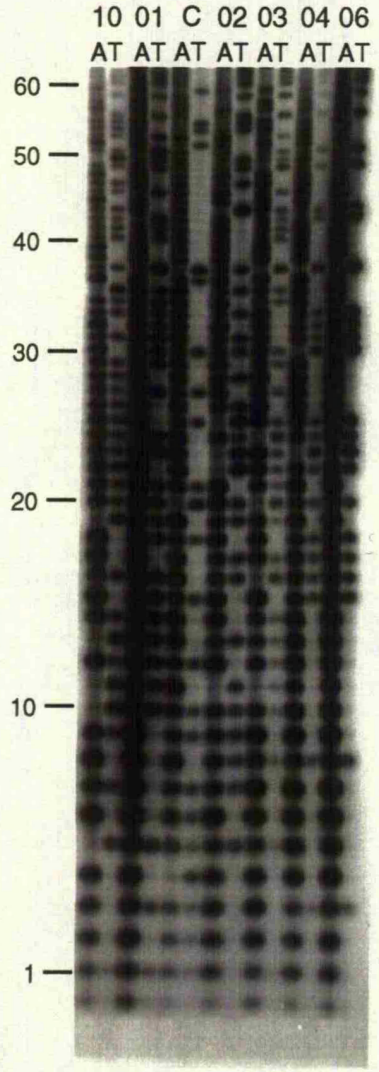
Figure 3.4. Reconstruction of the MVR codes of individual MS32 alleles by pedigree analysis.

A. Diploid MVR-PCR maps of members of CEPH family 1423 run alongside a standard of known code (C). Scoring commences at position 1 as indicated.

B. The ternary codes of father (142301), mother (142302) and a single child (142304). The paternal and maternal alleles transmitted to the child are labelled 1 and 3, and the non-transmitted alleles 2 and 4. Binary codes for all parental alleles can be deduced, assuming no recombination between parental alleles, at all positions except where the father, mother and child are heterozygous (code 3). Such ambiguities are indicated by "?".

C. Complete reconstruction of allele maps using codes from additional children from CEPH family 1423. The children show three different diploid codes, corresponding to three of the four possible combinations of parental alleles. Using these it is possible to unambigously define every code position along each parental allele. Haplotypes were extracted using software written by A.J. Jeffreys in VAX BASIC V3.4.

A. CEPH Family 1423



B. Single child

Individual	Alleles	MVR code position 0 5 10 15 20 25 30 35 40 45 50 55 60
Father CEPH142301	1 2	31313 11113 33333 31333 31331 13133 33333 12113 32113 11332 13113 11313...
Mother CEPH142302	3 4	11113 13333 51311 33133 12232 31313 31313 13111 11211 31331 11333 14313...
Child CEPH142304	1 3	11311 11311 41113 33313 13233 11112 33311 12111 13311 11313 11113 11313...
	allele 1	?aataa aaaaa aaaat ?ata? aat?a aaaat ?t?aa ataaa ataaa aa?at aaaa? aaaa?...
	allele 2	?taaat aaaat tttta ?aat? taa?a atata ?a?tt ataat ttaat aa?tt ataa? aata?...
	allele 3	?aaaaa aataa Oaaaa ?taa? att?t aaaat ?a?aa ataaa aataa aa?aa aaaa? aata?...
	allele 4	?aaaat atatt tataa ?aat? att?t tataa ?a?at aaaaa aataa ta?ta aatt? aOaa?...

C. Multiple children

Individual	Alleles	MVR code position 0 5 10 15 20 25 30 35 40 45 50 55 60
Father CEPH142301	1 2	31313 11113 33333 31333 31331 13133 33333 12113 32113 11332 13113 11313...
Mother CEPH142302	3 4	11113 13333 51311 33133 12232 31313 31313 13111 11211 31331 11333 14313...
Children:		
CEPH142303	2 3	31113 11313 53331 23132 33323 13133 11133 12113 33313 11233 13112 11212...
CEPH142304,06	1 3	11311 11311 41113 33313 13233 11112 33311 12111 13311 11313 11113 11313...
CEPH142310	2 4	31112 13132 23231 31123 33333 33331 31332 13113 33313 31323 13333 14313...
	allele 1	?aataa aaaaa aaaat aataa aataa aaaat tttaa ataaa ataaa aaaat aaaaa aaaaa...
	allele 2	?taaat aaaat tttta taatt taata atata aaatt ataat ttaat aattt ataat aatat...
	allele 3	?aaaaa aataa Oaaaa ttaat atttt aaaat aaaaa ataaa aataa aataa aaaat aatat...
	allele 4	?aaaat atatt tataa aaata attat tataa tatat aaaaa aataa taata aatta aOaaa...

Chapter 3 Figure 4

code 3 (at), these positions are scored as "?" (Fig. 3.4A). For families with more than one offspring, each having different combinations of parental alleles, the incomplete haplotypes of each parental allele can be determined separately for each child and then combined to define the ambiguous positions and thus give complete haplotypes (Fig. 3.4B). To guarantee unambigous haplotype extraction the minimum requirement is diploid codes from the mother, father and two children who share one allele in common. Although the use of family groups is limited both by the availability of DNAs and the high *de novo* mutation rate, which will lead to parental exclusions in ~2% of children, this approach was sucessfully used to determine the structures of Caucasian MS32 alleles from a large number of CEPH families, as well as Cellmark paternity trios and pedigrees from immigration casework (Jeffreys *et al.*, 1991a).

Allelic variability. In an initial survey the structures of 337 Caucasian MS32 alleles were defined, using either electrophoretically separated alleles, or more simply by extracting the MVR haplotypes of all four parental alleles from pedigree data as explained above. Haplotype comparison revealed that the vast majority (326) of these were different. 316 alleles were detected only once in the alleles surveyed, with 9 alleles sampled twice and one allele detected three times. The maximum frequency of any allele at this locus in Caucasians is therefore very low (3/337 = 0.009). If all alleles were equally rare, Poisson analysis indicates that ~3500 different MS32 alleles must exist in Caucasians to give this sampling frequency distribution. Given the MS32 mutation rate of ~1% per gamete and the current world population size of ~5 x 10^9 individuals the true level of allelic diversity in humans is likely to be enormous, with perhaps >10^8 different alleles. MVR mapping has the potential capacity to distinguish between all of these alleles; using codes a, t and O MVR-PCR can in theory distinguish between 350 (~7x10^{23}) different allelic states using information from the first 50 repeat units alone (Jeffreys *et al.*, 1991a).

Although the precise structure of most alleles is different, it had been observed previously that different alleles can share large, similar or identical, blocks of repeat units which are assumed to be closely related. These zones of haplotypic similarity may be used to make alignments between different alleles, forming groups within which alleles are assumed to share a recent common ancestor (Jeffreys *et al.*, 1990). All 326 different alleles were therefore compared, using computer programs based on staggered alignment or dot matrix analysis, to identify groups of alleles which showed significant similarities in their repeat maps (Fig. 3.5). 47% of alleles could be classified into 32 different groups in this way; each group contained 2-22 significantly related alleles, while none of the remaining 174 alleles showed any detectable matches with other alleles in the database (Jeffreys *et al.*, 1991a). Since this analysis some 500 alleles have been added to the database increasing the minimum estimate for the number of alleles at this locus in the Caucasian population to 6800. These new alleles have extended existing groups, created new ones and in some cases provided links between groups appearing previously to be unrelated. Examples of groups of related alleles are shown; these single-allele codes were derived from diploid code comparisons of families typed by myself in the initial pedigree survey (Fig. 3.5).

Most significantly, and in accordance with previous observations (Jeffreys *et al.*, 1990), variation between aligned alleles within these groups shows polarity, with the majority of interallelic differences in repeat copy number and variant repeat interspersion pattern clustering at the extreme beginning of the tandem array. This appears to hold true for all alignable groups. These data are consistent with the hypothesis that there is a mutational hotspot active in

Figure 3.5. Examples of groups of related Caucasian MS32 alleles.

Diploid MVR-PCR was performed on DNAs from pedigrees of the CEPH panel and paternity trios from Cellmark Diagnostics. Single allele codes were deduced from the ternary codes of mother, father and at least one child. MVR codes were analysed with software written by A.J. Jeffreys (Jeffreys *et al.*, 1991 & unpublished). Alleles sharing regions with closely related structures were identified by two approaches. Originally every pairwise comparison of repeat unit haplotypes was made for all alleles in the database. For each pair of alleles, comparisons were repeated for alleles misaligned up to 10 repeat units out of register. The proportion of matching positions and the contribution of **a**-type repeats to these matches, was calculated. A screening process then selected allele pairs which had a high proportion of matches, other than those due to the presence of large numbers of **a**-type repeats, for further examination. These were predicted to be more likely to represent *bona fide* matches, in contrast to alleles largely homogenised for **a**-type repeats which show misleading match frequencies (Jeffreys *et al.*, 1991). Later a pairwise dot matrix analysis of each allele code with all other allele codes was performed, searching for perfect 9 repeat matches. Matches consisting of more than 5 **a**-type repeats were discarded, while allele pairs showing a cumulative match score of greater than 20 over the best two diagonals were saved. Verification of the authenticity of these matches and final alignment of allele groups was done by eye.

The major group-specific region of related haplotype shared between alleles within a group is shown in red. Positions of divergence are shown in black. Regions of sub-group homology are shown in blue; they are also underlined where more than one region of sub-group homology is present in the same group. a = **a**-type repeat; t = **t**-type repeat; 0 = null or **O**-type repeat. Ambigous positions were not scored and are denoted by a "**?**", these often occured in alleles mapped using single offspring. Arrows (--->) above an allele highlight regions of intraallelic internal repetition. A row of stops (.....) indicates unknown haplotype extending beyond the end of the mapped region. Gaps (-) have been introduced to improve alignment.

Chapter 3 Figure 5

this region (Jeffreys *et al.*, 1990). A closer inspection of allele maps reveals zones of internal tandem duplication (arrowed in Fig. 3.5) similar to those already noted (Jeffreys *et al.*, 1990), suggesting that USCE or replication slippage may be involved in minisatellite evolution. There are often internal differences between regions with an otherwise identical haplotype; these mostly comprise small insertion/deletion events, the switching of single a-type and t-type repeat units and "conversion" between a-type, t-type and O-type repeats (probably due to the incorporation of additional sequence variants) without change in repeat copy number or disruption of alignments (eg. group 3, Fig. 3.5). Alignments within a group sometimes break down towards the 3′ end of the mapped region (eg. group 2, Fig 3.5), but without knowledge of full 3′ haplotypes it is not possible to determine whether this represents complete loss of homology beyond this point. This may be due to the occurence of large, but rare, internal deletions (Jeffreys *et al.*, 1990) of alleles within a group. Alternatively, this may represent the operation of interallelic exchanges, for example, the internal interallelic transfer of repeats into one of a pair of otherwise related alleles, or the transfer of large blocks of repeats between the 5′ ends of previously unrelated alleles. Determination of the 3′ MVR structures of such groups will be required to distinguish between these possibilities.

Mutation analysis. Southern blot allele length analysis of the families of the CEPH pedigrees, where parentage is beyond dispute, revealed 4 mutation events at MS32 (Armour *et al.*, 1989b). However, due to the problem of resolving alleles of similar length by agarose gel electrophoresis, which becomes increasingly difficult as allele size increases, it remained possible that mutation events involving small size changes had been overlooked and hence that the mutation rate to new length alleles had been underestimated. Analysis of the diploid MVR codes of parents and their children provided a means by which such events could be detected, since alleles changing in size by only a single repeat unit will throw the normal diploid code register out of alignment, thus creating discrepancies between the diploid code of a child with a mutant allele and the parent from which that allele was derived. Therefore, to quantify the mutation rate accurately, the diploid MVR codes of all the offspring of the CEPH families were compared with those from their parents. Seven children had codes showing multiple parental exclusions, indicating the presence of a mutant allele. These included all four germline mutation events previously identified by allele length analysis (Armour *et al.*, 1989b), and three new events involving single repeat unit (+29bp) additions (Jeffreys *et al.*, 1991a). In each case, the parental origin of the mutant allele could be defined by the presence of code positions specifically excluding one parent. Maternal and paternal mutations were observed at a similar frequency. It was possible to deduce the MVR map of each mutant allele from the diploid codes of non-mutant children (eg. Fig. 3.6) and thus examine germline mutation events in terms of internal structure changes. This analysis was extended to a further 75 families (in 60 of these DNA from only one child was available), but no further mutation events were detected. The 360 children typed from all families represent 720 gametes, giving a mutation rate of 7/720 = 1% per gamete detectable by MVR-PCR.

All seven mutation events detected were associated with an increase in repeat copy number, in most cases involving a small number of repeat units (Table 3.2), strongly suggesting ($p = 0.016$) a directional bias in the mutation process. Although MS32 alleles have an average length of 200 repeat units, the mutation events observed were extremely clustered showing preferential localisation to the first few repeat units of the allele. More detailed MVR analysis (Jeffreys *et al.*, 1994) later revealed that six mutations involved the exchange of repeats within the first 14 repeats while the remaining event was found to be a large subterminal duplication of the 34 repeat units between repeats 15 and 48. (Table 3.2). MVR-PCR is directed to the 5′ end of the array and may therefore be expected to bias

Individual	Alleles	MVR code position												
		5	10	15	20	25	30	35	40	45	50	55	60	
Father 141601	1 2	31333	13111	22331	33211	21131	13113	31311	34123	33311	33331	11223	11111	.....
Mother 141602	3 4	21433	32133	33333	11133	31333	11331	31111	11111	11113	11311	11311	11311	.....
	1	?taaaa	ataaa	ttaaa	tataa	taata	ataat	tataa	t0ata	tttaa	attta	aatta	aaaaa	.....
	2	?aattt	aaaaa	tttta	attaa	taaaa	aaaaa	aaaaa	aaatt	aaaaa	taaaa	aattt	aaaaa	.....
	3	?taata	ataaa	attaa	aaaaa	aaaaa	aaaaa	taaaa	aaaaa	aaaat	aataa	aataa	aataa	.....
	4	?ta0at	ttatt	taatt	aaatt	tattt	aatta	aaaaa	aaaaa	aaaaa	aaaaa	aaaaa	aaaaa	.....
Child 141606 mutant+3		21131	12111	13311	33311	11133	33113	21131	11111	11113	11311	13211	11331	.....
Exclusions:				p		p	e	e	p			e p		
	3	?taata	ataaa	attaa	aaaaa	aaaaa	aaaaa	taaaa	aaaaa	aaaat	aataa	aataa	aataa	.....
Deduced mutant allele		?taaaa	ataaa	aaaaa	tttaa	aaatt	ttaat	taata	aaaaa	aaaaa	aaaaa	attaa	aaata	.....
	1	?taaaa	ataaa	ttaaa	tataa	taata	ataat	tataa	t0ata	tttaa	attta	aatta	aaaaa	.....
				XXX	X									
	2			?aa	tttaa	aaatt	ttaat	taata	aaaaa	aaaaa	aaaaa	attaa	aaata	.....

Figure 3.6. Detection and characterisation of an MS32 mutation event by analysis of ternary MVR-PCR codes.

This example shows one of seven CEPH ternary code pedigree analyses that revealed the presence of a child with a mutant MS32 allele. The structures 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 7 positions of incompatability (red). 4 of these are specifically paternal exclusions (p), the other 3 are ambiguous exclusions (e) and do not indicate the parental origin of the mutant allele. There are no maternal exclusions, indicating that the child has inherited a mutant paternal allele and non-mutant maternal allele. The diploid code of the child is compatible with inheritance of 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. Comparison with paternal alleles 1 (green) and 2 (red) shows that the mutant allele commences with the code of allele 1 and then switches to the code of allele 2 after two a-type repeats of unknown origin (blue). 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. The first repeat unit does not produce a detectable signal in standard diploid MVR-PCR and therefore is not scored, for this reason single-allele codes commence with an ambigous repeat unit (?) This figure was adapted from Jeffreys *et al.*, (1991).

Table 3.2. Summary of the properties mutant alleles detected at MS32 in 286 offsping from the CEPH panel of families.

Mutant [1]	Parental origin (progenitor)	Detected on Southern blot [2]	Change in repeat copy number	Exchange position (repeats from hypervariable end)[3] Donor	Recipient	Putative exchange mechanism[4]
a	Maternal	-	+1	1	1	intraallelic
b	Maternal	-	+1	1	1	intraallelic
c	Maternal	+	+34	14	14	intraallelic
d	Maternal	-	+1	14	14	intraallelic
e	Paternal	+	+13	1	1	interallelic
f	Paternal	+	+3	1	1	interallelic
g	Paternal	+	+2	2	2	interallelic

1. Mutant assignments as in Jeffreys *et al.*, (1991a)

2. This survey detected all allele length change mutations previously detected by Southern blot analysis of *Alu*I digests of genomic DNA (Armour *et al.*, 1989b) plus three new hitherto undetected mutations resulting from gains of a single repeat unit. In all cases, the change in repeat unit copy number of the the mutant allele appearing in the progeny is consistent with allele-length changes detected by Southern blot analysis (not shown).

3. Exchange positions as shown in Jeffreys *et al.*, (1994). Recipient allele defined as the mutant allele detected in the progeny (the same as the donor for intraallelic events). Donor defined as allele from which additional repeat units appearing in the recipient were derived.

4. Mechanisms were infered from initial MVR mapping experiments. They have since been confirmed by sequencing (Monckton, 1993) and more detailed MVR analysis (Jeffreys *et al.*, 1994). Significantly ($p = 0.03$) all interallelic mutants are paternal.

This figure was adapted from Jeffreys *et al.*, (1991a, 1994)

pedigree analysis to the detection of mutation events here. The fact that the four events independently detected by previous allele length analysis were also located in this region confirmed the observation of mutational polarity. Since this is the region of the minisatellite already shown by allele alignment to exhibit maximum allelic variability, these results verify the hypothesis of the presence of a localised mutation hotspot. Mechanisms of unequal exchange (either presumptive sister chromatid exchanges or recombinations) between "donor" and "recipient" were invoked by defining the "recipient" allele as the parental allele contributing the relatively invariant end of the locus to the resulting new mutant allele. The seven mutations observed were entirely conservative, with no loss of information from the recipient allele. Two of the paternal events, for the first time, suggested the involvement of interallelic exchange in the generation of the new allele (Jeffreys *et al.*, 1991), this was later confirmed by further MVR analysis, which also showed that the remaining paternal germline mutation was interallelic (Jeffreys *et al.*, 1994) . I will return to a discussion of detailed characterisation of mutation events and consideration of possible mutational mechanisms in Chapters 6 and 7.

Determination of heterozygosity at MS32. Allele length estimates of minisatellites based on Southern blot analysis are error prone and make it impossible to distinguish between true homozygotes and heterozygotes where both alleles are of similar, or identical, size (Devlin *et al.*, 1990; Chakraborty & Jin, 1993). Enzymatic MVR mapping has been used to show that alleles of the same, or similar, length from individuals scored as homozygotes by Southern blot length analysis may in fact have different internal structures; however, this technique is limited to short PCR amplifiable alleles (Monkton & Jeffreys, 1991). Diploid codes provide an objective method for identifying homozygotes with alleles of any size, since an individual with two identical MS32 alleles will yield a diploid MVR map exactly the same as would be obtained for each allele singly, comprising only code types 1 and 2, or 6 if null positions are present, with no heterozygous a/t positions (eg individual 7, Fig. 3.2B.). Three individuals (one French, two Japanese) out of 334 surveyed showed homozygosity by this criterion, suggesting a mean MS32 heterozygosity level of 99.1% (Jeffreys *et al.*, 1991a). As predicted, these individuals showed a single band on Southern blot hybridisation of genomic DNA and gave a homozygous diploid code following MVR-PCR using an alternative flanking primer (data not shown). These results indicate that these individuals were not heterozygous for a second allele rendered unamplifiable by a 32D primer mismatch in the flanking DNA, a "null" allele too small to detect by Southern blot hybridisation (Armour *et al.*, 1992b), or two different length alleles with identical 5′ ends. However, it remains formally possible that these individuals are heterozygous for a deletion of the whole MS32 locus on one chromosome. By contrast, the majority (8/10) of apparently single band individuals initially detected by hybridisation with MS32 were in fact heterozygous for similar, or identical, length alleles, as shown by diploid coding (Jeffreys *et al.*, 1991a). Given the enormous allelic diversity observed at this locus it is likely that the majority of homozygotes arise through recent consanguinity; the number of different MS32 alleles observed suggest that the true heterozygosity in outbred individuals should be in excess of 99.9% (Monckton, 1993a). In a few cases an individual showing different length alleles by Southern blot hybridisation, gave an apparently homozygous diploid code as would be expected if different length alleles with the same 5′ structure were present (data not shown). However, MVR-PCR with an alternative flanking primer (or at a lower annealing temperature) showed that in all such cases this was due to the presence of heterozygous variant positions in the flanking DNA which caused an allele-specific flanking primer mismatch and thus prevented amplification from this allele (eg. Individuals heterozygous for the *Hinf*I site spanned by the 32D primer (Fig. 3.1A).

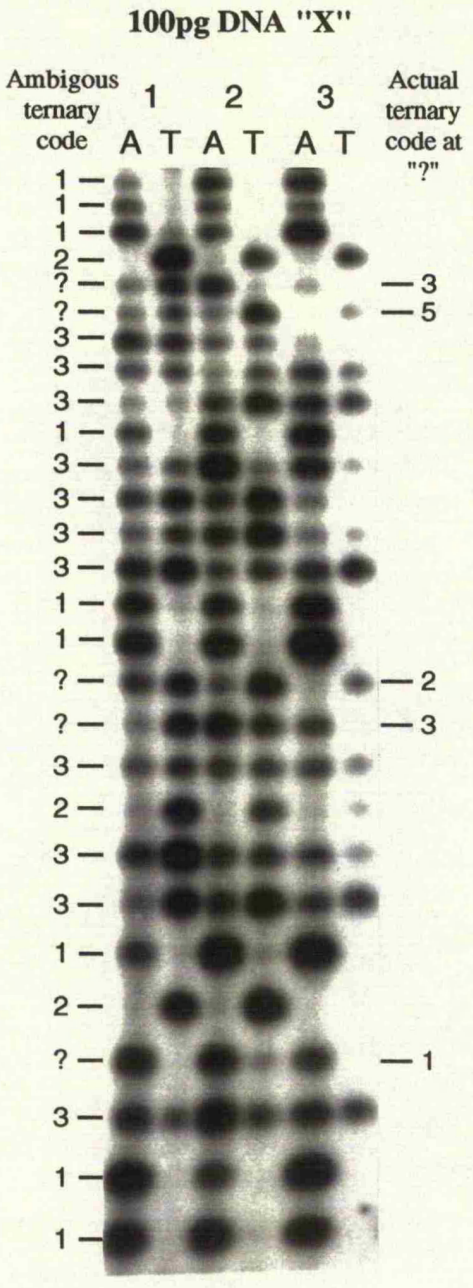


Figure 3.7. Ternary code information recoverable from trace amounts of human genomic DNA by MVR-PCR.

A "mystery" individual, whose diploid code was unknown, was selected from a collection of 450 people for whom diploid codes had already been determined. Three pairs of 100pg aliquots of genomic DNA from this individual (1, 2 and 3) were amplified by MVR-PCR for 28 cycles using 32-TAG-A (A) and 32-TAG-T (T) and the PCR products resolved on an agarose gel and detected by Southern blot hybridisation. Zero DNA controls gave no signal (not shown). The incomplete ternary code of the unknown individual was established from those repeat positions which gave concordant results in all three analyses. Repeat positions which gave ambigous typing results were scored as "**?**", as shown. These were probably caused by stochastic band "drop-out" and mispriming events due to the low number of input molecules. Note that band intensity fluctuations prevent the discrimination of codes **1** (**aa**) and **4** (**a0**), and codes **2** (**tt**) and **5** (**t0**) . The incomplete ternary code determined over the first 45 repeat positions was then compared with the 450 already in the database, ignoring the distinction between codes **1** and **4** and codes **2** and **5**. The correct individual was identified as the only database entry which showed a complete match with the incomplete MVR code. The complete diploid code was used to define the code states of previously ambigous positions, shown to the right of the figure.

Forensic applications. I tested the sensitivity of diploid MVR-PCR by generating MVR maps from increasingly dilute aliquots of human DNA from an individual whose diploid code had already been determined, but which was unknown to me. I then compared these diploid codes against those in the database to see whether this information could be used to successfully identify the correct individual. It was possible to generate reproducible profiles that unambigously identified the correct individual down to ~10ng DNA (data not shown) indicating that diploid MVR-PCR has comparable sensitivity to SLP analysis. Below this level band dropout, spurious bands and apparently random fluctuations in band intensity can arise, probably due to mispriming events and the stochastic loss of PCR products from the small number of input molecules. However, reliable consensus diploid codes were nevertheless obtained by comparing replicate MVR profiles from sub-nanogram amounts of genomic DNA (Fig. 3.7). MVR-PCR can also be applied to partially degraded DNA, since it recovers information from any DNA fragments long enough to include the flanking primer site, plus at least some repeat units (data not shown). Additional information could be recovered by using a flanking primer immediately adjacent to the start of the tandem-repeat array (eg. 32O, Fig 3.1A; see Table 3.2 for sequence). While degraded DNA will yield a truncated diploid code, this will still be compatible with database searches (a minimum of 17 repeats, ~500bp, are needed to distinguish all individuals in the current diploid code database), although with reduced discriminating power.

Discussion

Allelic diversity and mutation. MVR mapping at MS32 has provided a clear view of the extraordinary level of allelic variability that can exist at a human hypervariable locus; this is orders of magnitude higher than that detectable by allele length analysis. MVR mapping by PCR is technically much simpler and more efficient than the enzymatic method it superceded, is applicable to MS32 alleles of any length and provides information from genomic DNA which can be encoded digitally, both widening the possible applications of MVR mapping and allowing greatly increased throughput of data. The successful application of this technique has enabled a much more powerful investigation of allelic structure, evolution and turnover mechanisms at MS32, making this hypervariable locus the best characterised human minisatellite to date. Analysis of allelic structure and the use of allele alignment to identify closely related alleles has confirmed the previously observed polarity in variation. Furthermore, the application of MVR-PCR to genomic DNA in families with a *de novo* germline mutant allele has enabled a preliminary investigation of the mutational processes that give rise to such variability. Previous studies of minisatellite variation showed that allele length changes were largely, if not completely restricted to single alleles (Jeffreys *et al.*, 1990; Wolff *et al.*, 1988, 1989), suggesting USCE or replication slippage as possible mutation mechanims. The data presented here show that there is a mutation hotspot operating over the 5′ terminus of MS32 and provide the first direct evidence that interallelic recombination or gene conversion may have a major role in minisatellite instability. If the interallelic events were true recombinations this would represent a dramatic example of a human recombination hotspot, and revitalise earlier speculation that minisatellites may be actively involved in chromosomal processes such as homologue recognition, synapsis and meiotic recombination (Jeffreys *et al.*, 1985a, 1991a). The general bias towards size gain and the paternal origin of all putative interallelic mutation events may be significant consequences of the mutation process giving rise to most variation in repeat copy number at this locus, while the qualitative difference in the types of mutation event seen in the maternal and paternal germline may reflect the preferential involvement of different mutation processes in different contexts (see Chapters 6 & 7 for further discussion of possible mutation mechanisms).

Population analysis. Since the data presented in this chapter were collected a new means of generating single-allele MVR maps has been developed. This technique is based on the same principle as that used for MVR discrimination, but directs the terminal 3′ mismatches of pairs of allele-specific primers to polymorphic sites of base substitution in the MS32 5′ flanking DNA. By using the appropriate allele-specific flanking primer in MVR-PCR, the individual map of each allele can be generated direct from genomic DNA of individuals heterozygous at these position (Monckton *et al.*, 1993). The use of such primers has facilitated the rapid acquisition of further single-allele MVR codes and the allele database has now been extended to >1000 alleles. Some of these fall into already existing or new alignable groups while others are, as yet, apparently unrelated to any other allele in the database. As the allelic structures from individuals belonging to different populations are defined and grouped, analysis of their population distribution will become possible. The high (~1% per gamete) germline mutation rate to new alleles predicts that no allele present before the human radiation (~100,000 years ago, Cann *et al.*, 1987; equivalent to ~5000 generations, or 50 mutations) is likely to have survived unmutated to the present and therefore that alleles would be expected to be largely population specific. The majority of alleles mapped so far are Caucasian, Japanese and Afro-Carribean, and already some population specificity among alignable groups can be seen (Monckton, 1993b). However, the use of MVR maps as population-specific markers and the application of MVR data to population analysis has potential drawbacks. The high mutation rate at this locus means that most allelic structures surveyed have arisen relatively recently in human evolution, a period of time during which considerable admixture has occured. The problem may be further compounded by the presence of alleles largely homogenised for a-type repeat units (Jeffreys *et al.*, 1990, 1991a) and the criteria used for aligning alleles. Alleles where the mapped region is largely composed of a-type repeat units have been found in all populations surveyed and in one case two such alleles seem to have arisen by convergent evolution (Monckton, 1993b). Although the allele alignment software has been designed to ignore matches between stretches of a-type repeats, the threshold for selection of provisionally alignable regions is arbitrary and the final alignment and grouping of alleles are made subjectively, by eye. In this process some of the alignments suggested by the software are rejected as coincidental and small deletions or insertions are introduced in other alleles at positions selected to improve alignments. It is also possible that alignments of closely related alleles beyond the mapped region will be missed, for example it would not be possible to align diverged forms if an allele had undergone a deletion extending further into the allele than the mapped region at its 5′ end. If apparently population-specific groups of alleles are defined, a further problem will lie in distinguishing their origins; do they represent new structures arising after population divergence which have risen to significant population frequency by genetic drift, or ancient haplotypes differentially lost in some populations? These factors mean that a much more complete survey of allele structures from a wide range of populations will have to be made before any underlying relationships between them become apparent. Even if this monumental task is achieved it is likely that it will be extremely difficult to make may make reliable inferences about human population divergence, except in certain well defined cases involving closely related, homogeneous and localised groups, (eg. highly inbred and isolated populations, K. Tamaki, personal communication).

Forensic analysis: advantages. Diploid MVR-PCR provides a novel and simple method for generating unambiguous and highly discriminatory digital information directly from human DNA and has therefore potentially provided a very exciting new DNA profiling system. It offers many advantages over currently used DNA typing systems that involve allele length measurements and overcomes many of the other limitations associated with existing DNA typing technologies. Most importantly MVR-PCR obviates the problem of DNA profile matching

and allele size measurement based on error-prone DNA fragment length estimation. Code generation does not require standardisation of electrophoretic systems, is immune to gel distortions and band shifts, and does not require side-by-side comparisons of DNA samples on the same gel. Furthermore, MVR profiles contain considerable informational redundancy enabling code authenticity to be checked. In addition, the design of the MVR system means that it is also suitable for automation. For example, the use of non-isotopically labelled primers together with in-gel, or microcapillary tube, detection by laser scanning may allow the scoring of MVR codes directly into a computer database.

Preliminary investigations suggest that MVR-PCR is particularly well suited to many types of forensic analysis. It is applicable to very small quantities of DNA, as well as to degraded DNA and, importantly, to mixed DNA samples of the type often encountered in forensic casework (eg. victim plus rapist DNA recovered from semen-bearing vaginal swabs), particularly if pure DNA from one of the two individuals (eg. victim) is available. The latter scenario has been simulated in DNA mixing experiments, which showed that 10% admixture can be detected, and that comparison of the MVR-PCR profile of the "victim" with the profile of the mixed DNA samples can yield an ambiguous diploid code of the "rapist" (Jeffreys *et al.*, 1991a). Possible genotypes of the "assailant" can be deduced at all repeat positions where the "victim" is not code 3 (at), by checking for A- or T-track specific bands present in the mixture, but not in the "victim". 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 $2x10^6$ 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 $2x10^6$ "false suspects" failed to show any exclusions (99.9993% mean exclusion rate) (Jeffreys *et al.*, 1991a). The recovery of information from mixed samples has now been made easier and much more sensitive (down to 1% admixture) by the application of allele-specific MVR-PCR (Monckton *et al.*, 1993).

Current forensic applications of minisatellite loci calculate the probability that a match is significant based on estimates of allele frequencies, derived under the assumption that the population to which the typed individual belongs is at Hardy-Weinberg equilibrium. As mentioned in Chapter 1, evidence for population substructuring at some loci has called this assumption into question. The MVR mapping results presented here support the assertion that the majority of homozygote excess is accounted for by the inability to resolve different alleles of similar or identical size. The forensic application of MVR-PCR offers a way of avoiding this problem and may therefore be able to extract DNA profiling from the legal/statistical quagmire in which it has become embroiled. MVR codes are ideal for objectively determining a match between a forensic sample and a criminal suspect, since the determination of match frequencies using MVR-PCR would use observed phenotype frequencies, rather than genotype frequencies deduced from limited population databases under assumptions of Hardy-Weinberg equilibrium. Since there are far in excess of 3500 different MS32 alleles, ($>> 6x10^6$ diploid codes) it is likely that very large databases can be constructed before any significant saturation of MVR code types occurs. With such a large number of MS32 alleles in the human populations studied, diploid code variability, which is governed by the number and frequencies of different MS32 alleles in human populations, is expected to be very high, to the point of approaching individual specificity (over 500 diploid codes from unrelated individuals have been generated to date in this laboratory and all

are different). Such databases will 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. The digital coding system will allow the rapid generation of the very large communal population and investigative DNA databases required for forensic investigations, including profiles from previous offenders, missing persons, unsolved casework and population surveys. Furthermore, the standard format of MVR codes will facilitate rapid database searches and allow the dissemination of the information they contain between laboratories, using computers. Paradoxically, despite the potential to overcome some of the current legal problems of DNA profiling, the extraordinary power of this technique may also create new ones. Great care will have to be taken when presenting evidence of a complete code match between DNA from a scene of crime sample and a criminal, so as to leave it up to the jury to establish guilt or innocence.

MVR-PCR could also be used in parentage testing since the diploid code of a child containing a contribution from a non-parental allele will frequently show exclusionary mismatches with the parent, (see eg. Fig. 3.6). To determine the effectiveness of MVR-PCR in excluding non-fathers in parentage testing (non-maternity is seldom an issue), 28,635 Caucasian mother-child-nonfather trios were created (by a computer program written in Microsoft QuickBasic™ by A.J. Jeffreys) from the MVR diploid 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 remainder directionally ambiguous. 98.9% of non-fathers showed at least one paternal-specific exclusion, and 99.8% showed at least one exclusion in total (paternal-specific plus ambiguous) (Jeffreys *et al.*, 1991a).

Forensic analysis: limitations. Although MVR-PCR potentially offers considerable improvements in the field of forensic DNA typing, there are some potential problems in the forensic application of MVR code data that will have to be addressed before its use is sanctioned by the forensic community and the courts.

The use of diploid MVR codes to unambigously identify individuals does not hold for pairs of siblings who will have an approximately 1/4 chance of sharing the same parental alleles and therefore ternary MVR code, a limitation which also applies to other DNA markers. Diploid codes are highly reproducible and null positions in diploid codes generated from the same individual provide useful additional information for unambigous identification. The removal of null information to increase scoring reliability, by converting codes 4, (aO) and 5, (tO) to codes 1 (aa) and 2 (tt) respectively, does not significantly reduce the power of individual identification (Jeffreys *et al.*, 1991a; Fig. 3.3A). However, it is particularly important to be able to identify null positions accurately in parentage testing, since misscoring may lead to false exclusions between diploid codes. Null repeats are scored by virtue of lack of MVR information and although missing rungs in single-allele codes are easily detected, correctly identifying heterozygous null positions (code 4, aO; 5, tO) from the interpretation of band intensity differences in diploid codes is less reliable. Furthermore, this scoring does not distinguish between different types of null repeats which can be created by any possible repeat unit variant that prevents priming by the MVR-specific primers. The sequencing of null repeat units and subsequent design of null-specific primers has made it possible to unambigously identify most of these repeats, in particular a 28bp repeat unit (N-type) that accounts for 87% of null repeats in Caucasians (Tamaki *et al.*, 1992). When an N-type specific MVR primers is used in MVR-PCR it positively identifies the positions of these repeat types. Null repeat units are rare which makes them particularly useful for

identifying related allelic structures and adds to the power of paternity testing. If null information is removed from paternity analysis there is a significant drop in the mean proportion of non-fathers excluded (Tamaki *et al.*, 1992).

The possible effects of mutation must also be taken into account with any DNA typing system based on highly variable loci . As with the scoring of null repeats, this is not a significant factor in individual identification, but has implications for the application of MVR-PCR to paternity testing. The high *de novo* mutation rate at MS32 would be expected to give paternal mismatches of the kind seen in Fig. 3.6 in approximately 1% of cases, limiting the efficiency of paternity testing. In such cases it may be possible to use single-allele MVR-PCR to deduce the structure of the mutant allele and therefore show the relationship between father and child. In the absence of informative sites for allele-specific MVR in the child, a possible solution to this problem would be to use the codes from paternal and maternal alleles to try and reconstruct the observed mutant diploid code. Since it appears that most *de novo* germline mutations at MS32 are small one or other paternal allele would probably only have to be displaced by a few repeat units to acheive the correct registration and hence distinguish mutation from non-paternity. However, although such analyses might be highly suggestive, they would still require calculation of the statistical probability of obtaining a related allele to that of the alleged father from another individual in the population to provide convincing evidence for paternity.

It may be possible to overcome many of the limitations outlined above by extending MVR-PCR to other highly variable loci. Each additional locus would add to the statistical power of the data generated, and different loci, with slightly different characteristics, may complement each other in these analyses. This could increase the probability of distinguishing between siblings, reduce the potential problems raised by the need to correctly score null positions at MS32 and provide a means of distinguishing mutation from non-paternity. The development of MVR-PCR at a second hypervariable minisatellite locus is described in Chapter 4.

Chapter 4

DESIGN AND APPLICATION OF AN MVR-PCR SYSTEM AT THE HUMAN HYPERVARIABLE MINISATELLITE MS31

Summary

Minisatellite variant repeat unit mapping by PCR (MVR-PCR) has been successfully used to assay the interspersion pattern of variant repeat units along MS32 minisatellite alleles. I have now applied MVR-PCR to a second hypervariable locus with the aims of evaluating the general applicability of the approach, further investigating minisatellite allelic variability and mutation (see Chapters 5 & 6) and providing the additional discriminatory power needed for some of the potential forensic applications of the technique. The D7S21 (MS31A) minisatellite locus was chosen because it is one of the few hypervariable minisatellites isolated in this laboratory that conforms to all the necessary criteria for single-allele and diploid MVR-PCR. It is highly polymorphic, with 98% heterozygosity, based on Southern blot allele length analysis, and has repeat units which have two common sites of internal sequence variation, but no known length variants. An MVR system was developed to assay the interspersion pattern of variant repeat types along MS31A alleles in the same way as at MS32. Digital diploid codes can be produced from total genomic DNA, and both MS31A and MS32 can be simultaneously encoded by duplex MVR-PCR. The successful development of MVR-PCR at this second locus provides the opportunity to compare allelic structure and variation between different minisatellites and also enhances the potential for forensic application of this technique.

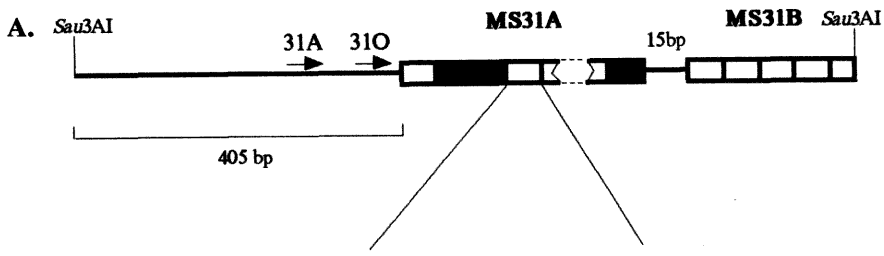
Introduction

The need to MVR map additional loci. The potential forensic uses of MVR-PCR are limited if it is only applied to one locus, for example MS32. Firstly, it is not possible to distinguish closely related individuals who share the same alleles and therefore diploid codes. Secondly the high germline mutation rate at MS32 (approximately 1% per gamete) and the presence of null repeat units complicate parentage analysis using MVR-PCR, since *de novo* mutation events or mis-scored heterozygous null positions could lead to false parental exclusions. Furthermore, MS32 diploid codes from unrelated individuals are occasionally rather similar, which could reduce the possibility of unambiguous individual identification, particularly from the first few repeat units in badly degraded DNA. To overcome these limitations and extend the range of application of MVR-PCR, particularly to parentage analysis, it is necessary to apply the technique to additional hypervariable minisatellite loci. Several loci which can be MVR mapped and have similar variability to that seen at MS32 should yield combined digital codes approaching complete individual specificity, thus providing the statistical rigour required for forensic application of this DNA profiling system. Ideally it should be possible to amplify these loci simultaneously, so that they can ultimately be used for multiplex MVR-PCR. Mapping of additional minisatellites is also needed to develop a fuller

view of their origins and evolution, and in particular to compare and contrast the turnover mechanisms that generate and maintain the enormous allelic variability of these loci. This may enable us to discover whether common processes are shared by loci exhibiting the same general properties, or whether these have arisen by different mechanisms and behave differently, despite superficial similarities.

Selection of suitable loci. A highly informative locus for MVR-PCR has to conform to certain criteria. It must be polymorphic, preferably with an allele length heterozygosity greater than 95%, to ensure that most or all alleles are rare. Repeat unit heterogeneity must not be too extensive, and sites of variation must be suitably positioned to allow the design of repeat unit specific primers. Finally, all primers used for MVR mapping must work at the discriminatory annealing temperature of the MVR-specific primers. While almost all human minisatellites show sites of internal sequence variation to which repeat-unit-specific primers can be directed, the majority, including the panel of 49 hypervariable human minisatellite loci isolated in this laboratory (Wong *et al.*, 1987; Armour *et al.*, 1989a), also have common repeat unit length variants. At some of these loci, for example MS205 (D16S309, Royle *et al.*, 1992), where most alleles are small enough to PCR amplify in their entirety, it is possible to isolate and map large numbers of single alleles (Armour *et al.*, 1993). Diploid MVR-PCR, however, is only possible if variant repeats with different repeat lengths do not exist at high frequency, since interspersed repeats of different lengths will throw the MVR ladders of the two constituent alleles out of register, making at least part of the diploid coding ladder uninterpretable. At the rare loci, for example MS32, where most or all repeat units are of the same length, a diploid map of the interspersion patterns of repeats from two alleles superimposed can be generated from total genomic DNA and encoded as a digital diploid code (Jeffreys *et al.*, 1991a). We have identified a further minisatellite, MS31A, which conforms to these criteria and therefore may be suitable for diploid MVR-PCR; MVR-PCR mapping primers have been designed for this locus (Neil & Jeffreys, 1993). Another hypervariable minisatellite locus pλg3, where most repeat repeat units are 37bp, but 33bp variants also exist, is also being evaluated for MVR-PCR (T. Guram, personal communication).

The MS31 locus (D7S21). The human minisatellite SLP MS31 was cloned from a λ library of size fractionated (5-15kb) *Sau3*AI inserts of human genomic DNA. It detects a hypervariable minisatellite locus with a *Hinf*I allele size range from 3.5kb to 13kb and an observed allele length heterozygosity of 98% in Caucasians, that reflects extreme variability in tandem repeat copy number (Wong *et al.*, 1987; Armour *et al.*, 1989b). This allelic diversity is the result of a 1% germline mutation rate to new length alleles, as estimated by Southern blot length analysis (Jeffreys *et al.*, 1988a). This locus has been localised to the subterminal region of the short arm of chromosome 7 (7p22-pter) by *in situ* hybridisation and linkage mapping (Royle *et al.*, 1988). The cloned *Sau*3AI fragment containing this locus (clone MS31) has been sequenced; it has 405bp of 5′ flanking DNA, followed by a number of 20bp minisatellite repeats (MS31A), these are separated by 15bp from an adjacent minisatellite (MS31B) which has a 19bp repeat unit of a sequence distinct from MS31A (Armour *et al.*, 1989a) (Fig. 4.1). There is no sequence information on the MS31B 3′ flanking DNA since this repeat array runs to the end of the cloned sequence, indicating the presence of a *Sau3A*I site, either in one of its repeat units, or in the first few bases of flanking DNA. Evidence from genomic restriction mapping suggested that MS31B is dimorphic in Caucasians (Armour *et al.*, 1989a), indicating that almost all of the length variation at D7S21 is due to tandem repeat copy number variation at MS31A.



B. **MS31A Repeat units** (20bp)

5′ AC / GT CCACCTCCCACAGACACT 3′

MS31A MVR-PCR Primers

Name	Sequence 3′ to 5′
31-TAG-AC	CA
31-TAG-GC	CG GGTGGAGGGTGTCTGTGAggcctggtacctgcgtact
31-TAG-AT	TA
31-TAG-GT	TG GGTGGAGGGTGTCTGTGAggcctggtacctgcgtact
31-TAG-A	A
31-TAG-G	G GGTGGAGGGTGTCTGTGAggcctggtacctgcgtact
TAG Primer	aggcctggtacctgcgtact

Figure 4.1. Organisation of the MS31 (D7S21) locus and MVR-Primers.

A. Variant repeats MVR mapped at MS31A are indicated by white (a-type) and red (t-type) boxes; broken box indicates the presence of varying numbers of repeats. PCR primers are indicated by arrows (see Table 4.1 for sequences of flanking primers).

B. MS31 repeat unit sequence and MVR-specific PCR primers. Red bases indicate t-type repeat unit and complementary positions in the repeat-specific primer.

Table 4.1. MS31A MVR-PCR primers.

Primer uses	Primer names	Primer sequences[3] 5′-3′	Final conc. μM
5′ flanking	31A	CCCTTTGCACGCTGGACGGTGGCG	1.000
	31O	GGAGGGGCCATGCCGGGGAC	1.000
MS31A four-state mapping[1]	31-TAG-AC	tcatgcgtccatggtccggAGTGTCTGTGGGAGGTGGAC	0.050
	31-TAG-GC	tcatgcgtccatggtccggAGTGTCTGTGGGAGGTGGGC	0.025
	31-TAG-AT	tcatgcgtccatggtccggAGTGTCTGTGGGAGGTGGAT	0.010
	31-TAG-GT	tcatgcgtccatggtccggAGTGTCTGTGGGAGGTGGGT	0.005
MS31A two-state mapping[2]	31-TAG-A	tcatgcgtccatggtccggAGTGTCTGTGGGAGGTGGA	0.050
	31-TAG-G	tcatgcgtccatggtccggAGTGTCTGTGGGAGGTGGG	0.025

1. Four-state MVR-PCR maps the positions of all possible variants of the two commonly polymorphic positions in the MS31A repeat unit.

2. Two-state mapping only assays the C/T variant position and displays two types of repeat unit. In both two-state and four-state mapping, amplifications were carried out in the Geneamp 9600 thermal cycler (Perkin Elmer Cetus), with denaturing at 94° for 30 sec, annealing at 68° for 30 sec and extension at 70° for 2.5 min for the first 10 cycles, after which the extension time was incremented by 20 sec per cycle for a further 10 cycles. Cycling was followed by a chase of 68° for 1 min and 70° for 10 min. See Materials and Methods for detailed MVR-PCR mapping protocol.

3. Uppercase denotes flanking primer sequences and the region of MVR-specific primers complementary to the MS31A repeat unit. Lowercase denotes the TAG sequence, which is identical to that used for MS32 MVR-PCR (see Table 3.1 for sequence).

Characterising variation at MS31A. Sequence analysis of the tandem array in the cloned MS31A allele revealed that, like most other sequenced minisatellite loci, there are polymorphic positions within the consensus repeat unit, generating minisatellite variant repeat units (MVRs). MS31A is atypical in that all repeat units so far characterised at this locus have the same length (20bp) (Wong *et al.*, 1987). Its repeat unit has two adjacent sites of base substitutional polymorphism; G/A followed by C/T. The G/A site creates an *Alw*NI RFLP, which was used along with a constant *Mln*I site in early attempts at enzymatic MS31 MVR mapping (Armour, 1990c). Although these studies revealed variation in internal allelic structures at this locus, this analysis was prone to even more difficulties than had been encountered using this technique at MS32, and was therefore abandoned (Armour, 1990d). The attributes of the MS31A repeat unit make this the only other characterised hypervariable minisatellite we know of, besides MS32, which satisfies all the MVR-PCR criteria (including diploid analysis) mentioned above. These features suggested that MS31A would be an ideal candidate for internal mapping of minisatellite repeat unit variation by applying the same MVR-PCR technique as used for MS32 (Jeffreys *et al.*, 1991a).

Design of an MVR-PCR system for MS31A. The structure of the D7S21 locus (Fig. 4.1) indicates that the 5′ end of MS31A alleles is more amenable to MVR mapping than the 3′ end. The proximity of MS31B to MS31A makes it difficult to design 3′ flanking PCR primers which could be used to efficiently map alleles from this end, particularly in light of evidence that there are polymorphic positions in the 15bp intervening sequence which such primers would have to span (Armour, 1990e). Another advantage of assaying internal repeat unit variation at the 5′ end of MS31A is that the existence of any polymorphic sites found in the flanking DNA can later be exploited in the design of allele-specific flanking primers, as has been successfully achieved at MS32 (Monckton *et al.*, 1993). Accordingly MVR-PCR primers corresponding to all four combinations of the two adjacent polymorphic positions were designed and used to generate MVR maps into the 5′ ends of single size-separated alleles (see Table 4.1 for primer sequences). From the preliminary results obtained (data not shown) it was possible to deduce that the C/T polymorphic site had roughly even numbers of "C" and "T" variants, making it ideal for MVR analysis. Furthermore, the C/T site is directly accessible for analysis by mapping into the 5′ end of MS31A from the adjacent flanking DNA. Access to this site from the 3′ flanking DNA would require degenerate MVR-PCR primers spanning the G/A variant position within the repeats. It was not possible to estimate the proportions of "A" and "G" variants at the adjacent site, since repeat units starting with AC and GC were not distinguished in this pilot experiment. Two MVR-specific primers, 31-TAG-A and 31-TAG-G, were therefore designed to MVR-map "C" and "T" variant repeats at the 5′ end of MS31A alleles. They each comprise 19 nucleotides complementary to the minisatellite repeat unit, preceded by a TAG sequence identical to that used for MVR-PCR at MS32 (Jeffreys *et al.*, 1991a; see Table 3.1 for TAG sequence) and differ only in either "G" (31-TAG-G), or "A" (31-TAG-A) at their 3′ terminus, to detect "C" and "T" repeat unit variants respectively (see Fig. 4.1 & Table 4.1 for sequences).

This work. In this Chapter I describe both single-allele and diploid MVR-PCR at MS31. I also show that MS31 and MS32 can be mapped simultaneously direct from genomic DNA in "multiplex" MVR-PCR. The results are considered in terms of the potential forensic applications of this technique, as well as allelic diversity at the 5′ end of MS31A. This work has been published (Neil & Jeffreys, 1993).

RESULTS

Single-allele MVR-PCR. The MS31 *Hinf*I allele sizes of a number of unrelated Caucasians had already been determined by Southern blot length analysis (Wong *et al.*, 1987). 34 single MS31A alleles were isolated from the genomic DNA of some of these individuals by preparative gel electrophoresis of the appropriate size fraction of an *Mbo*I digest. These alleles were then amplified using low concentrations of one or other MVR specific primer coupled with high concentrations of the TAG primer and a fixed 5′ flanking primer 31A, (Fig. 4.1; see Materials and Methods for MVR-PCR conditions; Table 4.1 for primer sequences). MVR-PCR products were resolved side-by-side by agarose gel electrophoresis and detected by Southern blot hybridisation with radiolabelled MS31 probe. Visualisation by autoradiography revealed complementary ladders generated from each of the two MVR-specific primers, extending from the flanking site to each variant repeat unit of a particular type (Fig. 4.2A). All of the alleles mapped in this initial survey gave different MVR codes.

As with MS32, each single-allele MVR-PCR map was encoded as a string of **a**-type or **t**-type repeat units. This coding ensured compatibility with computer software developed for analysis and manipulation of MS32 MVR-PCR allele codes. **a**-type repeat units are detected by 31-TAG-A and carry the "T" base at the polymorphic C/T site. **t**-type repeats carry the "C" variant and are detected by 31-TAG-G. As with MS32, a small proportion (around 1%) of repeat units failed to amplify with either MVR-specific primer, indicating the presence of additional null, or **O**-type, variant repeats (eg. positions 48 and 49, allele 5, arrowed in Fig. 4.2A). Coding commences from the second repeat unit, since the hybridisation signal from the first repeat is often too faint to allow reliable scoring; this start position was confirmed by reference to a standard genomic DNA sample of known code run on all gels. In many cases, allele codes could be read up to 100 repeat units into the array.

At some repeat positions the band intensity is reduced compared to the others, for example the top band of the region bracketed in Fig. 4.2A. This effect is reproducible and is presumably due to the presence of additional repeat unit sequence variants which hybridise less efficiently with the MVR-specific primers, quantitatively reducing amplification efficiency and hence band intensity. Attempts to overcome this phenomenon by altering PCR parameters, for example, annealing temperature, or using substances supposed to enhance PCR specificity, for example, tetramethylammonium chloride (TMAC) (Hung *et al.*, 1990) or formamide (Sakar *et al.*, 1990), were unsuccessful (data not shown). These, as yet uncharacterised, repeat types do not affect the ability to score single-allele codes and are not distinguished from normal **a**-type and **t**-type repeats. However, they can make it difficult to score accurate diploid codes from total genomic DNA (see below).

Diploid MVR-PCR. MS31A MVR-PCR can be applied to genomic DNA to reveal the profile derived from the superimposition of MVR interspersion patterns of both alleles (Fig. 4.2B). This can be described by a digital ternary code in exactly the same way as at MS32, in which ternary MS31A MVR codes are scored as **1** (**aa**), **2** (**tt**), or **3** (**at**). However, codes **4** (**aO**) and **5** (**tO**) cannot be used in MS31A diploid coding because the existence of **O**-type repeats, combined with positions of reduced band intensity, makes it impossible to use intensity differences to distinguish a genuine hemizygous null repeat position, (**aO** or **tO**) from homozygous positions, (**aa** or **tt**) where band intensity is reduced because a repeat unit from either, or both, allele(s) is poorly amplified. Such repeats may also lead to the mis-scoring of code **3** positions as either **1** or **2**. Examples of repeat positions likely to be mis-

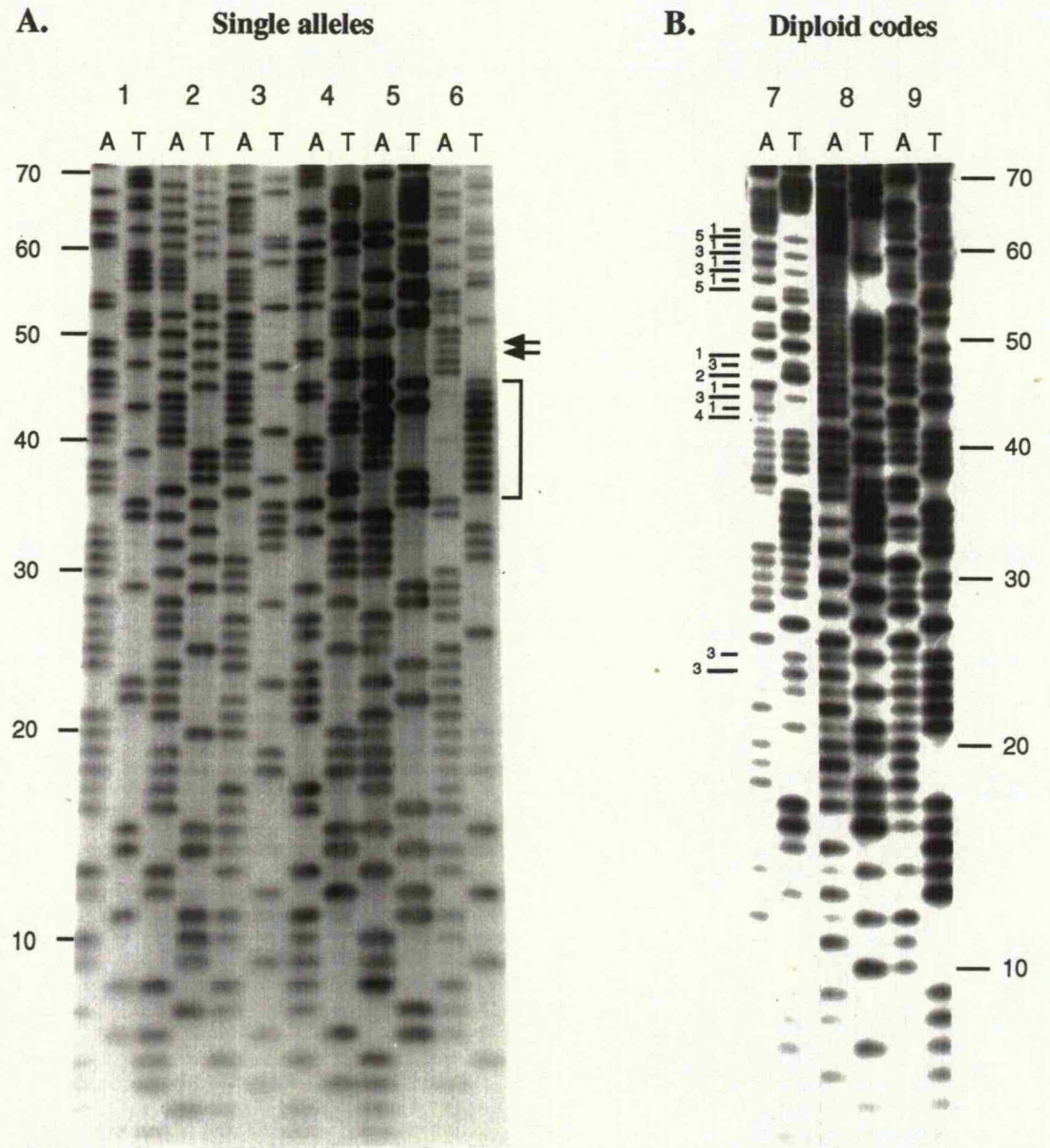


Figure 4.2. Examples of single-allele and diploid MS31A MVR-PCR.

MVR-PCR was performed on 100ng samples of genomic DNA (7, 8 & 9), or equivalent amounts of 2-10kb single alleles, isolated from genomic DNA by preparative gel electrophoresis of size fractions from *Mbo*I digests (1-6). Samples were amplified for 20 cycles using primers 31A plus TAG, together with either 31-TAG-A (A), or 31-TAG-G (T). PCR products were resolved by agarose gel electrophoresis and detected by Southern blot hybridisation. (see Materials and Methods for MVR-PCR protocol and Table 4.1 for primer sequences).

A. MVR-PCR on single alleles. The start position for coding was defined with respect to a standard of known code (not shown). The first few repeat positions are weakly detected in single-allele coding. Repeat units are scored as **a**-type (A), or **t**-type (T). Some positions fail to amplify with either tagged primer and are scored as **O**-type (eg. arrowed positions, 48 & 49 in allele 5). Not all repeat unit positions amplify with the same efficiency, leading to band intensity fluctuations (eg. **t**-type repeat intensity differences within the bracketed region of allele 6).

B. MS31A MVR-PCR on genomic DNA. Diploid codes from a father (8), mother (9) and their child (7). Diploid scoring on total genomic DNA commences at the second repeat (code position 1). The first few repeats in diploid codes are only weakly detected, but the start position for reading the code was defined by running a standard individual of known code on all gels (not shown). Ternary MVR codes are scored as **1** (**aa**), **2** (**tt**) and **3** (**at**). Examples of ambiguous positions are indicated to the left of the code from child 1, with the correct scoring for that position as deduced from single-allele codes from this individual. Misscoring of such positions can prevent deduction of single-allele codes from the diploid codes of a family.

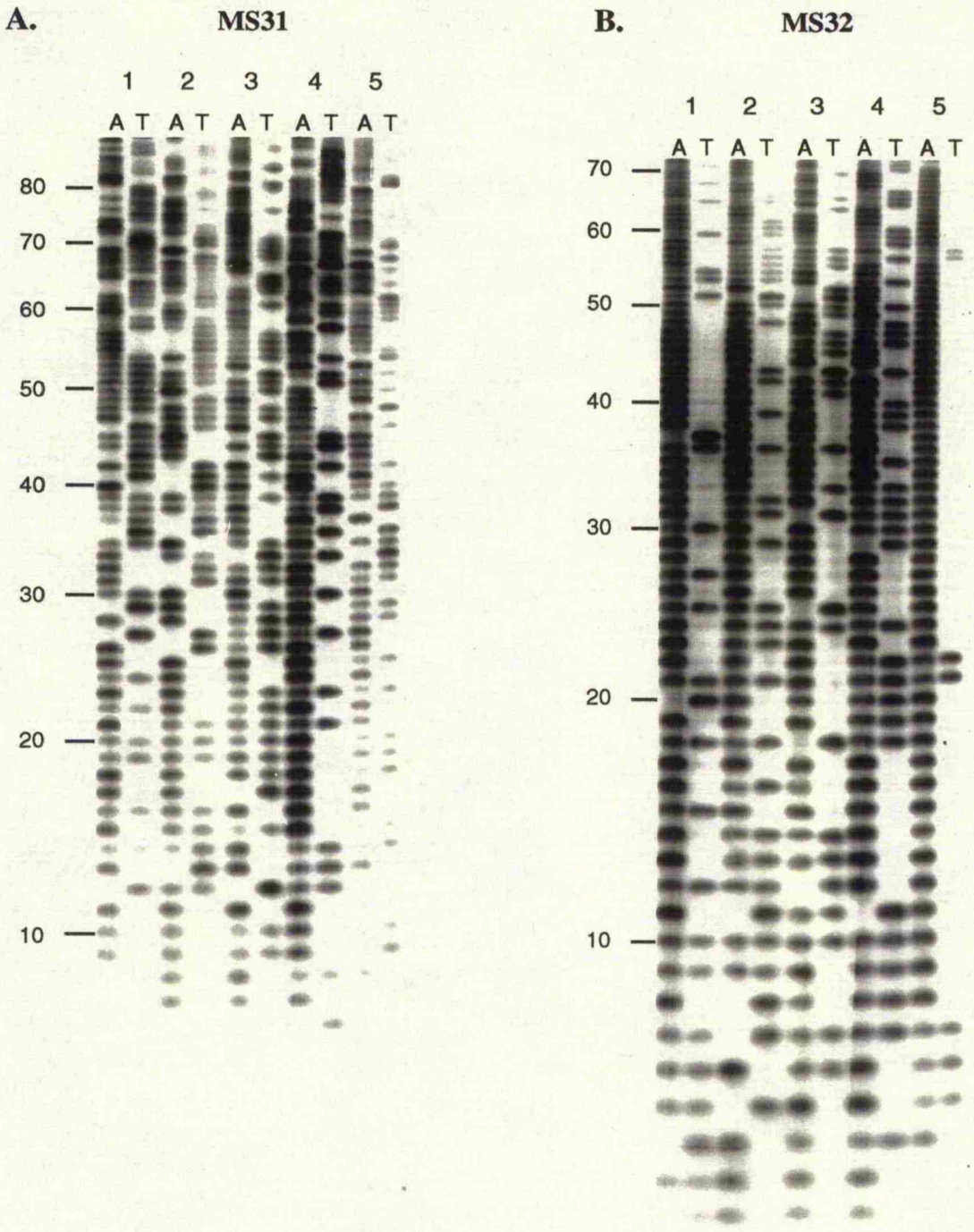


Figure 4.3. Diploid digital coding of genomic DNA by duplex MVR-PCR.

MS31A and MS32 diploid codes were simultaneously generated from the genomic DNA from 5 unrelated individuals (1-5) using flanking primers 31A and 32O and TAG plus 31-TAG-A and 32TAG-A (A), or 31-TAG-G and 32TAG-C (T) (see Table 3.1 for MS32 primer sequences; Table 4.1 for MS31A primer sequences). PCR products were resolved by agarose gel electrophoresis and detected by Southern blot hybridisation with MS31A (**A**) followed by probe removal and re-probing with MS32 (**B**).

scored in these ways are shown (arrowed in Fig. 4.2B), together with the correct scoring for that position as deduced from single-allele codes from these individuals. For this reason MS31A diploid coding could not be used for the deduction of single-allele codes or the identification of mutant alleles in families, and was therefore only applied to a few individuals.

Duplex MVR-PCR. Combinations of primers can be used to generate diploid codes from MS31 and MS32 alleles simultaneously. This "duplex MVR-PCR" has been tested using the flanking primers 31A and 32O (see Table 3.1 for 32O primer sequence), MVR-PCR primers 32-TAG-A, 32-TAG-T and 31-TAG-A, 31-TAG-G, and TAG. The same PCR parameters were used to MVR map each locus (see Materials and Methods & legend to Fig. 4.3). 31-TAG-A and 32-TAG-C were used in one PCR reaction with 31-TAG-G and 32-TAG-T in the other, to maintain the conventional order of **a**-type and **t**-type repeat unit lanes on MVR-PCR gels. Southern blot analysis by sequential hybridisation with MS31 (Fig. 4.3A), followed by MS32 (Fig. 4.3B), showed complete sets of PCR products from each locus with no evidence of inter-locus interference or cross-hybridisation, indicating that repeat units from both loci amplify, and are detected, independently.

Diploid code variability, allelic diversity and heterozygosity at MS31A. All of the diploid codes generated showed heterozygous **a/t** positions, as would be expected if this MVR system was accessing a highly variable region of this locus. However, these data were too few to be used to accurately calculate the mean observed level of heterozygosity at this locus as had been done at MS32. Instead a minimum estimate of heterozygosity was obtained from the single-allele analysis. Comparison of the 34 single-allele MVR maps obtained in this initial survey showed that all had different internal structures, again indicating considerable allelic diversity in internal MVR structure at the 5′ end of MS31A. The sampling distributions of different alleles can be used to estimate allelic diversity, $\theta = 4N_e u$, where N_e is the effective population size and u is the mutation rate. θ was determined from the number of different alleles n_a seen in a sample of i individuals, using a computer program (written by A.J. Jeffreys in Microsoft QuickBasic™). Under the infinite-allele model and assuming selective neutrality, $n_a = \Sigma_1^{2i} (\theta/\theta + i - 1)$ (Ewens, 1972) and homozygosity can be estimated by $1/1 + \theta$. A sample of 34 alleles, all of which are different gives a θ value of >540, $p > 0.95$ suggesting a heterozygosity of >99.8% [100 - (100/541)], in Caucasians. A more accurate estimation of heterozygosity at this locus, using a larger sample of alleles, is described in Chapter 5.

Discussion

Forensic analysis: advantages. The successful development of MS31A MVR-PCR provides a powerful adjunct to MS32 digital coding, particularly since both loci can now be typed simultaneously. Besides substantially increasing the speed with which reference diploid and single-allele code databases can be constructed, duplex MVR-PCR has many potentially important forensic applications. One out of four siblings are expected to share the same parental alleles and therefore diploid codes at a given minisatellite locus, diminishing the power of individual identification within families. With two unlinked loci the probability that a pair of children will share parental alleles at both is reduced to one in sixteen, thus duplex MVR-PCR increases the chances of distinguishing between siblings. This could be useful, for example, in immigration disputes, incest cases, or rape cases involving members of the same family. In forensic casework where the only available DNA is badly degraded it may only be possible to

map the first few minisatellite repeats at any locus. For instance, with an average DNA fragment length of 300bp only the first 10 repeat units of MS32 and the first 15 repeat units of MS31A would be accessible to MVR-PCR analysis. Duplex MVR-PCR would be expected to increase the amount of individual specific information which can be obtained from such a sample. In such cases it is possible that the ternary codes of two unrelated individuals at MS32 will be indistinguishable, however this is extremely unlikely to be the case at MS31A as well. MS31A codes have more heterozygous positions and can therefore be more informative over some coding regions; compare for example the MS31A and MS32 profiles of individual 5 whose MS32 code is largely dominated by repeat unit positions homozygous for a-type repeats. (see Chapter 5 for estimation and discussion of MS31A ternary code diversity). In such cases the second locus may provide important additional exclusionary power. As MVR-PCR is applied to further suitable minisatellites, multiplex MVR-PCR using three or more loci may become possible, as long as no cross-priming of repeat units occurs and PCR parameters are similar for all loci involved.

Forensic analysis: limitations. Although band intensity fluctuations at MS31A make diploid code comparison and analysis difficult, correct heterozygous null scoring is irrelevant for individual identification in a forensic context. The presence of reproducible band intensity fluctuations at MS31A may even enhance this application since it provides an additional level of variation. However, this limitation does create potential problems for applications of MS31A MVR-PCR diploid coding which rely on the interpretation of dosage, for example parentage testing or screening for mutant alleles. Paternity and relationship testing are the most important commercial applications of DNA typing and it is therefore in this arena that MVR-PCR must prove itself to be better than existing technologies, before its considerable promise as a general DNA typing system will be realised. The application of MVR-PCR to multiple hypervariable loci has the potential to overcome some of the limitations of its application in paternity analysis that were found at MS32. For example, as with SLP analysis, typing at a second locus may help to distinguish between a false exclusion due to mutation and non-paternity, since the simultaneous mutation of two independent loci will be very rare (~1/10000 for two loci with a 1% mutation rate). However, an erroneous paternal exclusions can also be caused by mis-scoring of a heterozygous null position. This is more likely at MS31A than at MS32 due to the band intensity fluctuations between individual code positions. An approach to solving this problem would be to eliminate null scoring from analyses of both loci and combine ternary code data to recover the exclusionary power lost in this process. In the absence of an extensive MS31A diploid code database it is not possible to quantify the benefit this procedure would have to the efficiency of paternity analysis, but it seems likely that MVR-PCR will have to be applied to additional loci to match, or exceed the power and efficiency of currently used techniques (A.J. Jeffreys, personal communication).

Allelic and diploid code variability. The unique single-allele and diploid MVR codes generated in this initial survey, and the minimum estimate of heterozygosity of >99% determined from MVR data, suggested that the 5′ end of MS31A has high level of variability and indicated that a large scale survey of allelic diversity at this end of the locus was warranted. However, the isolation and analysis of large numbers of single alleles would make this an extremely laborious and time consuming exercise. MS32 diploid codes from families (mother, father and at least one child) were used to deduce partial or complete MS32 single-allele codes of all four parental alleles (see Chapter 3). This approach greatly accelerated the construction of the MS32 single-allele database and also allowed the detection of subtle mutation events not detectable by Southern blot analysis, by comparing childrens' diploid codes with those of confirmed parents and looking for incompatible positions. Unfortunately this approach is not

straightforward at MS31A, since the non-uniformity of band intensities can lead to incorrect genotyping. This interferes with diploid code comparison and hence the deduction of haplotypes from pedigree data, by creating apparent parental exclusions, or incorrect allele codes. Individuals heterozygous for short MS32 alleles were identified by reduction of the diploid coding ladder to hemizygosity beyond the end of the shorter allele, with loss of heterozygous a/t repeat positions; such diploid codes may also be more difficult to score at MS31A. Sequence analysis of both the O-type repeat units and the presumed additional variants may enable additional MVR-primers specific to these repeat unit types to be designed (Tamaki *et al.*, 1992). When used together with the existing MVR-primers these may then generate ternary codes where previously null positions can be positively identified and scored reliably. These diploid codes would then be suitable for the forensic applications mentioned and could also to used to deduce the maps of single alleles segregating within families. The difficulties outlined above made it a priority to develop techniques that could efficiently access single-allele codes at MS31A. MS32 this had already been achieved with MS32 by using the more direct approach of allele-specific MVR-PCR (Monckton *et al.*, 1993; Chapter 3). I now sought to adapt this technique for use at MS31A, this work is described in Chapter 5.

Chapter 5

INVESTIGATING ALLELIC STRUCTURE AND DIVERSITY AT MS31A: FLANKING SEQUENCE ANALYSIS AND ALLELE-SPECIFIC MVR-PCR

Summary

Initial single allele and diploid maps of the 5′ end of MS31A alleles suggested a high level of structural variability, with characteristics similar to those seen at MS32. In the absence of extensive diploid code analysis I now sought to quantify allelic diversity at this end of MS31A alleles directly and compare the nature of interallelic structural variation at MS31A with that seen at MS32. In particular, I wanted to look for evidence of polarity in structural variation of the type predicted to arise from localised mutation processes. It is not possible to deduce reliable single MS31A allele codes from family diploid codes, as was done with MS32, and the alternative of PCR amplification and recovery of single alleles is limited to the minority of MS31A alleles that are small enough to amplify in their entirety to levels detectable by ethidium bromide staining. Instead, size selection of single alleles from restriction digests of genomic DNA was used to isolate several alleles for mapping, despite being tedious and labour intensive. To increase the efficiency of single-allele coding another approach, which had already been successfully used at MS32, was adapted for use at MS31A. This method exploits polymorphic positions in the DNA flanking the hypervariable end of the minisatellite to design allele-specific flanking PCR primers. These enable the mapping of single alleles directly from the genomic DNA of informative individuals. Sequencing of the MS31A 5′ flanking DNA from several chromosomes revealed three positions of base substitutional polymorphism. PCR based restriction assays and allele-specific primers were designed for each of these sites. Population analysis showed that each of these sites is in Hardy-Weinberg equilibrium, and determination of flanking haplotypes showed that there is significant, but not complete, linkage disequilibrium between them. Single-allele MVR-PCR at MS31A, using both size-separated alleles and allele-specific primers, revealed an extreme level of allelic variability, far in excess of that detectable by allele length analysis. 182 MS31A alleles were mapped and, with two notable exceptions, all were different. As at MS32, computer analysis of allelic structures revealed that several alleles share regions of related internal structure. Internal segments of code similarity were assumed to derive from a common ancestral allele and alleles which shared such regions were grouped on this basis. The flanking haplotype of alleles within these groups is usually the same. Some of these groups show polarity in allelic variation reminiscent of that seen at MS32, providing evidence for a localised variability "hotspot" at MS31A.

Introduction

Investigating allelic diversity. Because of the high mutation rates to new length alleles at hypervariable minisatellite loci like MS31A and MS32, a large number of different alleles are expected to exist in the current human population. The application of MVR mapping to single alleles at MS32 demonstrated extreme levels of allelic variability (Jeffreys *et al.*, 1990), to the extent that even length isoalleles in unrelated individuals and pseudohomozygotes could usually be distinguished (Monckton & Jeffreys, 1991). The observed variation in diploid codes at MS32 (500 codes all different) is a reflection of this massive underlying allelic diversity. However, the true number of different alleles is undoubtedly much greater than this (Chapter 3; Jeffreys *et al.*, 1991a).

Large scale surveys of allelic structures at MS32 revealed that although the majority of MS32 alleles are unique many share similar regions of MVR code haplotype, presumed to derive from a common ancestral allele, which can be used to align these alleles into groups (Jeffreys *et al.*, 1990, 1991a; Chapter 3). Such alignments show clear evidence for polarity of internal allelic variation at MS32. Within aligned groups of alleles, regions of related haplotype extend inwards from the 3′ end, while most interallelic variation is confined to the extreme 5′ end. The majority of observed mutation events at MS32 (Table 3.2) are also restricted to the 5′ end of alleles, providing evidence that polarity in allelic variation arises through the action of a local mutational hotspot. By characterising mutant and progenitor alleles and analysing their flanking haplotypes, it is possible to infer possible mechanisms for the mutation processes that contribute to the evolution of this minisatellite (Jeffreys *et al.*, 1991a, 1994).

The MS31A hypervariable minisatellite locus was selected for MVR-PCR because of its many similarities with MS32. Both have comparable heterozygosity, mutation rates and allele size ranges, and like MS32, MS31A does not exhibit extensive repeat unit length variation. The initial MVR-PCR analysis of MS31A single-allele and diploid codes (Chapter 4) suggested that the internal structure of this locus may also be analogous. With such large numbers of alleles at these loci, interallelic relationships only become apparent when many allele structures have been characterised. Therefore, in order to make a more detailed comparison of the two loci in terms of allelic diversity, structural variability and polarity in allelic variation, a large-scale survey of the internal MVR structures of MS31A alleles was necessary. Analysis of family diploid codes at MS32 allowed the deduction of single-allele codes and also revealed mutant alleles that had not been detected by Southern blot length analysis, because the size changes involved were too small (see Chapter 3). This approach was not possible at MS31A, for the reasons discussed in Chapter 4. Therefore it was necessary to perform MVR-PCR on single MS31A alleles to determine their structures and to characterise *de novo* mutation events at this locus.

Single-allele MVR-PCR. There are several procedures that can be used to access single-allele information at minisatellite loci. As shown in Chapter 4, either allele from a suitable individual can be isolated by physical separation on the basis of size, using restriction digestion followed by preparative gel electrophoresis, and then subjected to MVR analysis. However, this method cannot be applied to individuals with alleles that cannot be resolved by agarose gel electrophoresis and is too laborious and tedious to be used for a comprehensive survey of allelic structures. Further disadvantages are the requirement of a preliminary experiment to estimate allele sizes and the need for relatively large quantities of DNA (≥5μg total genomic DNA). More efficient PCR based strategies, for example single molecule dilution (Jeffreys *et al.*, 1990), or amplification to levels detectable by ethidium bromide

staining and resolution by agarose gel electrophoresis (Jeffreys *et al.*, 1988b) followed by band removal, can also be used to isolate single minisatellite alleles. The drawback with these approaches is that they can only be applied to alleles small enough to amplify in their entirety to levels where they can be visualised by ethidium bromide staining. PCR amplification of the MS31A locus is known to be difficult for all but the shortest alleles (<5kb, Armour, 1990d), of which there are few at this locus; therefore the use of these strategies in a large scale survey of allelic diversity was also precluded. A more efficient single-allele MVR-PCR technique, amenable to both population analysis and mutation screening and characterisation, was therefore required for investigation of MS31A.

One possibility was to digest DNA from individuals heterozygous for a flanking RFLP with the appropriate restriction enzyme followed by MVR mapping with a distal flanking primer, to allow amplification from only the uncut allele. An initial experiment suggested that this strategy was indeed feasible (Fig. 5.1), however, not all polymorphic sites create RFLPs, and this method would not be able to map the cut allele. A more efficient approach is to identify polymorphic positions in the DNA flanking the minisatellite and design allele-specific flanking primers complementary to the different allelic forms of such variants. Allele-specific primers are identical, except for a 3′ terminal mismatch that corresponds to the variant flanking base. At an appropriate annealing temperature this enables them to discriminate between the two allelic forms present in the genomic DNA of heterozygous individuals (Newton *et al.*, 1989). This technique can be applied to any flanking polymorphism to allow the selective mapping or amplification of single alleles direct from total genomic DNA. Allele-specific MVR-PCR can be performed on as little as 10ng of genomic DNA to directly map alleles of any length, without the need for allele separation prior to mapping, and has been successfully used to map single alleles at MS32 (Monckton *et al.*, 1993).

Allele-specific MVR-PCR is particularly useful where the alleles of interest are similarly sized and hence difficult to resolve by agarose gel electrophoresis or, if they are present in different individuals, distinguish by allele length analysis. For example, in the case of an individual inheriting mutant and non-mutant alleles of a similar size it can be used to selectively map either allele, provided there is a heterozygous flanking polymorphic position. It can also be used to identify and characterise *de novo* mutation events, by comparing single allele codes of parents and their offspring, even where mutant alleles are too close in size to their progenitors to be scored as such by Southern blot allele length comparison. If the parent contributing a *de novo* mutant allele to an individual is informative at a flanking variant position the progenitor allele can also be mapped (see Chapter 7). Besides being used to create large allele databases this technique may also have forensic applications, for example selectively amplifying one or both of the assailant's alleles from victim/assailant DNA mixtures (Monckton *et al.*, 1993).

This work. An *Alu*I RFLP at the MS31A locus had been discovered previously (Armour, 1990c). I located and characterised this polymorphism by amplifying and sequencing the MS31A 5′ flanking DNA from human genomic DNA. Two additional flanking variants were identified during this analysis and assays for each of the polymorphic flanking sites and their chromosomal haplotypes were developed. A combination of separated alleles and allele-specific MVR-PCR were used to generate a single-allele database for MS31A. 60 of the Japanese single allele codes were kindly provided by Dr. Keiji Tamaki. Malaysian alleles were typed by Dr. Chong-Lek Koh. The structures of the mapped alleles were compared to identify regions of homology between them. Some of this work has been published (Neil & Jeffreys, 1993).

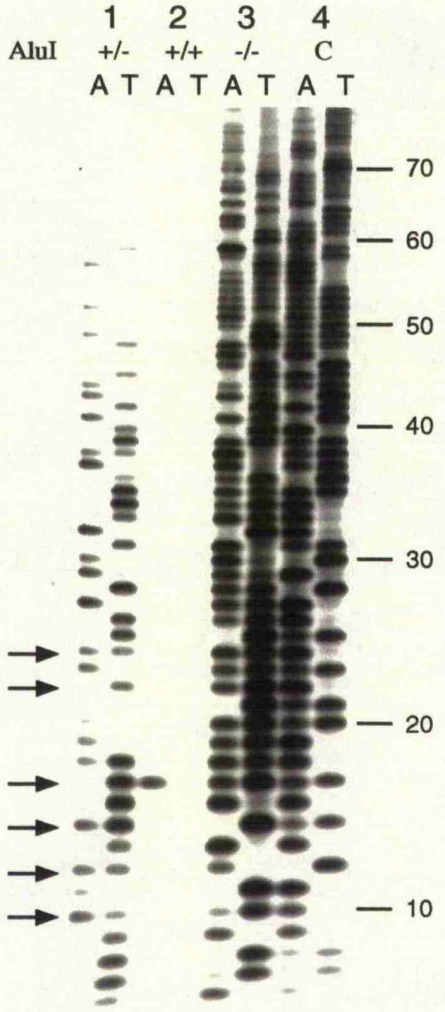


Figure 5.1. MVR-PCR following cleavage at a heterozygous RFLP.

MVR analysis of *Alu*I digested genomic DNA from 3 individuals (1-3) previously characterised as; 1, an AluI+/- heterozygote, 2 and 3 AluI+/+ and AluI-/- homozygotes respectively. 1µg of DNA from each of these individuals were digested with 10units of *Alu*I and the digests diluted 1/10 to prevent subsequent inhibition of PCR. 10ng DNA from each digest was then used to provide the template for MVR-PCR using flanking primer 31O (see Fig. 5.2 for primer sequence). A non-digested standard DNA, of known MVR code (4), was also included as a control. The products were resolved by agarose gel electrophoresis and detected by Southern blot hybridisation to radiolabelled MS31 probe followed by autoradiography. Arrows indicate diploid positions in the otherwise single-allele code from individual 1. These bands and those amplified in individual 2 indicate that digestion was not 100% complete.

Results

1. Detection, characterisation and analysis of MS31A 5′ flanking polymorphisms

Detection of an *Alu*I RFLP. In an earlier study, MS31 allele sizes were estimated by Southern blot length analysis of *Alu*I digested DNA from several unrelated Caucasians, in order to select the smallest alleles for PCR amplification of the whole locus (R. Neumann, I. Patel & A.J. Jeffreys, unpublished results). It was noted that the size of some of the PCR products was not consistent with the size determined by Southern blot length analysis, but differed by a constant amount. Restriction mapping showed that this was due to the presence of a polymorphic *Alu*I site ~400bp inside the *Sau*3AI fragment spanning the MS31 locus, positioning it either in the vicinity of the first MS31A repeat, or towards the 3′ end of the tandem repeat array, defining a variant repeat unit. The heterozygosity of this polymorphic position was estimated to be 35% in Caucasians (Armour, 1990d; data not shown).

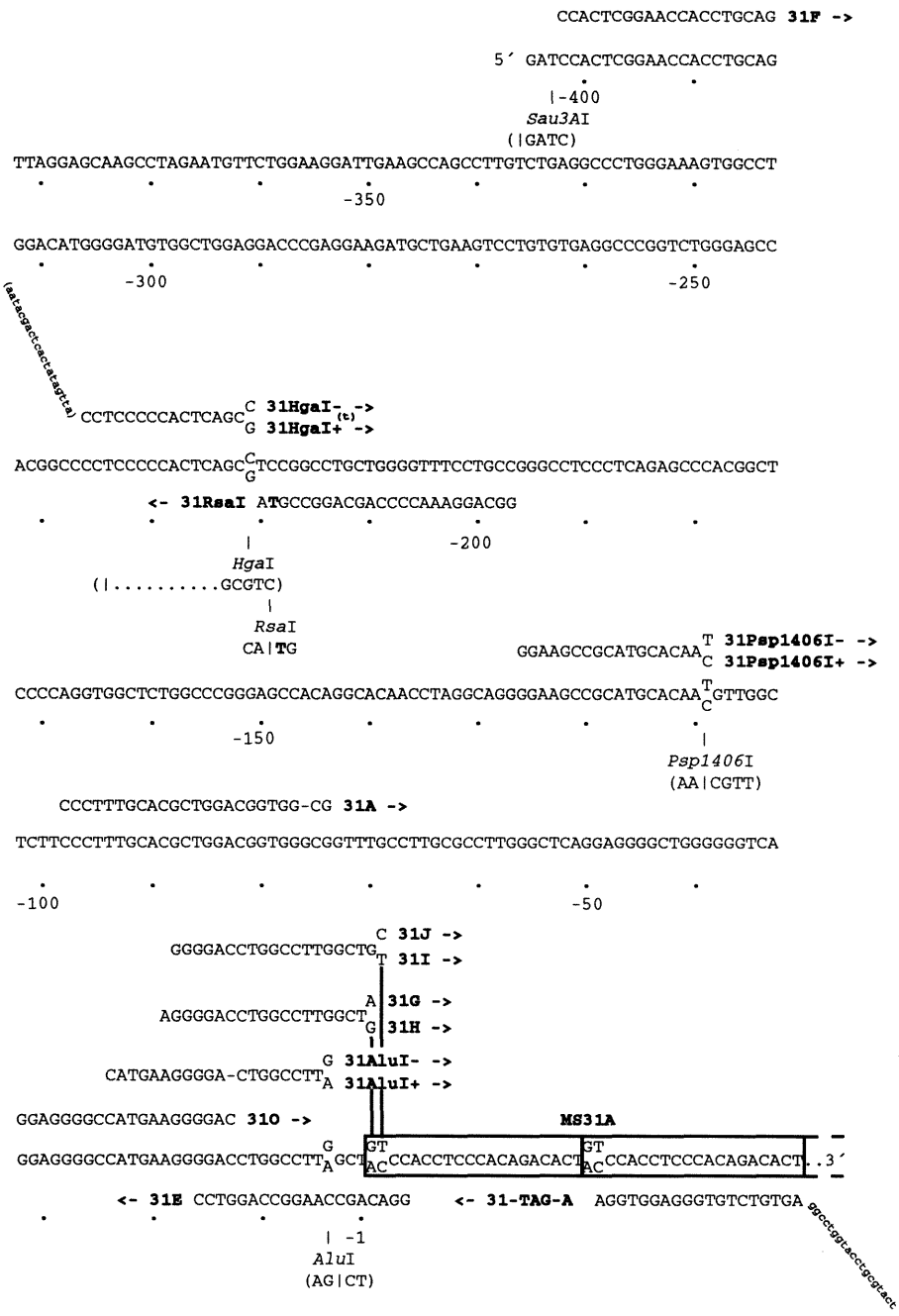
The sequence of cloned MS31A 5′ flanking DNA revealed a candidate cryptic *Alu*I site 403bp from the 5′ *Sau*3AI site and 2bp from the first repeat unit (Fig. 5.2). To determine whether a variant position here was responsible for the polymorphism, DNA from 3 individuals, characterised by restriction mapping as $AluI^{+/+}$ and $AluI^{-/-}$ homozygotes and an $AluI^{+/-}$ heterozygote, was analysed. Genomic DNA from these individuals was digested with *Alu*I and then used as the template in an MVR-PCR reaction using the flanking primer 31O which binds 10bp 5′ to the candidate cryptic *Alu*I site (Fig. 5.2). The $AluI^{-/-}$ homozygote gave a normal diploid code, the $AluI^{+/+}$ homozygote gave no MVR map and the $AluI^{+/-}$ heterozgote gave a largely single allele code from the $AluI^{-}$ allele. (Fig. 5.1) These results confirmed that the *Alu*I polymorphism is located between primer 31O and the minisatellite. The presence of diploid positions (arrowed) in the single allele code and bands in the $AluI^{+/+}$ homozygote lanes indicated that *Alu*I digestion was not complete and consequently that this method may not be suitable for large-scale single-allele coding.

Isolating and sequencing the MS31A 5′ flanking DNA. In order to design allele-specific primers for MS31A MVR-PCR it was necessary to identify the molecular basis of the polymorphic *Alu*I site. This was done by PCR amplification and sequencing of the region containing this site from individuals homozygous for $AluI^{+}$ and $AluI^{-}$ alleles. To screen for further variant positions, the ~400bp of MS31A 5′ flanking DNA for which PCR primers were available was isolated and sequenced from an additional 10 chromosomes derived from unrelated individuals. These included those carrying both $AluI^{+}$ and $AluI^{-}$ alleles, to ensure that variant positions in linkage disequilibrium with different alleles at the *Alu*I site would not be missed. Ideally I wanted to identify polymorphisms with a heterozygosity of at least 10%, that would provide reasonable numbers of informative individuals for use in large-scale allele-specific MVR-PCR. Comparison of the sequences from a total of 12 chromosomes gives a 72% [$100\ (1 - 0.9^{12})$] chance of detecting variants with a population frequency of 0.1 and predicted heterozygosity of 18% [$2\ (0.1 \times 0.9)$], assuming Hardy-Weinberg equilibrium.

To incorporate as much flanking DNA as possible, including the *Alu*I site, into a PCR product which could then be sequenced, primers 31F (Fig. 5.2) and 31-TAG-A and 31-TAG-G, (Fig. 4.1) were used. 30 cycle PCR amplifications between 31F and 31-TAG-A, and 31F and 31-TAG-G primers, all at 1μM in the absence of TAG, with annealing at 68°C and extension for 30 seconds, produced PCR products extending across the flanking DNA and into the first few

Figure 5.2. The MS31A 5′ flanking sequence, RFLPs and primers.

The sequence of cloned MS31 5′ flanking DNA was determined by Armour *et al.*, (1989). Additional Taq-cycle sequencing of double stranded DNA, generated from the total genomic DNA of unrelated individuals by PCR amplification between primers TAG and 31F, revealed three polymorphic positions of base substitution. The 5′ flanking sequence of the top strand of MS31A is shown 5′-3′ (top to bottom, left to right) including first two MS31A repeat units (boxed). Numbers below the sequence indicate the number of bases from the minisatellite; the first base of the first repeat unit being counted as 0. Polymorphic positions (bases above and below main sequence), restriction enzyme sites (vertical bar "|" below sequence), flanking primer and allele-specific primer sequences, are also shown. The restriction enzyme for which the polymorphic sites create an RFLP are shown below each, with the recognition sequence for the enzyme and a vertical bar (|) indicating the position of cleavage. The name of each flanking primer is shown (bold) together with an arrow at its 3′ end indicating the direction of priming. Pairs of allele-specific primers are identical except for the 3′ most base as shown. Vertical lines below the 3′ ends of primers 31G, 31H, 31I and 31J show the position of the first minisatellite repeat to which this base corresponds. Primers HgaI^{+t} and 31-TAG-A have 20nt extensions, (bracketed, smaller font) that are not complementary to the flanking sequence. The bold **T** in primer 31RsaI is a deliberate mismatch with the flanking sequence, used to engineer a polymorphic restriction site for the enzyme *RsaI* into PCR products generated from this primer. Gaps (-) in primers 31AluI and 31A indicate positions were there is an inadvertent mismatch between the primer and the published flanking sequence (Armour *et al.*, 1989).



Chapter 5 Figure 2

repeat units of MS31A. Following electrophoresis, the resulting MVR maps were visualised by ethidium bromide staining (data not shown). Single bands from heterozygous MVR positions were removed from the visible MVR map, in order to isolate flanking DNA from single chromosomes, so that any variant positions would be hemizygous and therefore more easily detected by sequencing. The DNA was recovered by electroelution and ethanol precipitation and dissolved in 20µl water (see Materials and Methods). 3µl of this template was then reamplified using the primers 31F and TAG and the same PCR parameters as previously, to provide large quantities (~1µg) of double stranded sequencing template, which was recovered in the same way.

The 5′ flanking DNA was sequenced from each end in a 10 cycle Taq sequencing reaction, using the same PCR parameters as in the original amplification with either primer 31F, or TAG, end-labelled with γ^{33}P or γ^{32}P (see Materials and Methods). Sequence comparison revealed three sites of base substitutional polymorphism in the MS31A 5′ flanking DNA. The first is an A/G transition, located 4bp 5′ to the first minisatellite repeat (-4A/G), which gives rise to the *Alu*I RFLP. The second is a C/T transition 109bp from the first repeat unit (-109C/T) which creates a *Psp1406*I RFLP and the third is a C/G transversion 221bp from the minisatellite (-221G/C) generating a polymorphic site for the enzyme *Hga*I (Fig. 5.2).

Determining the genotype of flanking polymorphisms. A simple assay was developed to determine the genotype of an individual at each polymorphic position. In all assays a region of flanking DNA spanning the site to be genotyped was first amplified directly from total genomic DNA by PCR and then digested with a diagnostic restriction enzyme. The resulting fragments were then resolved by electrophoresis through a high percentage agarose gel and visualised by ethidium bromide staining (see Materials and Methods & legend to Fig. 5.3).

*Alu*I RFLP. To genotype the -4A/G polymorphic site using *Alu*I it is necessary to generate a PCR product containing this site. The *Alu*I site is so close to the minisatellite that one of the primers must overlap most of the first MS31A repeat unit and it was possible that amplification using such a primer would generate a ladder of additional products that may have interfered with the assay, by priming at internal repeats. In order to minimise this potential problem, a PCR assay that was intended to bias amplification towards products from the first repeat unit was designed. I chose 31-TAG-A as a primer because this was known to recognise the most common repeat unit type (a-type), which was therefore more likely to be the first repeat unit. By using 31-TAG-A at higher concentration than in MVR-PCR (1µM vs 50nM) I hoped that it would also amplify non-specifically from the first repeat unit if this was t-type. To favour the amplification of short products from the first repeat unit the TAG primer was not used in this assay and a high cycle number (35) and short extension time (30sec) were employed.

When DNA extending from the flanking region into the minisatellite array was amplified from total genomic DNA using 31-TAG-A and flanking primer 31A (Fig. 5.2), only PCR products extending into the first few repeat units were amplified to levels detectable by staining with ethidium bromide. The 137bp fragment corresponding to amplification from the first repeat unit was always present and was the predominant PCR product. Cleavage of an AluI$^+$ (-4A) allelic PCR product with *Alu*I generates a 96bp DNA fragment extending from the 31A primer site to the *Alu*I site and a 41bp product too small to be detected in this assay. AluI$^-$ (-4G) alleles are not cut, and heterozygotes show both cut and intact PCR products. Examples of this assay are shown in Fig. 5.3A.

Figure 5.3. Assays for three MS31A 5′ flanking polymorphic sites.

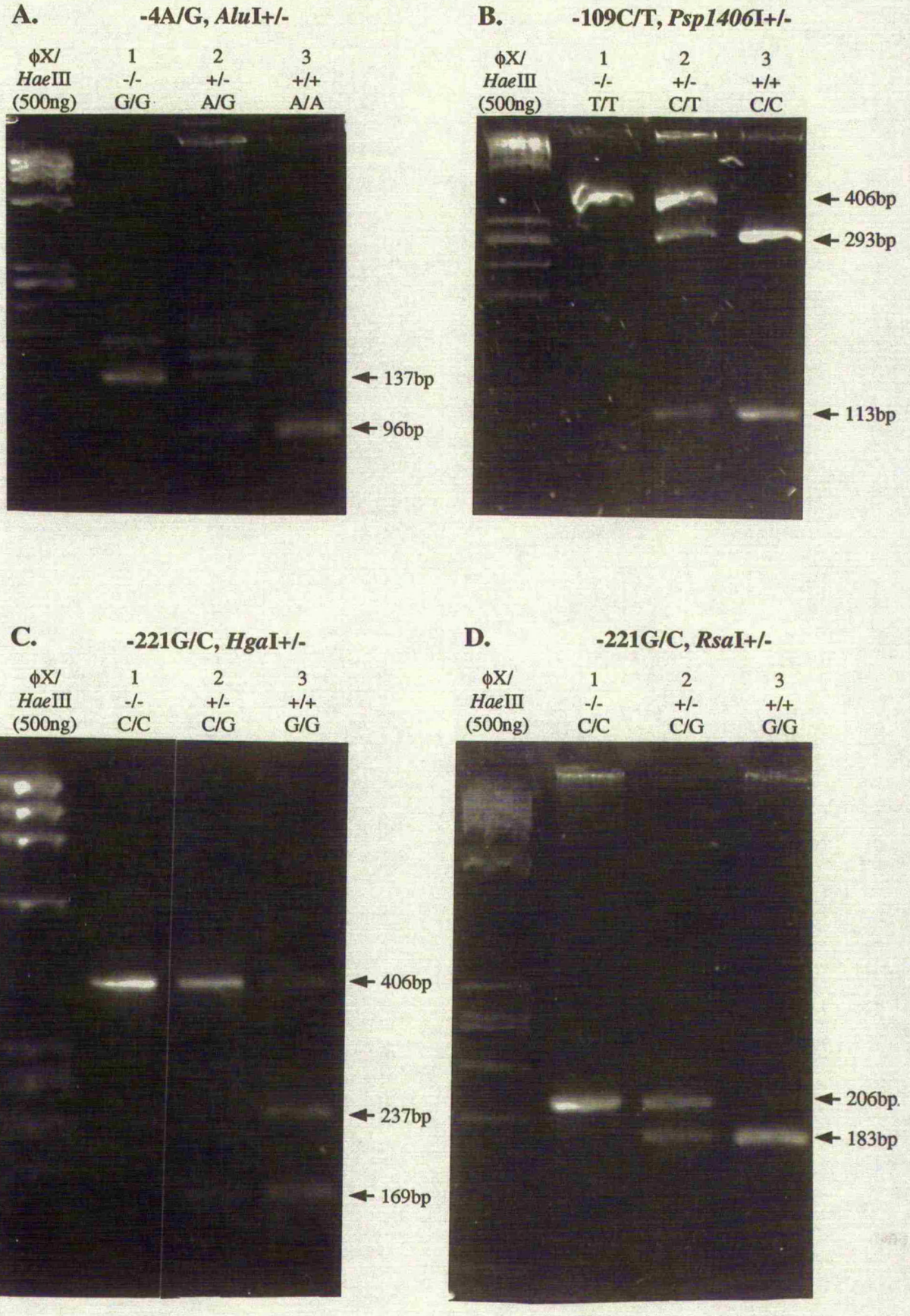
All assays use PCR amplification of a DNA fragment spanning the site to be genotyped, followed by digestion with a diagnostic restriction enzyme. PCR was performed in a Geneamp 9600™ thermal cycler (Perkin Elmer Cetus). 7µl PCR reactions contained; ~100ng genomic DNA, the standard PCR buffer, 1µM of the primers indicated and 0.25 units of Taq polymerase. Each cycle consisted of denaturation at 94° for 30sec followed by annealing for 30 sec at the appropriate temperature and extension at 70° for 30 sec. PCR products were digested by adding 3 units of the appropriate enzyme, along with 1µl of the manufacturers recommended 10x reaction buffer and 1µl 10mM spermidine trichloride. The *Psp*1406I digest also included 1µl 1mg/ml BSA. In all assays the digestion products were resolved by electrophoresis through a 5% NuSieve™ GTG agarose (FMC) gel in 0.5 x TBE buffer and visualised by ethidium bromide staining.

A. -4A/G, *Alu*I polymorphism assay. Amplification between 31-Tag-A and 31A for 35 cycles, with annealing at 70°, was followed by digestion with *Alu*I. The 137bp band corresponding to amplification from the first MS31A repeat unit is cut into two fragments of 96bp and 41bp (not visible) by *Alu*I if the -4A variant is present. The three individuals shown are therefore scored as; 1, -4G/G, 2, -4A/G and 3, -4A/A respectively.

B. -109C/T, *Psp*1406I polymorphism assay. 30 cycle amplifications using 31E and 31F, with annealing at 66°, were followed by digestion with *Psp*1406I. The 406bp product amplified between 31E and 31F is cut into two fragments of 293bp and 113bp if the -109C variant is present. The three individuals shown are therefore scored as; 1, -109T/T, 2, -109C/T and 3, -109C/C respectively.

C. -221G/C, *Hga*I polymorphism assay. 30 cycle amplifications using 31E and 31F, with, with annealing at 66°, were followed by digestion with *Hga*I. The 406bp product amplified between 31E and 31F is cut into two fragments of 237bp and 169bp if the -221G variant is present. The three individuals shown are therefore scored as: 1, -221C/C, 2, -221G/C and 3, -221G/G respectively. The -221G/G digest is incomplete showing that this enzyme was unsuited to this assay. *Hga*I is also very expensive, so a cheaper and more efficient assay for this polymorphism was developed (Fig. 5.3D).

D. -221G/C, *Rsa*I polymorphism assay. 30 cycle amplifications using 31F and 31RsaI, with annealing at 68°, were followed by digestion with *Rsa*I and resolution and detection of products. The 206bp product amplified between 31RsaI and 31F is cut into into two fragments of 183bp and 23bp (not visible) if the -221G variant is present. The three individuals shown are therefore scored as: 1, -221C/C, 2, -221G/C and 3, -221G/G respectively.



Chapter 5 Figure 3

Table 5.1. MS31A flanking polymorphism allele frequencies and heterozygosities in 80 Caucasians and 93 Japanese.

		Caucasian			Japanese		
Position	Allele	Number	Frequency	Heterozygosity	Number	Frequency	Heterozygosity
-4	A	19	0.12	0.21	45	0.24	0.37
	G	141	0.88		141	0.76	
-109	C	39	0.24	0.37	85	0.46	0.50
	T	121	0.76		101	0.54	
-221	G	93	0.58	0.49	132	0.71	0.41
	C	67	0.42		54	0.29	

Table 5.2. Observed and expected genotype frequencies for the -221G/C, -109C/T and -4A/G MS31A 5′ flanking polymorphisms in 80 Caucasians and 93 Japanese.

	Caucasian		Japanese	
Genotype	Observed (O)	Expected (E)	Observed (O)	Expected (E)
-4A/A	2	1.2	5	5.4
-4A/G	15	16.9	34	33.9
-4G/G	63	62.0	54	53.7
χ^2	0.76 (0.5>p>0.2)		0.03 (0.95>p>0.9)	
-109C/C	5	4.6	20	19.7
-109C/T	29	29.2	45	46.2
-109T/T	46	46.2	28	27.1
χ^2	0.04 (0.99>p>0.95)		0.07 (0.7>p>0.5)	
-220G/G	25	26.9	50	46.8
-220C/G	43	39.0	32	38.3
-220C/C	12	14.1	11	7.8
χ^2	0.86 (0.5>p>0.2)		2.57 (0.2>p>0.1)	

***Psp*1406I RFLP.** The assay for this site simply amplifies the flanking DNA between primers 31E and 31F (Fig. 5.2) and uses *Psp1406*I to genotype the polymorphic site. The 406bp product from a -109C allele is cut into two fragments of 293bp and 113bp by *Psp1406*I (Fig. 5.3B).

***Hga*I RFLP.** The -221G/C site can be assayed by amplifying the MS31A 5′ flanking DNA between primers 31E and 31F and then digesting the 406bp PCR product with *Hga*I. If the *Hga*I site (-221G) is present the product will be cut into two fragments of 169bp and 237bp (Fig. 5.3C). However, this enzyme is inefficient in this assay (as shown by incomplete digestion of the 406bp fragment in the -221G/G homozygote) and is also very expensive. I therefore designed an alternative assay that uses a PCR primer, 31RsaI (Fig. 5.2), with a base mismatch (penultimate 3′ base shown in bold, T) to force a point mutation into the DNA adjacent to the polymorphic site in the PCR product. This creates a restriction site for *Rsa*I (GTAC) in products amplified from the -221G allele. In these alleles the 206bp product amplified between 31RsaI and 31F is cut into fragments of 183bp and 23bp by *Rsa*I (Fig. 5.3D).

Population surveys of flanking polymorphic positions. These tests were used to define the genotypes of all three polymorphic positions in large numbers of unrelated individuals from Caucasian and Japanese populations, both to define the heterozygosities of these polymorphisms and to identify individuals suitable for allele-specific MVR-PCR. The three sites are polymorphic and at Hardy-Weinberg equilibrium in both populations surveyed (see Table 5.1 for allele frequencies, Table 5.2 for genotype frequencies; χ^2 null hypothesis of Hardy-Weinberg equilibrium for all sites ≤ 2.57, $p > 0.1$). The -4A/G polymorphism is the least variable, (21% heterozygosity in Caucasians, 37% in Japanese) while the -109C/T site (37% Caucasian, 50% Japanese) and the -221G/C site (49% Caucasian, 41% Japanese) have higher heterozygosities.

Determination of MS31A 5′ flanking haplotypes. There are several ways of defining the haplotypes of the 5′ flanking polymorphisms of particular MS31A alleles. The simplest of these is pedigree analysis, which was used to deduce flanking haplotypes for the MS31A alleles segregating in 40 of the large CEPH kindreds. This study confirmed Mendelian inheritance for all three polymorphic positions. For unrelated individuals this information could be obtained by allele-specific MVR-PCR, using primers for each heterozygous position, or sequencing the 5′ flanking DNA of mapped alleles. However, it was more efficient to assay flanking haplotypes directly by PCR amplification and restriction analysis and then use allele-specific MVR-PCR to link them to MS31A alleles.

Double heterozygotes can be haplotyped by amplification between an allele-specific primer at the most distal heterozygous flanking position, followed by digestion with the appropriate enzyme for the proximal heterozygous site. In each assay a negative control was included to ensure that no non-specific amplification from the allele-specific primer had occured. To assay -221/-4 double heterozygotes it is necessary to amplify between either 31HgaI$^+$ or 31HgaI$^-$ and 31-TAG-A. However, amplifications using these primer pairs yielded very small quantities of product. This problem was overcome by using an extended version of 31HgaI$^+$, 31HgaI^{+t} (this primer was designed for another purpose and has a tail complementary to the T7 17mer primer sequence) which amplifies more efficiently with 31-TAG-A. In -221/-4 double heterozygotes, amplification between 31HgaI^{+t} and 31-TAG-A generates a 294bp product from -221G alleles only. The 294bp product from -221G/4A alleles is cut into two fragments of 254bp and 41bp by *Alu*I, while that from -221G/-4G alleles is not cut (Fig. 5.4A). -109/-4 double heterozygotes were haplotyped by amplifying between Psp1406I$^+$ and 31-TAG-A followed by digestion with *Alu*I; -109C/-4A chromosomes give products of 124bp and 41bp, -109C/-4G chromosomes give a single product of 165bp (Fig. 5.4B). -221/-109 double heterozygotes were haplotyped by amplifying between HgaI^{+t} and 31E followed by

Figure 5.4. Haplotype assays for MS31A 5′ flanking polymorphic sites.

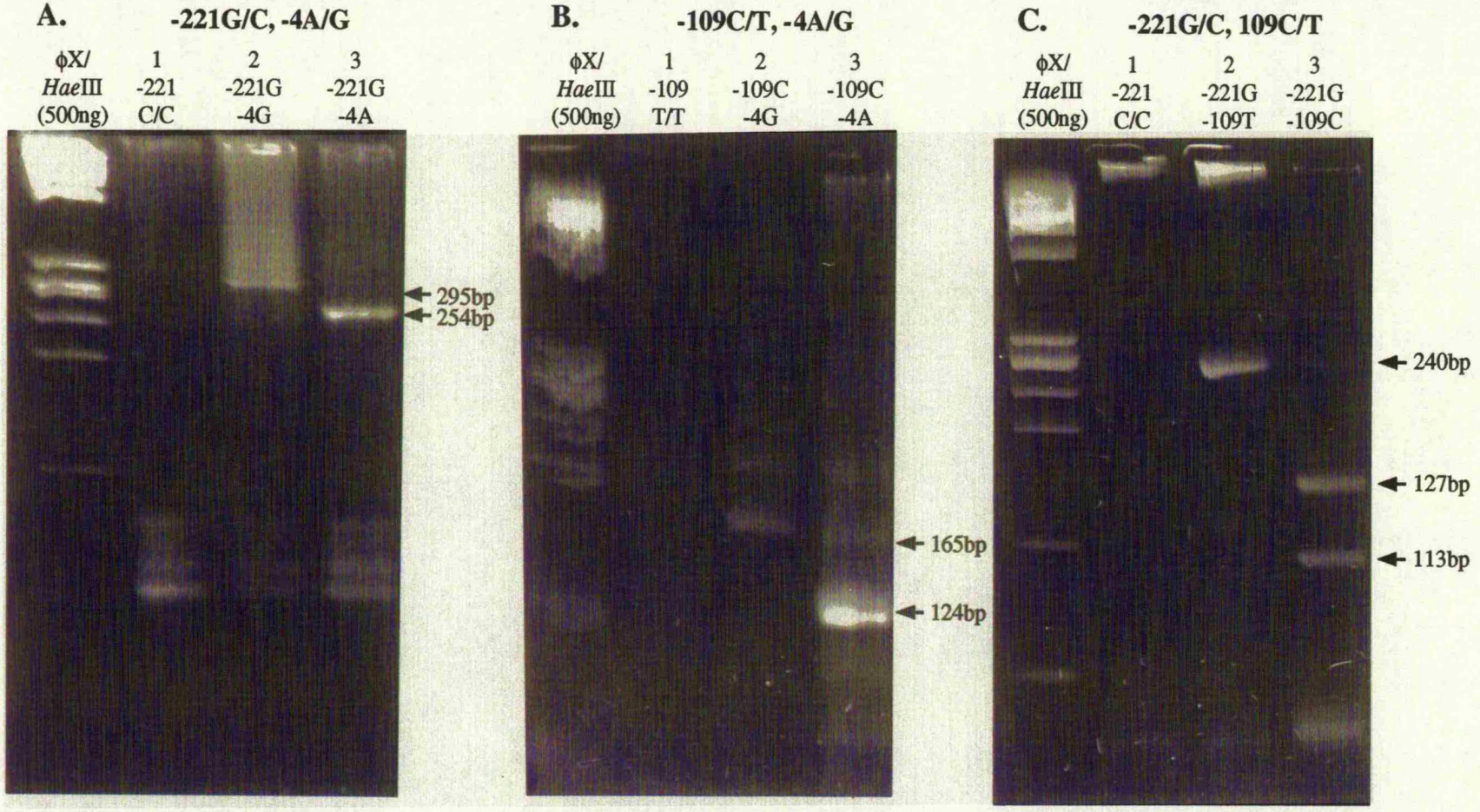
Flanking haplotypes of double heterozygotes where phase was unknown were determined by PCR amplification between an allele-specific primer directed to the 5′-most heterozygous position and a primer complementary to a sequence 3′ of the second polymorphic position, followed by digestion with the appropriate restriction enzyme. Two such assays were performed to establish phase in triple heterozygotes. PCR reactions, restriction digestion, resolution and detection of products were performed as described in the legend to Fig. 5.3. Control individuals, homozygous for the variant position discriminated against by the allele-specific primer were included in each assay, to check that the products scored were generated from one chromosome only.

A. -221G/C, -4A/G double heterozygote haplotype assay. Amplification between 31HgaI+t and 31-Tag-A for 35 cycles, with annealing at 70°, was followed by digestion with *Alu*I. The 295bp band corresponding to amplification from the first MS31A repeat unit is cut into two fragments of 254bp and 41bp (not visible) by *Alu*I if the -4A variant is in phase with -221G . The three individuals shown are therefore scored as; 1, -221C/C, 2, -221G/-4G and 3, -221G/-4A.

B. -109C/T, -4A/G double heterozygote haplotype assay. Amplification between Psp1406I+ and 31-TAG-A for 35 cycles, with annealing at 66°, was followed by digestion with *Alu*I. The 165bp band corresponding to amplification from the first MS31A repeat unit is cut into two fragments of 124bp and 41bp (not visible) by *Alu*I if the -4A variant is in phase with -109C . The three individuals shown are therefore be scored as; 1, -109T/T, 2, -109C/-4G and 3, -109C/-4A.

C. -221G/C, -109C/T double heterozygote haplotype assay. Amplification between 31HgaI+t and 31E for 30 cycles, with annealing at 66°, was followed by digestion with *Psp1406*I. The 240bp amplification product is cut into two fragments of 127bp and 113bp by *Psp1406*I if the -109C variant is in phase with -221G. The three individuals shown are therefore scored as; 1, -221C/C, 2, -221G/-109T and 3, -221G/-109C.

The haplotypes of triple heterozygotes were determined by performing the two assays detailed in **B.** and **C.** above.



Chapter 5 Figure 4

Table 5.3. MS31A 5′ flanking haplotype frequencies for the -221G/C, -109C/T and -4 A/G flanking polymorphisms of 158 Caucasian and 186 Japanese chromosomes.

	Caucasian			Japanese		
Haplotype	number		frequency	number		frequency
-221 -109 -4	expected	observed	observed	expected	observed	observed
G C A	2.6	10	0.06	14.6	32	0.17
G C G	19.4	24	0.15	46.2	40	0.22
G T A	8.4	8	0.06	17.1	6	0.03
G T G	61.3	50	0.31	54.2	54	0.29
C C A	1.9	0	0.00	6.0	2	0.01
C C G	14.0	5	0.03	18.9	11	0.06
C T A	6.1	0	0.00	7.0	5	0.03
C T G	44.4	61	0.39	22.1	36	0.19
TOTAL		158			186	
χ^2	44.3 (p<0.001)			44.1 (p<0.001)		
χ^2 -221/-109	18.2 (p<0.001)			14.4 (p<0.001)		
χ^2 -221/-4	14.0 (p<0.01)			5.3 (p<0.05)		
χ^2 -109/-4	10.5 (p<0.001)			21.4 (p<0.001)		

Table 5.4. Repeat unit composition of 99 Caucasian and 79 Japanese MS31A alleles.

	Caucasian		Japanese	
Repeat	Number	Frequency %	Number	Frequency %
a-type	3681	55.2	2745	50.0
t-type	2906	43.6	2681	48.8
O-type	79	1.2	65	1.2
Total	6033	100	5491	100

Most alleles (excluding those <70 repeats long) were mapped over the first 70 repeat units (mean for all alleles 70.0, S.D. 2.5).

digestion with *Psp1406*I; -221G/-109C chromosomes give products of 147bp and 113bp, while -221G/-109T chromosomes give a product of 260bp (Fig. 5.4C). To haplotype triple heterozygotes the phase of -221G and -109C/T and -109C and -4A were both determined separately.

These assays were applied to unrelated Japanese and Caucasian individuals who were heterozygous at two or three of the polymorphic positions and for whom flanking haplotypes could not be deduced from pedigrees. Observed haplotype frequencies for Caucasians and Japanese are shown (Table 5.3) and compared to expected haplotype frequencies, calculated from allele frequencies assuming random association between the sites. In both populations significant linkage disequilibrium exists between all the polymorphic sites, (χ^2 (7df) ~44, significant deviation from null hypothesis of random association). However, linkage disequilibrium only appears to be absolute for the haplotype -221G/-4G in Caucasians. All possible flanking haplotypes were observed in the Japanese population. These haplotype frequencies can be used to predict the proportion of individuals heterozygous for at least one flanking polymorphic site (1- proportion homozygous at all sites) and therefore suitable for single-allele mapping by allele-specific MVR-PCR. This combined heterozygosity over all three flanking polymorphic positions is 72% in Caucasians and 80% in Japanese.

2. Allele-specific MVR-PCR at MS31A.

Design of allele-specific primers. Three pairs of flanking primers differing only in their 3′ terminal base, A (31AluI$^+$) or G (31AluI$^-$), C (Psp1406I$^+$) or T (Psp1406I$^-$) and G (31HgaI$^+$) or C (31HgaI$^-$), which corresponds to the variable base at each flanking polymorphic position, were designed for allele-specific MVR-PCR (see Fig. 5.2 for sequences). Their length was tailored in an attempt to make all allele-specific primers discriminate at the same annealing temperature (68°C). The AluI$^+$ and AluI$^-$ primers have an additional mismatch with the MS31A flanking DNA, due to an error in their sequence specification; this did not compromise their use in allele-specific PCR and was therefore not rectified. Annealing temperatures were titrated for each primer to find those most suitable for allele-specific MVR-PCR. 31AluI$^+$, 31AluI$^-$, HgaI$^+$ and HgaI$^-$ work best at an annealing temperature of 68°C, while Psp1406I$^+$ and Psp1406I$^-$ discriminate optimally at 66°C (data not shown). When employed as the 5′ flanking primer in MVR-PCR, using the appropriate annealing temperature, these primers discriminate almost completely between the two alleles in -4A/G, -109C/T or -221G/C heterozygotes, allowing selective mapping of one or other MS31A allele from total genomic DNA of individuals heterozygous for any one of these polymorphisms (Fig. 5.5).

Diversity of allelic MVR structures at MS31A. Both size separated alleles and allele-specific MVR-PCR were used to map 182 MS31A alleles from unrelated individuals, of mainly Caucasian and Japanese origin, over the first 70 repeat units, or in their entirety if shorter (Fig. 5.6). Almost all of the alleles mapped had different structures, indicating extreme variability. 79 Japanese alleles were mapped and 77 of these were different, these figures give an estimated θ value of 1010 (see Chapter 4 for calculation of θ values), suggesting a heterozygosity of ~99.9%. If all alleles were equally rare, Poisson analysis (using software written by A.J. Jeffreys, in Microsoft Quickbasic™) indicates that >1000 different Japanese MS31A alleles must exist to give this sampling frequency distribution. All 99 Caucasian alleles mapped were different indicating that the number of different alleles, and hence heterozygosity, may be even higher in this population. Only one individual from the CEPH panel of families was scored as a homozygote by Southern blot length analysis; although genotyping revealed that this individual was heterozygous at the -221G/C site, heterozygosity in MVR structure is yet to be confirmed.

Figure 5.5. Allele-specific MVR-PCR at MS31A.

The optimal discriminatory annealing temperature for each of the allele-specific primers was determined by annealing temperature titration (data not shown). This figure shows the discrimination by each allele-specific primer at its optimal annealing temperature. Three individuals with the genotypes: 1, -221G/G (HgaI+/+), -109C/C (Psp1406I+/+), -4A/A (AluI+/+), 2, -221G/C (HgaI+/-), -109C/T (Psp1406I+/-), -4A/G (AluI+/-) and 3, -221C/C (HgaI-/-), -109T/T (Psp1406I-/-), -4G/G (AluI-/-), were chosen for analysis. MVR-PCR was performed as described in Materials and Methods. DNA samples were amplified for 20 cycles using the allele-specific primers specific for cut, (+, **A**) and uncut (-, **B**) forms of the RFLPs, plus TAG together with either, 31-TAG-A (A), or 31-TAG-G (T) (see Table 4.1 and Fig. 5.1 for primer sequences). PCR products were separated by agarose gel electrophoresis and detected by Southern blot hybridisation. In each case the allele-specific primer generates a diploid code from the homozygote with a complementary flanking sequence, a single-allele code from the individual heterozygous for the flanking polymorphism to which the primer is directed, and no products from the individual homozygous for a 3′ primer mismatch in the flanking DNA.

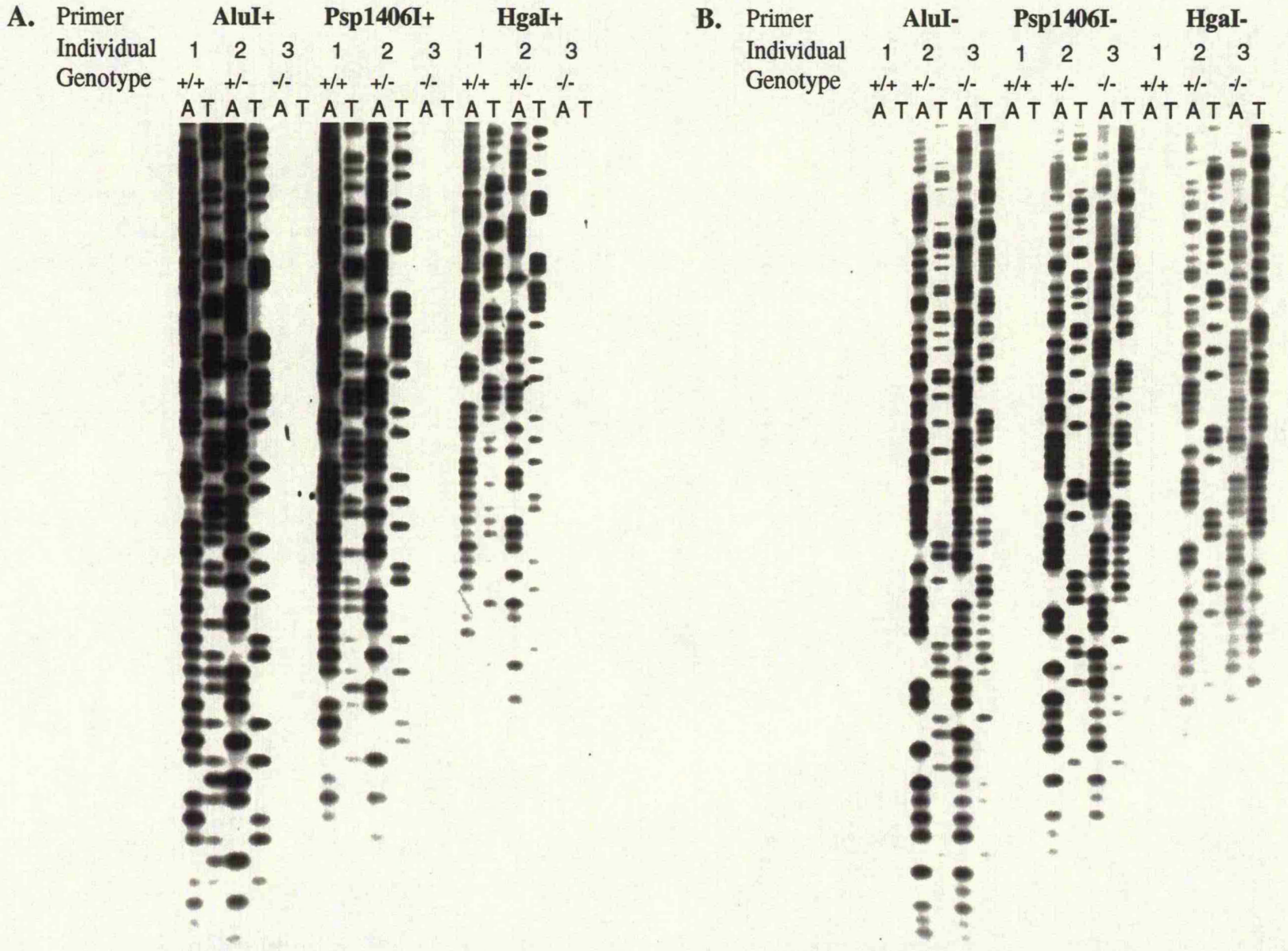

A.
Primer
Individual
Genotype
AluI+
Psp1406I+
HgaI+
1 2 3 1 2 3 1 2 3
+/+ +/- -/- +/+ +/- -/- +/+ +/- -/-
A T A T A T A T A T A T A T A T A T
B.
Primer
Individual
Genotype
AluI-
Psp1406I-
HgaI-
1 2 3 1 2 3 1 2 3
+/+ +/- -/- +/+ +/- -/- +/+ +/- -/-
A T A T A T A T A T A T A T A T A T

Figure 5.6. MS31A allele database.

A. Groups of MS31A alleles aligned by dot matrix analysis. To identify related alleles which share extensive regions of map similarity, all possible pairwise comparisons of 109 allele codes were made by dot matrix analysis and the diagonals searched for perfect eight repeat matches. Pairs of alleles showing at least 20 matching positions over the best two diagonals were selected and checked by eye for relatedness and alignment. For each allele its ethnic origin; Caucasian (c), Japanese (j), Malaysian (ma) or Bangladeshi (b)], tandem array length (to the nearest 5 repeats, as estimated from Southern blot length analysis, nd = not done), flanking haplotype and MVR haplotype are shown. The predominant group specific portion of each haplotype is shown in red. Regions of sub-group homology are shown in blue, and positions of divergence are shown in black. Gaps (-) have been introduced to improve alignments. a = a-type repeat; t = t-type; 0 = null or O-type repeat; ? = ambiguous (position not scored); ..... = allele continues beyond mapped region; << = end of a short allele. Coloured arrows above the first allele in group 2 show three examples of regions of intraallelic internal repetition.

B. Unalignable alleles. For each allele its ethnic origin, flanking haplotype and MVR haplotype are shown, as above. These alleles have no regions of homology detectable by dot matrix alignment. The most homogeneous MS31A allele, which is also the shortest Caucasian allele mapped to date, is highlighted in red.

C. Short Oriental alleles. These alleles are presented as in Fig 5.6A and alignments between them were detected in the same manner. Vertical bars (|) indicate alleles which are indistinguishable. Although all the alleles in this group are highly related they were subdivided into two groups, 19A and 19B, on the basis of differences in structure at the 3′ end (shown in blue in group 19B).

A. Grouped Alleles

Group	Race	Estimated Allele Size (repeats)	Flanking Haplotype -220	-109	-4	MVR-haplotype
Group 1						
	c	nd	C	T	G	?aaaaattaaataaaaaattatataatatttaaaatttttatataatataaatttaaattaatttattaat
	c	235	C	T	G	tattatattatttattatataaattaattttatataatataaatttaaattaatttattaattaat
	b	240	C	T	G	?aaattatataatataaatttaaattattaaataataataattt-----attaatt-aattatattaattaatat
	c	325	G	C	G	???taattattaatattatttt-atttaaaatattaaatttat----attaatt-aa-tatattaattaatatat
	c	nd	C	T	G	?taataattattaatattatttttaatttaaaatattaa-ttatt---attaatttaa-tatattaattaatata
	c	265	G	C	A	aataaaa0ataaaatattaatttaaaatattaaaaataataataatatt-aa-tatattaattaatata
	j	nd			G	????????a?aataatattatattattatataatta-tt-aa-tatattaattaaaatattataa-aat??atataaa
	c	220	G	T	G	?ataatatttaaaaataatatatatataaataa-ttatt---attaatt-aa-tatattaattaatatattat
	j	nd			G	????ataa?atttatatttatatttaattaataaatataataaattatt---attaatt-aa-tatattaattaatatattat?ataatatatat
	j	nd			G	???????atatttatatatatataatttttataaatataatatatttataattatt---attaatt-aa-tatattaattaatatattat?ataatatatata<<
	c	195	G	C	G	taaattaaaatataaataataaaataatatatatta-tt-aa-tatattaattaatatattataataatatat
	j	nd			G	?tttttaaattaattaattatatttaattaataaatataataaattatttttatttaa-tatattaattaatatattataataatatatataa-tat
	j	nd			G	??tttaatttaaaattatttaatatttaa-tatattaattaatatattataata-tatatataaatattaatttttataat.
	c	250	G	C	G	?aatat0attattttaaatatttaaatattaattaatatattataataataaaaaatataatttatt
	c	210	C	C	G	?tattaataataataaatataattaataaattaataaatttattaattaatatattat--taatataatt
Group 2						--------> --------> ----------->------>------> ---------->
	c	140	G	C	A	?ttaaaat0taaaaaatattatttttttatattttattaaatattaaaat0tatta-ttttttatatttata-000aaa00aa00-atttt
	c	105	G	C	A	?ttatattataattatatattattaaaat0tattaatttttttatattt-atat000aaa00aa00-atttt
	c	100	G	C	A	taaatatatattattaaaat0tattaattt--ttatattttata-000aaa00aa00-attttaatttttattt
	c	95	G	T	A	ttaatatattattaaaat0tattaatttttttatattttata-000aaa00aa00-attttaatttttattt
	c	90	G	C	A	tatattattaaaat0tattaatttt-ttatattt-ata-000aaa00aa000attttaatttt-attttt
	c	105	C	C	A	??tatattaaaat0tattaatttttttatattttata-000aaa00a-000attttaattttt-tt
Group 3						
	j	nd			A	atttaaaaaaaaaaaattttaatatttttttttaaattatatttttaatttatttattttattttt-attatttttaatattaaaaaatt
	j	nd			A	ataaataaatttatatatttatatttta--ttatttatttatttttata--ttttaatattaaaaaattt-aaattataaaatattttaatataatat
	j	nd			A	?tataatatattattttatatttat-ttatttatttttttttttattattttaatattaaaaaattt-aaattataaaatattttaatataatat
	j	nd			A	?tttaatttttattttttttt-aatatttt-aatattaaaaaattttaaat-ataatataattttaatataatattatt
Group 4						
	c	260	G	T	G	?ttataattaattatatttaaataaataaattattaataataaattaa
	c	430	C	T	G	aaaataaaatattaaataaatttaatttaaaaatttataaattaaattattaataataaattaattaa
	c	235	C	T	G	?taatttaaataaatttaatttaaaaataatatatataaaataaa-taaaataataaataaaaaaaaaaa-taaa-taattattattaa
	c	255	C	T	G	?aaaaattaaataaaatataatatataaaatataaaataaaata-taaataaaaaaaaaaaaataaaataattattattaattaatatatt
Group 5						
	c	300	G	T	G	ttaaattataattaaaaaataaaaaatataaaaattaaaaaatataaaaattaaataaaaaatataatat
	c	280	C	T	G	?aaataataaaaaatataaaaattaaaaaatataaaaattaaataaaaaatataatataataat
	c	250	C	T	G	aatataaaaattaaataaaaaatataatataataataaaatataatataatataatataaaaataaatta
Group 6						
	c	220	G	C	A	?aaaaatatataataataaaaaaattattttattattaatattatttatataaattaatttaaaatatta
	c	220	G	C	G	?tt?aata0taattaaaaaattttattattaatattatttaaatttaaaatatt
	c	240	G	C	G	aaaaaaaaattttattattaatatt-tttaattttaattaatattattttatttaaaattttaattaa

Continued over

Figure 5.6 MS31A allele database.

A. Grouped Alleles

Group	Race	Estimated Allele Size (repeats)	Flanking Haplotype -220	-109	-4	MVR-haplotype
Group 7						
	c	170	G	C	A	??attaataatttataattataattaattttaattt--aaaaatttaaaaataa0atatttaaataatatt
	c	240	C	T	G	aaatatatatataaatataattaattttaatttaaaaaaatttaaaaataaaatattttaaattaaat
	c	155	G	C	A	??aaatatataaatttaattatttttatt-aaaaaatttaaaaataaaatatttaaataatattaaatt
Group 8						
	c	340	C	T	G	taaataataaataaataataaaattt--attaaatattattaaatatatatatatatatattaaaataatatat
	c	430	C	T	G	??ataaaataaaataaaattttaattaaatattattaaatatatatatttatat-ttaaa-taatatatat
	c	320	C	T	G	tatataaaatttta-ttaaatattattaaatttatatatatatttattaaaat
Group 9						
	j	nd	G	C	G	?aaataaataaataaataaataaataaaattataaaataaaatataaaatta--taaaataaaata-taaaa
	j	nd			G	??????taaataaattat--aaattataaaataaaata--aaattaaataaaataaaaat0tattattattaatatattatttatatattatataatatatat
	j	nd			G	???????ata?aataaaataaaattataaaataaaatataaaatta--taaaataaaataataaaaatataaaatataaaatttaaattaattat
Group 10						
	c	185	G	C	G	taatataataataaaaattaaataaaataaattttataaataaaaataaaaataaaatattaaaaatataaaatataaatttaa--ttatattataataatatt
	c	180	G	C	G	t??tataataata----------aaataaaaataa-aataaaatattaaaaatataaaatataaatttaaaattata0tataataatat
Group 11						
	c	205	G	T	A	?aaaaattaaaataaaaaaaaaataaatatataaattaaattaaataaatattattattaattaatatattaattatatttattaat
	c	195	G	T	A	taataaaaaataaaatataaattaaattaaataaatattattattaattaatatattaattatatt-attaat
Group 12						
	j	nd	G	C	A	?aaatttttttattaaaattaataaaataaaatattaaaaatataaaattataaaattaatttattataa
	j	nd	G	C	A	atatttttttattaaaattaataaaattaataaattattaataataattatatatattaaataaattta
Group 13						
	c	100	G	C	G	?aaaatttaattatatataaaaaaaaaatttaattatataaaaatataaattaaaaatttaaaaatataa00attattattt
	c	255	G	T	G	aaaaattaaattatataaaaatataaattaaaaatttaaaaatataa00attattatttatatttttta
Group 14						
	j	nd			A	attaattaaataaaaaaaaaaaaaaaaaaataaattttttaattttatttat-taaaatttaaataaaatatttttttttaatatttt-aattttaa
	j	nd			A	??taattaaat-aaaaaaaaaaaaaaaaataaattttttaattttatttatataaaatttaaataaaatatttttttttaatatttt
Group 15						
	j	nd			A	??ttattattaataaataaaaattaaaattattattaataaataaataatataaattaatataaatataataaattaaaaatttaatatatataaaaataatataataaaatttaaaatttaa
	j	nd			A	?ataatttattttaataaataaataatataaattaatataaatataataaattaaaaatttaatatatataaaaataatataataaaatttaaaatttaa
Group 16						
	j	nd			G	??????atttataatataataataatatattttataaaataataaaatttataaaaaaattttatataataaaaaaata--taaattttaatt
	j	nd			G	?atataataaaatttataaaaaa-ttttatataataaaaaaaaatattaaatttaattaaatattatttaaaatatttattattaaattataataa
Group 17						
	j	nd	C		A	?tataaaaaaataataataaaaattttataaaa-taataaaaaaaataaataaaaataatatttaaat
	j	nd			A	?tattaaatattataaaaatttt-ataaaaataataaaaaaaataaataaaaataatatttaaataaaaataatatttaaa
Group 18						
	c	140	G	T	G	tattattaaaataaattaaatttaatttaattttaaatttaaaatttaattaat
	c	120	G	T	G	t?aaaattaaatttttaattttaa--ttaaaatttaattaataatttattttaatttttttaatatt

Continued over

Figure 5.6 MS31A allele database.

B. Unalignable Alleles

Group	Race	Estimated Allele Size (repeats)	Flanking Haplotype -220	-109	-4	MVR-haplotype
	c		G	T	G	atatttttaaaatatttaaaaaattataatattttaattataatattataatattataaataaatatttta.....
	c		G	T	A	?ataaaaaaaaatttaaaaatttaaatataataataatataaataaaaaaaaataaattatattaattta.....
	c		C	T	G	tattaaatttaatatttaaaaatattaaatataatttttataatatattaataaaatattataaaatta.....
	c		C	T	G	ttaaatataataatttaaaaaattaaaaataaaattaaataaataaataatttaatttttaataattatt.....
	c		C	T	G	ttatattaaaataaaaaaaaattaaatttaattaaataatttatttaaattaaataaattataaattaaa.....
	c		C	T	G	taaataataaatttaaattaaataaaaaattattatataataatatatataataatatatataatat..... .....
	c		C	C	G	tatatataatatatatatataaaatatatatatatatataatataaatataaaaaaataaaaaataaat.....
	c		C	T	G	tattaataataaaaattaaatataaatatatatttaaata0aaataatattaa0aaata0aaaatat.....
	c		G	T	A	?aaatttaaaatatattaataatattttattattttaataaaattaaaattatatattaaaaataaattaa.....
	c		C	T	G	aataataaaaaatttaaaattaaaattaaattaatttatttaaaatttatttatttaaaatttattta.....
	c		G	T	G	ttaataaaaaattaataataaaaaattaaatttattatttaaataataaatttttttattaataata.....
	c		C	T	G	aataaatatttaattaaaataaaataaatatatatatttaaaaatatatatattaaaatatatatatat.....
	c		C	T	G	taatataaaaatattaatttaaaataatatttatattaaatttaattaatttataaaatattaatttaa.....
	c		G	C	A	taaataaataataataaaaaaaaaataaaatttaatttttttttttaaaaataaatttattattaatatta.....
	c		C	T	G	aaaaaaaaatataaaaataaaaaaattaaataaaattaattaaataaaaaaaaaattaaaaaaaaaaa.....
	c		G	T	A	aaattaaataaattaaatatttatttttaaaaaaaaaaaaaaaaaaaaaaaataaaaaaattaaataaaaaa.....
	c		G	C	A	aataaattttattttttaattattaatatttaOttataa000aataaa00aa000atttattatttatat.....
	c		G	C	G	tttatttttttttaataattaataatttaaaaaaaatttaaaatttaaatttttaaaaaaatataaaatttta.....
	c		G	T	G	aaaattttaattaaattaatttttttaaattaatttattaaaattaaaaattttttaatttaaatttatt.....
	c		G	C	G	tataatatatataatataatatatataaaaaaataaataaattaaaaataaaatatatataaaaaaaaaa.....
	c		G	T	G	aaataataaaaatataatttaattaaaatataaaaaaaaataaatttatttttttttttttaattaattt.....
	c		C	T	G	?at?atataaatttaattatttttattaatttaattatttaattaaattaataat0tataaaa0aaaaaa.....
	c		G	T	G	aataatatttttttttttttttttttttttttttttttttttttttttttattaaaa<<
	c		G	T	G	t???taatataaaataaaatttaattaatataaatatta0aaatattattttaaatttttaatataatt.....

Continued over

Figure 5.6 MS31A allele database.

B. Unalignable Alleles

Group	Race	Estimated Allele Size (repeats)	Flanking Haplotype -220	-109	-4	MVR-haplotype
	c		G	C	A	??ttaataatttatta0aa0aaatatatttaaaaaaaaaaatttaaataatttattaaataaaaattaaaaaa.....
	c		C	T	G	aaattttatatttaataatttatttttattatttttaataaatattaatttaataaatattaaataataatt.....
	c		G	C	A	?tataatttaatttaaaaataaaaaaaaatattaatttaattaaatattatttttaataataaaaaaaaaaa.....
	c		G	T	G	?ttttttaaaaaaatttttaaaaaatttataaatttttttttttttttatttatttatattttttttaatttatttata.....
	c		G	T	A	?aaaaaattatatatataataaaaatataaaaatataaaaatttaatttaaaattaaattaatatattatt.....
	c		C	T	G	taaatatattaatataaattaatatataaatataaatatttttatttaaaaaatttttaaataaaataatt.....
	c		G	T	G	?ttaataaaatataaatataa0atataattatataaattaaataaaatataaaattat0ttaaattaaa.....
	c		G	C	A	?aaatatataattaaaaatttaaaaaatatataaatataaattatttttatttatatttttttatttttata.....
	c		G	C	A	taaaaaaaaaaatatataaaaaaaaaaatataaaattaaaatttttttaatttaaaaaaaaaatattaaaatttt.....
	c		C	T	G	?aaattaaataaattaattataaattaaaataaataaattattattaattaattaattatttaatttttt.....
	c		C	C	G	??atttattaattaatatataaataatttatattaaaatttataatatatatattaaataatatattatt.....
	c		G	C	G	??atatatatataattaaataatataaaaaaaaaataaaatattattataaattaattaaatataaaatt.....
	c		C	T	A	?ttataaatattaaaaaataaaaattatttaaaatttaaaattattaataaattaatttatatataata.....
	c		G	T	G	??taattaatatttaatattaataatatttatttttattttattaaatttaaataaattaataaatttt.....
	c		G	C	G	??tataaaaataataatttaaattaatttttttataatatatttataatatatattatttaaaattta.....
	c		G	T	A	aaaaaaaaaataaaaaattaaatttaaattaaattaaaattaattatatatttattattttaaataataat.....
	c		G	T	A	??aattaattataaaaaatttaatatatattttattttttttaaataataaattttttaataaatatttt.....
	c		G	T	G	?tattaaatattaattttaattaatttat0aaaaataataataataataataaaaatat0atataataataaataat.....
	c		G	C	A	?taattaatataaaataattaatataa0tataaaaa0tatataata0tataaaaaa0tatataaaatattatt.....
	c		C	T	G	??taaatataaataaataaatatattttaaattaaattaatatatatattaaaataatatattattaa.....
	c		G	T	G	??taaaataaaataaaataataatatttattttaaataataataaaatttaaatataaataataatta.....
	c		G	T	G	taaaattaaaataaaaaaataaaaaaataattaaaaaaatataaaataaaaatataaaataaaaaaatat.....
	c		C	T	G	?aaaaaaaaaaaaaaaaaaattaattatttttattttataataattaattaatttaaattaaaataattta.....
	b		G	T	G	?aaataaaatttaaattaaaatatataaattaaataaaataaaataaaatataaataa?attattattt.....

Continued over

Figure 5.6 MS31A allele database.

B. Unalignable Alleles

Group	Race	Estimated Allele Size (repeats)	Flanking Haplotype -220	-109	-4	MVR-haplotype
	b		G	C	A	?aaaaaaattaaaataaaattaaaaaaaaaaaaaaaatataaaatttaaaaaaaaataaaaaaatataa.....
	j		G	C	A	attaaatttttaatttttttaatttattaaatttaaatttaatttttatttaaatttttaaaattttaatt.....
	j		G	C	A	?ttataaaatataaatatataatattattatttaattaatatttttttattaaaattaaattaatattaa.....
	j		G		G	?ttttattttttattttaaataataaaaaaatattatttttaataaataatatttaaaaatttatt.....
	j		C	C	G	taataataaaatataatattttataaatataat0tatatatat0tataa0tatattaaaat0at.....
	j		G	C	A	???taataattaaaaaatataaaataaataattaaaaattatataataataattaaaaatttaa?ata.....
	c		G	T	A	??aattaaaaaaaataaattatattatttttttttaattatttttttattttttttttttttttttta.....
	c		G	T	A	?aaaaaataatataaaatttaaaaatataaaaaaaatattaatttaaatataataatattataaataaa.....
	c		C	T	G	?atatttaaaaatataataaaataaa??a?aaaaaa?tatataaaataaaaaaaaaaaaaaaaaaaaatt.....
	c		C	T	G	taattaattattaaaattaataatttaaataattaataatttattaataataaattaatttatttaatt.....
	c		G	C	A	?attatataatataatataataaatataataaattaaaaatatttaaatttattattaaaataaattat.....
	c		G	C	G	?ttaaaaaaaatattatttatttaaaatatattaatttaaatttaatttttttattttattttaatatt.....
	c		G	C	A	?aaaaaattaaaaatatataaatataattatttttttattatttttaatttaaaaatttaaaataaaa.....
	j		G	C	A	tataaaaatttaaaatatataaaattaaaaataaaatatttttatttattaatatttttattata.....
	j		G	T	G	atatttatttttaaaaaaataaatattaaattattataataattattatattattttaatttatat.....
	ma		G	C	G	atttaaatataatatatataaataaataaaattatataatatatataatatataataaaatatat.....
	ma		C	T	G	attaaaaattaaaataaaattatttatttttaattttaatatttaattattaaaattatttt.....
	j				A	?aattttattataaaataaaaaaaaaattaaaaataaatttattaaatttttattttttataaaattataaa.....
	j				A	?aaaattaaataaaaaattaaataaatataaaaaattaaataaatataatataaatttaaaaattattt.....
	j				A	?ttaaatttaaattaaattaattttattaaattaataaaatatttaataattattattaattaataatattaatt.....
	j				A	?taaataaatttaataaatatttaaataaatatttaaatttttataattttaaaatttattaatatataatat.....
	j				A	?aattaaattaattttatttatatttaaaaaaataatttaaaaaatttaaaaataatatttaaaaaatataa.....
	j				A	?tttattttttttttttaaattaattaataattttttttaattattaattattattattaatattt.....
	j				A	?tattttaaattaaattaatattatttaaaatttttttttttttaatttttttaattatttatttattt.....

Continued over

Figure 5.6 MS31A allele database.

B. Unalignable Alleles

Group	Race	Estimated Allele Size (repeats)	Flanking Haplotype -220	-109	-4	MVR-haplotype
	j				A	attaaattaaaaaaaaaaattattatttttttttaaattttattattatataataat0ataattataattttattttaataaattaatatttttat.....
	j				A	atttaattatatttaaaatataatatttaattattaaattaatatttatataaaatttttaatataatttattaaatttttatttatat.....
	j				A	taaaaaaaaaaaattaaaaatataatttatttatataaattataaataaaatatttatttatttattaatatttt0ttataa0000aataaa00aa.....
	j				A	?aaataaaaatataaaaaaaaatatttaaaaatataaaaattaaattaatttatttaatattaatatttaaaatttattatttatttataatattattatttt.....
	j				A	ttaaaattaaattttttattttttttttattttttt0ttttttatttaaaattaaatttttttattttttttattatttttttttttttttaatt.....
	j				A	ttttaaatttaaaatttaaaaaaaatttaaatttaaaaaataaaaattatatttaaaaaattaaaattataaattaaaattataattaaaatt.....
	j				A	?tataaaatattaaaaaaatattttaattattaaaatataaataaaataatttattaataaattaattttaattaatatattataataatatatatatttttattaa.
	j				G	???????atttattttttttataattaaaaatatattataaattaaaattattatttttaataaatttta.....
	j		G			?????taataaaaattttttttattaaatatatttttttataaattatttataaataataataattattatttattaatatttaat.....
	j				G	??atataaaaaaa0ttttaaa0tat0aataaaaaaaaataaaaaaatataaataaaaaaataataataaataaaaaataaataataaataaaaa.....
	j				G	??aaaaaaaaaaataataaaaaaaa0aaaaaaaaaaaaaaaaaattataaaaaaaaaaaaaaaaaaaattttataataat?tttt.....
	j				G	?tataaataattataaatattaaaaaatattatttaattatttaataataattttatttattataaataaattt?aataatt.....
	j				G	???tataaatataaaaataaaattattaaaaaataaaatattaaaaatttaaaatttaaaattaataattaaaatttaatat.....
	j				G	??tattaatttttttaaattttaaaattaaattaaataatatttattaatatatttttaaataatttttaaataataaatatttttt.....
	c		G	T	G	?tataatattaaaaaaaatatataaaattaaaaaaaaaaaaaaaaaaaaaaaaaatttaaaaaattaattaaaaa.....
	c		C	T	G	?aa?t?aatatataataaaaattaattaatatatttaataaatatattatattaatatatataataa.....
	c		C	T	G	???taattattaattaataaaattaataaaatatttattaataatttttattatttattatttaatttt.....
	c		G	C	G	??tt?taataatttaataatttttatttatttaataaaatttttaaaaaaaaaaaaaattaaaataatttt.....
	c			C	A	?????t??ttatttttatta0atattaataaataaataaatttatttatttattaaatttaaatttaat.....
	c			T	G	???aaaataaaaaataaaaaaaaaatatttatttaatataat0tatttttttaaattaataat0tataaa.....

Continued over.

Figure 5.6 MS31A allele database.

C. Short Oriental Alleles

Group	Race	Estimated Allele Size (repeats)	Flanking Haplotype -220	-109	-4	MVR-haplotype
19A						
	j	53	G			?????att-tattatttatattaatatttatt-attaattttatatttttta0a0a<<
	j	53	G	T	G	?tttaattatattatttatattaatatttatttatta-ttttat-tttttta0a0a<<
	b	51	G	T	G	??tatttaattatattatttatattat--tttatttat-aatttt----ttttta0a0a<<
	j	51			G	?????ttta-ttatttttat------atttatttatttattaattttata-ttttta0a0a<<
	j	51			G	??ttattttat-atttatattaatatttatt-attaattttata-ttttta0a0a<<
	j	52	|		G	??ttaa--ttattatttatattaatatttatttattaattttata-ttttta0a0a<<
	j	52	|	G	T	G	????aa--ttattatttatattaatatttatttattaattttata-ttttta0a0a<<
	j	51		G			?????ttattatttatattaatatttatttattaattttata-ttttta0a0a<<
	j	49		G			???taa-tttattatt-atattaatattaatttt--aattt-ata-ttttta0a0a<<
	j	50		G			?????atttt-atttttatatttatattta-ttatt-attttata--tttta0a0a<<
	j	52		G	T	G	????attttattatttatattaatatttatttattaattttata--tttta0a0a<<
19B						
	j	50		G	T	G	?ttaattttat--tttattttaatatttatttatta--ttt-atata--tata0a0a<<
	j	52				G	???taattatattatttatattaatatttatttatta--ttt-at---tttata0a0a<<
	j	50		G			??????tatattatt-atattaatatt-atttatta--ttt-att--tttata0a0a<<
	ma	54		G	T	G	?attttttttttttttttttatattatt---------atttat--tttttattt-atat--ttata0a0a<<
	j	49				G	??????tatattatttatattaatatttt-------ttattt-atatatttata0a0a<<
	j	48				G	???????tatattatttatatta----------ttttttattttatttatttata0a0a<<
	j	45	|			G	???c?attatattatttatatta-----------ttt-ttttatatatttata0a0a<<
	j	45	|			G	???taattatattatttatatta-----------ttt-ttttatatatttata0a0a<<
	j	46		G	T	G	??ttaattatattatttatatta----------ttttttttttatat--ttata0a0a<<
	j	46				G	?????tattatattatttatattatttatttttttttt-------------ta0a0a<<
	j	47		G			????attttattatttatattaatatttatttttttttttt----------ta0a0a<<

Chapter 5 Figure 6

Repeat unit composition of MS31A alleles. The repeat unit composition of alleles was analysed for each of the two populations surveyed and was found to be similar in both, with roughly equivalent numbers of the two repeat unit types assayed (Table 5.4). The proportion of O-type repeats was ~1%, similar to that found at MS32. The distribution of null repeats across the alleles in the database is not even, alleles usually have either several null repeats, or none at all. Dot matrix analysis of the organisation of the two repeat types assayed within alleles showed that there are few clusters of a particular repeat type and no evidence of a subset of homogeneous alleles largely fixed for one repeat unit type (data not shown). The most homogenous allele observed (highlighted red, Fig. 5.6B) has a run of 39 t-type repeats.

Repeat unit distribution along MS31A alleles. A compositional scan along the mapped portions of alleles in the database showed that the probability of having a given repeat unit type at a given position is similar for all positions and comparable to the overall frequency of that repeat unit (data not shown). The first repeat unit is not scored in MVR analysis, because of a generally poor signal on autoradiographs caused by weak probe hybridisation, but it was predicted that it would be no less variable than internal repeats. Allele maps (Fig. 5.6) show that the second repeat unit, the first scored in MVR-maps, shows similar variability to internal repeat positions (44% a-type, 55% t-type). Based on this assumption, two pairs of allele-specific flanking primers, 31G and 31H (Fig. 5.2) were designed to discriminate between variant bases (A/G) at the first position of the first MS31 repeat unit of -4G MS31A alleles, in order to map single alleles from individuals heterozygous at this position. These primers were used to amplify entire MS31A alleles from 78 Caucasian AluI- homozygotes to identify heterozygous individuals, however, all individuals tested were homozygous, with a "G" at this position in Alu1$^-$ alleles (data not shown). A second pair of primers 31I and 31J (Fig. 5.2) were therefore designed to discriminate between variant bases at the second position of the first MS31 repeat unit of -4G MS31A alleles. Significantly ($p<0.001$) when these were used for allele-specific amplification of the entire MS31A locus of Caucasian -4G alleles (data not shown), 77 out of 78 alleles amplified with primer 31I only. Therefore the first repeat unit of MS31A alleles is almost always a-type (second base 5′-3′ T). The first repeat unit of the one exception was t-type (second base 5′-3′ C). For this reason neither of these primer pairs was subsequently used for allele-specific MVR-PCR.

MS31A ternary code diversity. Given the extreme allelic diversity seen at MS31A and the fine scale interspersion pattern of the two repeat unit types it was predicted that MS31A diploid codes might be generally even more informative than those seen at MS32; (eg. compare the MS31A and MS32 profiles of individual 5, Fig. 4.2B). To test this prediction a computer program was used to synthesize 54 artificial diploid codes from combinations of MS31A single-allele codes, selected at random from the database. These synthetic diploid codes were compared over the first 50 repeat units to find the mean number of differences between them (Fig. 5.7). Since this analysis had already been performed for MS32 (Fig. 3.3) using real diploid codes, the results of the two surveys could be compared. Contrary to the prediction, MS31A diploid codes (generated from the sample of single allele codes in the database) were generally no more informative than those at MS32. At both loci the number of differences between diploid codes is quasi-normally distributed, with a mean of ~30 differences over the first 50 repeat units, and diploid code information from the first 20 repeat units is sufficient to distinguish all diploid codes.

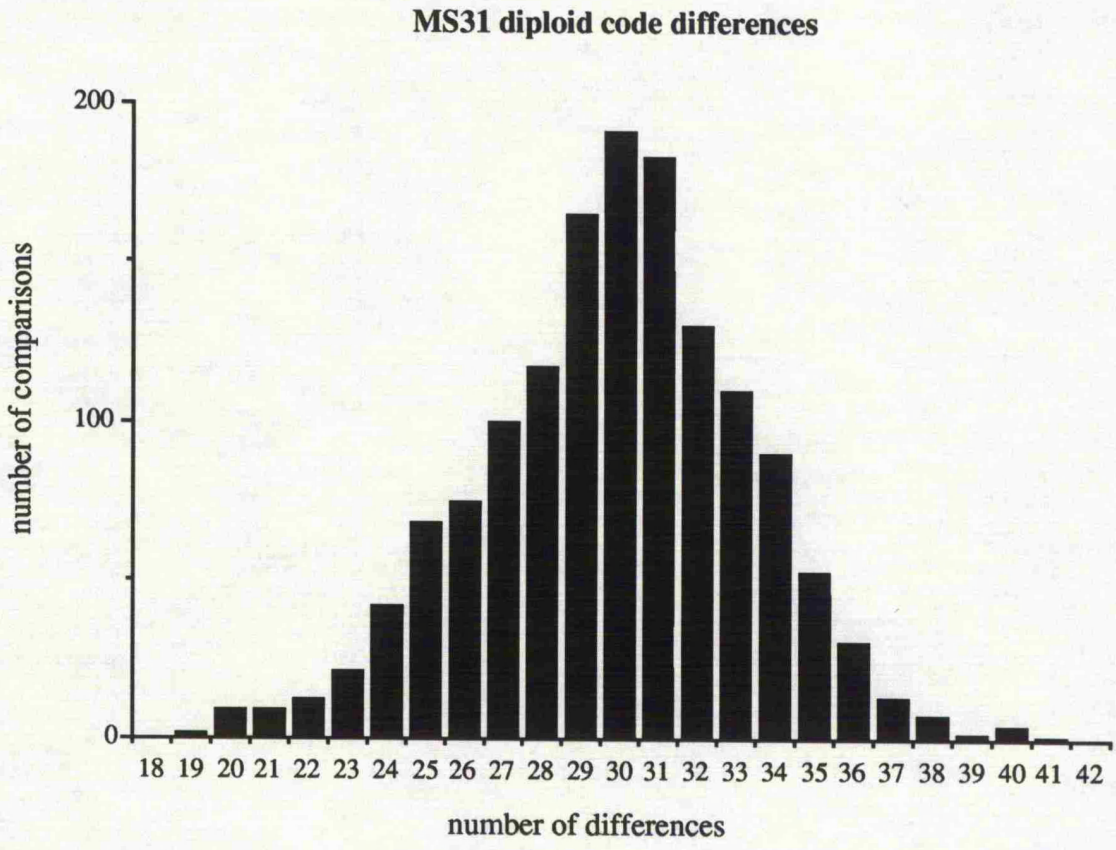


Figure 5.7. Prediction of individual variation between MS31A diploid MVR codes.

54 synthetic MS31A diploid codes were generated by a computer program (written by A.J. Jeffreys) that combined different pairs of alleles, selected at random, from the MS31A allele code database. 1431 pairwise comparisons of these diploid codes were then made and the number of differences between them calculated.

Comparative analysis of allele structures at MS31A. All possible pairwise comparisons between allele maps were made by a computer program (written by A.J. Jeffreys in Microsoft Quickbasic™) that uses dot matrix analysis to detect regions of MVR-map similarity between different alleles (see legend to Fig. 5.6 & Materials and Methods). Allele maps sharing such regions were selected and the zones of haplotypic similarity were aligned. Alignments were verified by eye and improved in some cases by the introduction of gaps. Alleles sharing large regions of MVR structure were grouped on this basis. 82 out of the 182 mapped alleles fell into 19 groups containing from 2 to 22 apparently related alleles per group (Fig. 5.6A, C). Each of the remaining 100 alleles has a structure with no obvious similarity to any other allele yet mapped. Groups of alleles tend to share flanking haplotypes, further confirming the authenticity of alignment. All groups, except 1 and 19, are population specific.

Very closely related alleles tend to show most interallelic variability in repeat copy number and interspersion pattern at the extreme 5′ end of the tandem repeat array (Fig. 5.6A). They are usually almost identical along the rest of the mapped region, which in group 2 extends nearly as far as the 3′ end, generally have similar numbers of repeat units and share the same 5′ flanking haplotype (eg. groups 2, 6 and 10). Other groups, some of which have alleles with different flanking haplotypes and allele lengths, show patches of identity interrupting, or interrupted by, unrelated segments (eg. groups 6, 7 and 9). Alignments within such groups sometimes break down towards the 3′ end of the mapped region (eg. group 6). Among the groups of aligned MS31A alleles there are many examples of small internal differences between regions with an otherwise identical haplotype; these mostly comprise one or two repeat unit deletion/insertion events and the switching of single a-type and t-type repeat units without change in repeat copy number or disruption of alignments. Groups 7 and 10 have examples of a to O and t to O switches respectively, which are probably due to the incorporation of repeat sequence variants that block priming, or to recent base substitutional mutation.

Southern blot length analysis of over 100 unrelated Caucasian individuals showed that the majority of the shortest MS31A alleles cluster around 2kb (~100 repeat units). At least some of these are closely related and belong to group 2 (Fig. 5.6A). Only two alleles shorter than this have been identified in Caucasians, one of which has been mapped (highlighted red, Fig. 5.6B). The other is very small (~15 repeat units) and difficult to MVR-map. Preliminary data (not shown) suggests that it has an abnormal structure, largely comprised of null repeats.

A group of short and unusual Japanese alleles. The Japanese have a higher frequency of shorter MS31A alleles than Caucasians. 25% of Japanese alleles mapped to date have around 50 repeat units. Strikingly all of these appear to be very highly related, belonging to group 19 (Fig. 5.6C). Within this group there are several examples of alleles with the same length, and two examples of alleles with indistinguishable internal structure. Although the alleles in group 19 are clearly very closely related, there is less evidence of 5′ polarity in allelic variation between alleles in this group, than in other groups. Rather, the alleles are distinguished by the switches of repeat type and deletions/insertions that are commonly observed in other groups of MS31A alleles, although in this group deletion/insertions can involve up to a quarter of the allele's full length. This is the only group of alleles for which 3′ MVR structure has been completely mapped; the last six repeat units are the same in all alleles in this group.

Discussion

MS31A allelic diversity. The alleles in the current database are not a representative sample of those in the populations from which they were selected. Several of the Caucasian alleles were selected on the basis of size and those outside the normal size range (4-13kb) were often found to be related (eg. group 2. Fig. 5.6A). Many single allele codes were generated by allele-specific MVR-PCR from the -4A/G site and since -4A alleles are less frequent (0.12 in Caucasians, 0.24 in Japanese) than -4G alleles, they are over-represented in the database (0.31 Caucasians, 0.44 Japanese). This is reflected in the similar numbers of groups of -4G and -4A alleles, 11 and 9 respectively. This effect is not as pronounced for alleles mapped from the -221G/C site because there are more even numbers of the two alleles here. Further allele-specific mapping using all three flanking polymorphic positions should help to redress any imbalances that these sampling errors have caused. Although the MS31A database is still small and is also biased, with these caveats in mind, it is still informative to look at the structures of the alleles that have been mapped to date (Fig. 5.6, A, B & C) and draw some inferences about the behaviour of this minisatellite from them.

This preliminary survey of allelic variability at MS31A has revealed extraordinary levels of MVR code variation, greater than at the other minisatellites analysed by MVR mapping in this laboratory (Jeffreys *et al.*, 1991a; Armour *et al.*, 1993). Actual MS31A heterozygosity is far higher than the 98% estimated by Southern blot allele length analysis (Armour *et al.*, 1989b) and greater than the heterozygosity of 99.1% observed by MVR at MS32 (Jeffreys *et al.*, 1991a). Based on current human population size and the estimated mutation rate of 1% per gamete it is likely that the number of alleles worldwide is very large, perhaps $>10^8$.

Almost without exception MS31A alleles from both populations surveyed have a better balance and finer scale interspersion of the two repeat types assayed along their mapped regions, compared with alleles mapped at MS32 (Jeffreys *et al.*, 1991a; Monckton, 1993c) (Table 5.3; Fig. 5.6). The majority of alleles are devoid even of regions of significant MVR code similarity and there is no evidence of a subset of homogeneous alleles largely fixed for one repeat unit type (data not shown), as is seen at MS32 (Jeffreys *et al.*, 1990, 1991a). There are fewer short alleles (less than 50 repeat units) at MS31A than at MS32, which means that the potential problem of scoring codes beyond the end of short alleles (see Chapter 4) will not be encountered often in MS31A diploid MVR-PCR. Short MS31A alleles are more common in Japanese than in Caucasians, an observation which also applies to MS32 (Monckton, 1993b). Although individual MS31A allele codes may be generally more informative than those at MS32, diploid codes from MS31A are not more diverse than those generated from MS32 (compare Figs. 3.3 and 5.7). Surprisingly, it is predicted that MS31A diploid codes will be no more likely to distinguish two unrelated individuals than those from MS32. This indicates that the increased allelic variability of MS31A does not significantly increase its discriminatory power beyond the already high level of informativeness already afforded by MS32 diploid coding.

Forensic applications. The potential forensic uses of diploid MVR-PCR have already been discussed (Jeffreys *et al.*, 1991a; Chapters 3 and 4). Allele-specific MVR-PCR may extend the range of these applications, particularly where mixed DNA samples are encountered, for example DNA recovered from vaginal swabs in rape cases. It has already been shown that ambiguous diploid codes from MS32 can be obtained by standard MVR-PCR down to approximately 10% admixture. When a pure sample of one of the DNAs contributing to the mixture is available, for example that from the victim, a high level of discriminatory power can be achieved (Jeffreys *et al.*, 1991a; Chapter 3). Allele-specific MVR-PCR allows the specific amplification an allele from DNA comprising less than 10% of

such a mixture, providing that it differs in the genotype of at least one flanking polymorphic position (Monckton *et al.*, 1993). Depending on the genotypes of the two DNA samples, either single-allele or diploid code information from one of the individuals contributing to the mixture may be obtained by using a suitable allele-specific primer. MS32 single allele codes obtained in this manner were used to interrogate the diploid code database with good exclusionary power. Most alleles exclude all false suspects; however, the small subset of MS32 alleles largely homogenised for a-type repeats are less informative and reduce overall discriminatory power (Monckton *et al.*, 1993). It is likely that MS31A allele-specific MVR-PCR would also be amenable to the analysis of mixed DNA samples, although this application has not been investigated. Allele-specific MVR-PCR at both loci would be expected to increase exclusionary power, particularly in those cases where MS32 alleles were less informative.

Analysing allelic variation. The development and application of allele-specific MVR-PCR at MS31A has been of vital importance in further characterising allelic structure and variability at this hypervariable locus. It will now be possible to map the MS31A alleles of large numbers of unrelated individuals from diverse populations, both to investigate relationships between MS31A alleles, and also to compare structural variability, and the mechanisms that give rise to it, with other hypervariable minisatellite loci. Most interestingly, many apparently very closely related MS31A alleles tend to show MVR-haplotype variation preferentially restricted to the beginning of the tandem repeat array. This gradient of variability within groups of alleles is similar to that seen at MS32 (Jeffreys *et al.*, 1990, 1991a; Monckton, 1993c), where polarity is due to the localisation of most spontaneous mutational changes in repeat copy number, and therefore allelic structure, to the start of the tandem repeat array (Jeffreys *et al.*, 1991a; Chapter 3). The data from MS31A are consistent with the hypothesis that there is also a mutational hotspot active at the 5′ end of this minisatellite, since this would be predicted to cause the observed polarity in variation between groups of related alleles (Jeffreys *et al.*, 1990). It is therefore possible that a similar mutation process is responsible for the generation of the majority of allelic variability at MS31A and MS32.

At MS32 the region distal to the variability hotspot shows limited polymorphism in map structure (Jeffreys *et al.*, 1990, 1991a; Monckton, 1993c). There is little information concerning 3′ map variability at MS31A and therefore no data on mutation at this end of the locus. Since grouped alleles are often of similar lengths, regions of shared MVR map similarity may extend to the 3′ end of the alleles. This is indeed the case for the short alleles in group 2, whose maps extend close to the 3′ end, and those in group 19, which were completely mapped and have identical extreme 3′ haplotypes. However, extrapolations drawn from groups 2 and 19 must be treated with caution, since these short alleles lie outside the normal MS31A allele size range and may therefore have atypical structure. Alignments within groups sometimes break down towards the 3′ end of the mapped region (eg. group 6, Fig. 5.6A) which could be due to the occurrence of occasional intraallelic deletions or internal interallelic recombination or conversion events. As no information is yet available on the level of variability at the 3′ ends of MS31A alleles these possibilities are indistinguishable. In order to investigate 3′ variation and mutation at MS31A it will be necessary to overcome the potential difficulties of developing an MVR mapping system for this end of the locus (see introduction to Chapter 4). This can probably be achieved by obtaining additional, 3′ sequence information that will enable the design of additional flanking primers.

A group of unusual alleles. The alleles in group 19 (Fig. 5.6C) provide an interesting departure from the general observations concerning allelic variability at MS31A. This group contains several alleles of the same length, including two pairs of apparently identical MS31A alleles. In the case of the 45 repeat allele there are two possible

explanations; firstly, this may represent the erroneous double entry of a single-allele code from the same individual into the database, for example due to mislabelling of a sample. Secondly, this allele may be present in two closely related individuals. In the case of the two indistinguishable 52 repeat alleles the first of these explanations can be ruled out, since the other allele from each of these individuals has been mapped, and is different. However, although the 52 repeat alleles are from DNA samples collected at different times, it still conceivable that they are from relatives. Assuming that both of these pairs of alleles are from unrelated individuals, their maximum population frequency is still low (2.5%). As a whole the alleles in group 19 constitute 25% of all Japanese alleles mapped; the fact that these are all -4G alleles indicates that this is not a result of sampling bias. By comparison the highest proportion of Caucasian alleles mapped to fall within a single group is 9%, representing the 9 Caucasian alleles in Group 1. This group exhibits considerable 5′ length and MVR map variation, typical of alignable groups at both MS31A and MS32. By contrast there is much less 5′ variation in structure between alleles in group 19. It is unfortunate that in most cases the first few repeat units of these alleles were difficult to map and so provide no information about extreme 5′ MVR structure. Despite this, it appears that these alleles differ largely by virtue of non-polar changes in internal MVR-haplotype. Amplification of MVR maps to levels detectable by ethidium bromide staining (Jeffreys *et al.*, 1993), which enables the first few repeat units to be scored more easily, would help to test this assertion. The features exhibited by this group would be expected to arise if these alleles were refractory to the mechanism of repeat unit turnover that generates most interallelic variability at MS31A. This could explain both the apparent lack of polarity and elevated population frequency of alleles within this group. It is perhaps noteworthy that the most obvious example of a polar difference in this group is from a different ethnic group (Malay). Interestingly over 50% of the short Japanese alleles mapped at MS32 also fall into a single group and have an unusual structure, being largely homogenised for a-type repeats.

Population analysis. As more alleles are mapped and grouped, and their flanking haplotypes determined it may become possible to make deductions about the derivations of allele lineages, and hence the relationships between human populations, particularly when this information is analysed in conjunction with MVR data from other hypervariable minisatellites. When individuals are assigned to large population subgroups, in this case Caucasian or Japanese, most of the groups of related MS31A alleles assembled so far are found to be population specific. This was also found to be the case with MS32 alleles (Monckton, 1993b). However, given the extreme allelic variability of this locus and the small numbers of alleles sampled from only two populations involved, it is likely that as further alleles are mapped some groups will expand to incorporate alleles of differing ethnic origins. Population-specific MVR haplotypes may represent either new haplotypes arising after divergence, or ancient haplotypes which have been differentially lost, through drift, in some populations. If MVR maps are to be used to draw inferences about population dynamics it will be important to be able to differentiate between ancient and modern haplotypes, in order to distinguish these scenarios. The data collected at MS31A suggest that this may be possible. Group 1 (Fig. 5.6C), contains alleles from both the major populations surveyed, suggesting that the related segment of these haplotypes is ancient, pre-dating the branching of lineages that gave rise to present day Japanese and Caucasian populations some 50,000 years ago (Nei & Roychoudhury, 1974). Alleles in this group show considerable 5′ MVR map and flanking haplotype variation, with alignments confined to relatively small patches of haplotypic homology in some cases. These features are also consistent with an ancient origin for the region of MVR-map homology shared between these alleles. In contrast alleles in the other groups generally display large zones of MVR map identity, almost exclusively terminal divergence and conservation of flanking haplotype, features which imply more recent derivation. Similar observations have been made for groups of related alleles at MS32 (Monckton, 1993b).

Mutational inferences. Groups of related alleles at both MS31A and MS32 tend to share common flanking haplotypes (see Monckton, 1993c for grouped MS32 alleles with flanking haplotypes). This implies that the operation of turnover mechanism(s) that give rise to the majority of allelic diversity at these loci is largely confined to the beginning of the minisatellite array itself and does not usually extend into the flanking DNA. At MS31A 5′ variation between related alleles does not usually include the polymorphic marker only 4bp from the first repeat unit, which itself appears to be approaching fixation. Therefore it seems that the 5′ ends of these loci are not hotspots for simple interallelic unequal recombination, but rather that other processes are more likely to be the predominant mechanism(s) for generating allelic diversity. Intraallelic events could arise from USCE or slippage and evidence for the involvement of interallelic events at MS32 suggests that localised gene conversion may also be involved. However, occasional switches of flanking haplotype within groups of related alleles (eg. groups 1, and 7, Fig. 5.6A) are observed at both of these loci. At MS31A all possible 5′ flanking haplotypes are found in the Japanese population showing that linkage disequilibrium between the sites is not complete. The haplotype -4A/-221C alleles was not found in Caucasians, but this is predicted by allele frequencies to be the least common haplotype and the total number of haplotypes determined was small, so it seems likely that as more Caucasian alleles are surveyed this haplotype will be found in rare cases. It is dangerous to speculate about the mechanisms giving rise to assortment of flanking haplotypes without knowing the ages of these groups of alleles. If the alleles in a group are ancient, such differences could represent rare interallelic recombination events; without information from more distal flanking markers this will be impossible to investigate. If these groups have more recent ancestry, as the population specificity of groups suggests, differences in flanking haplotype may be caused by minisatellite mutation events that extend to the flanking DNA.

A closer inspection of allele maps both within and between groups can also provide clues as to the sorts of processes that may be involved in the generation of new length alleles at MS31A. Some alleles have zones of internal tandem duplication (eg. regions arrowed above the top allele in group 2, Fig. 5.6A) similar to those already noted at MS32 (Jeffreys *et al.*, 1990), suggesting that USCE or replication slippage may play a part in MS31A evolution. There is one possible example of a conversion like event; in group 7 two closely related alleles with the same flanking haplotype and similar length share an internal region of similarity with a third allele of markedly different length and flanking haplotype. Most alignable MS31A alleles also show small differences scattered internally along regions of alignment. These differences usually involve the gain or loss of one, or a few repeat units, or the apparent switching of one repeat unit type for another, without a local change of repeat copy number. For example the alleles in group 14 would be indistinguishable, but for three single repeat unit insertions/deletions. There are more variants of this type at MS31A than at MS32, suggesting that they may be generated by processes distinct from those operating at the putative mutation hotspot at the beginning of the array, which seem to be common to both loci. They could, for example, be generated by a slippage type mechanism, with a mutation rate that increased as repeat unit length decreased. This would explain why MS31A, with a 20bp repeat unit, shows more of these variants than MS32 which has a 29bp repeat unit.

To gain a proper insight as to the mechanism/s of repeat unit turnover at MS31A it is necessary to analyse mutation processes directly by characterising *de novo* mutation events, in particular to assess the contribution of any putative mutation mechanism to the creation of a localised variability hotspot at the beginning of the tandem repeat array. Analysis of MS31A mutation is presented in Chapter 6 and compared with mutation at other loci in Chapter 7.

Chapter 6

MUTATION AT MS31A

Summary

In common with other human hypervariable minisatellites, MS31A has a mutation rate to new length alleles high enough to be detected by Southern blot length analysis of pedigrees. 29 MS31A germline mutation events were detected in this way. These show a considerable bias toward mutation in the male germline, which has a mutation rate of 1.4% per gamete, compared to the female germline, where the mutation rate is 0.3% per gamete. Comparison of the MVR maps and 5′ flanking haplotypes of single mutant alleles and their progenitors has enabled individual mutation events to be studied in detail. This analysis suggests that simple unequal recombination is not the dominant mutation mechanism at MS31A, but rather, that there are at least two distinct mutation processes operating at this locus. The first appears to be specific to the male germline, and exclusive to gains in size caused by the introduction of small numbers of repeats into the 5′ end of the tandem repeat array. This confirms the hypothesis of ultravariability at this end of MS331A alleles, already suggested by comparison of related alleles. Both intraallelic and interallelic size gain mutations are seen, with no evidence for exchange of flanking markers associated with interallelic exchanges, which are sometimes complex. Therefore, it seems that the variability hotspot observed at the 5′ end of MS31A alleles is caused by the localised action of male germline specific, unequal conversion-like processes. The second category of mutation event involves the deletion of varying numbers of repeat units. Deletions are not restricted to the 5′ end of the tandem repeat array and were only observed in the female germline. A few examples of somatic mutation at this locus were also observed. Some of these were detected in DNA derived from lymphoblastoid cell lines and probably reflect instability of this locus during cell culture. However, one was confirmed as a genuine somatic event, probably occuring in the early embryo. MVR mapping showed that this was likely to be an intraallelic reduplication, again close to the 5′ end of the progenitor allele This mutation provides the only known example of a human exhibiting mosaicism for a mutant minisatellite allele. Although there are no data concerning mutation events too small to be resolved by Southern blot length analysis, the mutation profile of MS31A in the male germline appears to be remarkably similar to that seen at MS32. This observation suggests that the major mechanisms of repeat unit turnover shaping the evolution of these different hypervariable minisatellite loci, and therefore possibly others, may be the same.

Introduction

Minisatellite mutation rates. The most distinctive feature of human hypervariable minisatellites, including those studied in this work, is the presence of large numbers of different length alleles, containing different numbers of repeat units, that results in high levels of heterozygosity in human populations (Wong *et al.*, 1986, 1987; Nakamura *et al.*, 1987a; Armour *et al.*, 1989b, 1990; Vergnaud *et al.*, 1991). It is reasonable to assume that these minisatellites are without phenotypic effect and therefore that this extreme variation in allelic tandem repeat copy number arises from a high spontaneous germline mutation to new length alleles. At some loci this mutation rate is so high that it can be measured directly by pedigree analysis (Jeffreys *et al.*, 1988; Armour *et al.*, 1989b). For example the rate is about 1% per gamete at MS31A and MS32, and can be as high as 15% per gamete (Vergnaud *et al.*, 1991). Direct measurement of mutation rates in germline and somatic tissue is of direct relevance to forensic and legal applications of SLPs. Germline mutations will produce apparent exclusions in paternity testing and somatic mutation could in principle produce differences in DNA phenotypes between tissues (eg. blood and sperm) in the same individual. More important, from the point of view of this thesis, is the need to analyse *de novo* mutation events in order to unravel the molecular processes which generate variability at these loci.

Sequence considerations. There has long been speculation as to the mechanisms generating such high levels of variation at these loci and the possible role, if any, of minisatellites in the human genome. Initial observation of the repetitive nature of these sequences and their high mutation rate to new length alleles led to speculation that allelic variation may have arisen through unequal exchange between misaligned minisatellite alleles; it was proposed that this was a result of the functional involvement of minisatellites in promoting chromosome synapsis and/or meiotic recombination. At first this hypothesis was supported by a wealth of circumstantial evidence. Among the first few single locus minisatellite repeat sequences discovered, were those isolated by virtue of cross-hybridisation to the probes 33.15 and 33.6 derived from a tandemly repeated sequence in the first intron of the myoglobin gene (Jeffreys *et al.*, 1985a). Although each of these had a different sequence, they all shared an apparently conserved G/C rich core region of 11 to 16bp which exhibited a high degree of sequence similarity to the χ recombination signal of *E. coli* and was therefore hypothesised to act as a recombinogenic signal in the human genome (Jeffreys *et al.*, 1985a). However, as further G/C rich minisatellites, including MS31A, were cloned, following detection by cross-hybridisation to the polycore probes, increasing divergence from the consensus sequence was observed (Wong *et al.*, 1987). New core sequences were proposed to accommodate these findings, reducing the supposedly conserved core consensus to a few Gs and Cs (Dover, 1989), but the discovery of minisatellites containing A/T rich repeats, with no obvious similarity to the proposed core (Huang & Breslow, 1987; Vergnaud *et al.*, 1991), threw its relevance into considerable doubt. Despite these observations, it remains possible that the core has significance as a sequence-specific component of minisatellite instability for a subset of loci. Sequence similarities between the insulin HVR and the 3′ α-globin cluster HVRs have also been noted (Jeffreys *et al.*, 1987b), but these have little homology to the proposed core sequence, suggesting that certain classes of sequence may be predisposed towards forming minisatellites. The common feature of these and the core containing minisatellites is marked purine-pyrimidine strand asymmetry. Therefore strand composition, rather than precise sequence, may cause the instability exhibited by these loci. Studies of plasmids containing a number of direct 12bp GC-rich repeats with pronounced strand asymmetry from *Herpes simplex* virus type 1 have shown that these sequences can adopt unusual conformations *in vivo*. This has been called anisomorphic DNA and can exert physical stress on the DNA helix, leading to strand

bending or double strand breakage (Wohlrab *et al.*, 1987). Therefore it is possible that localised distortion or disruption of the DNA in this way, perhaps at the interface between anisomorphic and normal DNA is at least partly responsible for the high levels of mutation seen at some minisatellite loci. It may be significant that anisomorphic DNA sequences are bound by a conformation-specific, as opposed to sequence-specific, nuclease (Wohlrab *et al.*, 1991). However, the observed similarities between minisatellite loci may equally be an artifact of ascertainment bias due to the probe sequences used to isolate them by cross hybridisation, or be the consequence of a propensity for certain G/C rich sequences to become minisatellites given the correct wider molecular environment. The biased chromosomal distribution of human minisatellites may be a reflection of the latter possibility.

Clues from genomic location. *In situ* hybridisation and linkage mapping revealed that hypervariable human minisatellite loci are not uniformly scattered along human chromosomes, but show a strong, though not exclusive, tendency to cluster in the proterminal regions at, or near, the ends of genetic linkage maps (Royle *et al.*, 1988; Armour *et al.*, 1989a; Nakamura *et al.*, 1988; Lathrop *et al.*, 1988; Armour *et al.*, 1990; Vergnaud *et al.*, 1991). The subtelomeric regions of human chromosomes have been shown to contain DNA with the highest G/C content and are known to have a high recombination rate and to be the sites of initiation of chromosome synapsis and pairing during meiosis (Solari, 1980; Laurie & Hulten, 1985). These features, and the observation that minisatellite core probes hybridise preferentially to chiasmata in human bivalents, were further suggestive of a link between minisatellites and recombination (Hulten, 1974; Laurie & Hulten, 1985; Chandley & Mitchell, 1988). Similar conclusions are suggested by the genomic distribution of minisatellite loci. While all the (non-acrocentric) autosomal telomeric regions appear approximately equally rich in minisatellites, there is marked contrast between the two ends of the X chromosome. Despite intense efforts to map the X chromosome, very few polymorphic minisatellites have been isolated from its sex-specific region, which has no partner in male meiosis (Donis-Keller *et al.*, 1987; Fraser *et al*, 1989; Armour *et al.*, 1990; Consalez *et al.*, 1991). By contrast, the pseudoautosomal region, a region of high recombination in male meiosis, is very rich in minisatellite loci (Cooke *et al.*, 1985; Page *et al.*, 1987). There are two broad explanations as to the nature of the link between minisatellites and recombination and the origins and possible functions of hypervariable minisatellites in these dynamic regions of the human genome (see Jarman & Wells, 1989). Firstly, minisatellites might evolve as highly unstable, and coincidentally G/C rich, repeated sequences due to the local action of recombination or other mechanisms in the subtelomeric regions. On the other hand, minisatellites may be "hotspots" for meiotic recombination and evolve as a consequence of their own, possibly sequence directed, recombinational proficiency. Different lines of evidence have been converging for some time to that suggest that the latter explanation is not the case.

Species comparisons. If minisatellites were intimately involved in such fundamental processes as chromosomal recombination they would be expected to be present and behave in the same way in closely related species. PCR amplification and sequence analysis have shown that homologues of several hypervariable human minisatellites are indeed present in several primate species. However, these are often short and monomorphic (Gray & Jeffreys, 1991), implying that these primate sequences do not participate in the same molecular turnover processes as their human counterparts. This argues against, although it does not exclude, a conserved functional role for minisatellite sequences, but does not rule out the possibility that some minisatellites may evolve to high repeat copy number due to the proximity of local recombination hotspots.

Indirect mutational analyses. Inferences of possible minisatellite mutation mechanisms were initially made by considering the general properties of minisatellite mutations observed in pedigree analysis. Southern blot length analysis was used to study the inheritance of 5 hypervariable minisatellites in the CEPH panel of large families, and relatively high rates of germline mutation, detected as clonal, heritable changes in allele length, were found at some of these loci (Jeffreys *et al.*, 1988a). The mutation rates of these loci were found to increase with heterozygosity, consistent with a neutral mutation /random drift model for their evolution. Analysis of allele size changes showed that small changes, involving gain or loss of a few repeats, were responsible for most mutations and that gains and losses of repeat units occurred with approximately equal frequency (Jeffreys *et al.*, 1988a). These observations ruled out a major role for intramolecular recombination, which can only decrease allele size, in minisatellite mutation, but were consistent with the involvement of intraallelic mechanisms, such as USCE (with reciprocal products, Smith, 1976) and/or replication slippage (Tautz *et al.*, 1986; Levinson & Gutman, 1987), or interallelic processes, like unequal recombination and gene conversion (Dover, 1982). In this study mutation rates in the male and female germline appeared to be approximately equal. Since mature spermatocytes have undergone many more postzygotic cell divisions (~400) than oocytes (~24) (Vogel & Rathenburg, 1975), this observation argued against a role for mitotic recombination or replication slippage, both of which would be assumed to depend on the number of mitotic divisions involved, and implied that mutation is restricted to one stage of gametogenesis, possibly meiosis (Jeffreys *et al.*, 1988a). However, 19 out of 23 paternal mutations and 17 out of 20 maternal mutations in this study all occured at one of the five loci, MS1 which has a markedly higher mutation rate than the others (~5% vs ~1%), severely biasing this analysis. More recent evidence has revealed that several loci, including MS1, MS31 and other minisatellites used in this initial study, have a pronounced bias towards size increase mutations arising preferentially in the male germline (Olaisen *et al.*, 1993; Henke *et al.*, 1993 & personal communication). An exceptional example is the hypervariable human minisatellite, CEB1 (D2S90), isolated by Vergnaud *et al.*, (1991), which has a very high, but male-specific, germline mutation rate of ~15% per gamete (female mutation rate ~0.3% per gamete). This evidence has revitalised speculation that much of the mutation occuring at these loci might be mitotic in origin.

Direct mutational analyses. More concrete deductions about the possible nature of minisatellite mutation mechanisms can only be made by direct analysis of *de novo* mutation events. The hypothesis that simple unequal recombination is involved can be experimentally investigated, since it would be expected to result in the exchange of flanking markers. This prediction was tested in several cases by determining the linkage phase of markers flanking mutant minisatellite alleles. A single *de novo* mutation event occuring in the maternal germline at the human minisatellite YNZ22 was shown not to involve exchange of immediately flanking markers, suggesting that it was intraallelic (Wolff *et al.*, 1988). This group also made a more comprehensive study of 12 mutations, occuring at the hypervariable human minisatellite MS1 (D1S7), in families from the CEPH panel. They showed that flanking markers within the surrounding 10cM interval were not exchanged at a significantly elevated rate during mutation (Wolff *et al.*, 1989). Vergnaud *et al.*, (1991) also found no evidence for an increase in the exchange rate of distal markers associated with mutation events at the CEB1 locus. These results suggested that simple unequal recombination is unlikely to be the mechanism generating the majority of length variants at these loci, although a possible minor contribution was not excluded. However, these studies had limited resolving power because the flanking markers analysed were frequently many cM away from the locus undergoing mutation. Since patterns of repeat unit turnover and markers much closer to each locus were not examined, more complex and localised interallelic exchange mechanisms, for example unequal interallelic conversion, could not be ruled out. Furthermore,

no information was provided concerning intraallelic processes that may have been involved, for example USCE or replication slippage.

Mutation analysis by MVR-PCR. Studies of variant repeat interspersion within minisatellite arrays can be used to investigate mutational mechanisms in greater detail, both by comparing internal allelic variability in populations, and by the analysis of structural changes occurring during *de novo* length-change mutations. MVR-PCR provides the means by which both of these objectives can be achieved. Although MVR analysis of single alleles from unrelated Caucasians reveals very high levels of allelic variability at MS32 and MS31A, several of the alleles sampled so far at each locus do appear to be related. These are almost identical along most of their length, with the major region of difference between them apparently confined to one end. This observation implies that the extraordinary levels of allelic variability revealed by MVR mapping at both of these hypervariable minisatellite loci must be maintained by a high-rate *de novo* mutation process that is polar with respect to the orientation of the minisatellite allele. In order to identify the underlying mutation processes, it is necessary to identify both mutant and progenitor alleles and compare their structures at the molecular level. Detailed analysis of mutations involving interallelic exchanges can elucidate whether the mutation has occurred by a simple unequal recombination, including exchange of flanking markers, or whether the exchange is confined to the repeats.

MVR mapping was first used in mutation analysis to study deletion alleles, isolated from size fractionated sperm and blood DNA by PCR amplification of single molecules (Jeffreys *et al.*, 1990). Deletions were present in both tissues at a frequency of ~0.07% per haploid genome, with small deletions more frequent than larger ones and a significant clustering of deletion endpoints towards the 5′ end of alleles (designation of 3′ and 5′ ends at MS32 has been reversed since the publication of this paper). No examples of of interallelic exchange were found and a low level of mosaicism in sperm DNA indicated a premeiotic origin for at least 40% of the new mutants detected.

MVR-PCR at MS32 has provided a very useful tool for the detection and detailed molecular dissection of *de novo* minisatellite mutation events (Jeffreys *et al.*, 1991a, 1994; Chapter 3). Small mutations, not resolvable by Southern blot length analysis, were identified, allowing the more accurate quantification of mutation rate, and the comparison of the MVR maps of mutant and progenitor alleles allowed the more accurate assignment of possible mechanisms to mutation events at this locus. All seven of the mutations detected by pedigree analysis at MS32 involved gains of small integral numbers of repeat units close to the 5′ end of the tandem repeat array; most of the events were probably intraallelic, but two provided the first direct evidence of possible interallelic unequal exchange. In one of these, the first 11 repeats from one allele were inserted into the first repeat unit of the other, with no loss of information from the recipient allele. Two extra repeats, of unknown origin were found in the recipient at the 3′ insertion boundary, and a flanking marker 269bp 5′ to the minisatellite (Hump1, Monkton, 1993d) was not exchanged. In the other event it appeared that three repeats from one allele were inserted into the first repeat of the second. Unfortunately, in this case flanking markers 5′ to the progenitor allele were uninformative (Jeffreys *et al.*, 1991a). These results indicated for the first time that the generation of new alleles by interallelic exchange is not always a process of unequal recombination, but that unequal interallelic gene conversion may play a significant role in the generation of minisatellite ultravariability.

All of the experimental results described above suggest that simple unequal recombination is not the predominant mechanism of human hypervariable minisatellite mutation, but that other interallelic and intraallelic processes are involved. These have been best characterised at the MS32 locus (Jeffreys *et al.*, 1990, 1991a, 1994) where it appears that allelic variability is shaped by two major mutation pathways. One is predominantly intraallelic and results in deletions, sometimes of relatively large numbers of repeat units, which occur at similar frequency in germline and somatic tissues. These deletions remove repeats from internal regions of minisatellite alleles, but show a tendency to occur toward the 5′ end. The other exhibits much greater polarity; its action is localised to the extreme 5′ end of the tandem repeat array and results in the gain of small numbers of repeat units. These may be derived from interallelic exchanges involving a short stretch of repeat units, possibly in a gene conversion-like process, or from the same allele, perhaps by USCE or replication slippage. This process appears from pedigree analysis to operate at a higher frequency in the germline than that causing deletions and is proposed to be the source of the majority of allelic diversity at this locus.

The Gap Expansion Model of minisatellite mutation. Features of MS32 mutation such as lack of exchange of flanking markers, size gain bias, unidirectional terminal polarity and conservative insertion of interstitial sequences into a terminal site in the recipient allele, were incompatible with preliminary models of minisatellite mutation involving recombination and/or conversion, for example that proposed by Wolff *et al.*, (1991). This led to the formulation of an alternative hypothesis, the gap expansion model (GEM) (Monckton, 1993e) to explain the mechanism by which size gain mutation events may occur. The GEM is based on the double stranded break repair (DSBR) model of homologous meiotic recombination proposed by Szostak *et al.*, (1983) (see also Sun *et al.*, 1991a). It proposes that minisatellite length gain mutations are initiated by DSBs that occur close to the beginning of the tandem repeat array and result from their aberrant repair. If the free ends of such a break diffuse apart (gap expansion) and then misalign on the homologous chromosome or sister chromatid, which provides the repair template, this will result in the generation of new material, (ie. a size increase through unequal conversion; Fig. 6.1A) during subsequent DNA synthesis and repair.

The polarity of deletion endpoints at MS32 suggests that it is possible that deletion mutations are initiated in the same way (ie. a DSB) as small size gain mutations, but result from processing via an alternative intramolecular repair pathway. If the free ends of such a DSB diffused laterally together, before misalignment and repair, a deletion of material would result (gap collapse, Fig. 6.1B). *In vitro* experiments have been used to show that DSBs in mitotic cells are preferentially repaired via a non-conservative single-stranded annealing (SSA) mechanism (Lin *et al.*, 1984). In this process 5′ exonucleases digest the 5′ strand next to a DSB, leaving single stranded 3′ overhangs. Regions of intramolecular homology then anneal, non-homologous DNA is excised and duplex DNA is regenerated by gap filling DNA synthesis and ligation. Tandem repeat arrays have many units of internal homology and the potential for SSA repair mediated collapse of an interstitial DSB is presumably high. Repair of random DSBs occuring within the minisatellite array in this way would also result in the deletion of repeat units (Fig. 6.1C). The low level of mosaicism and equal frequencies of deletions in somatic and germline tissues seen at MS32 (Jeffreys *et al.*, 1990) suggest that a proportion of deletions occur during mitosis. Either, or both, of these mechanisms could account for deletions of repeat units of the sort seen at MS32.

Figure 6.1 Aberrant repair of DSBs occuring within minisatellite tandem repeat arrays.

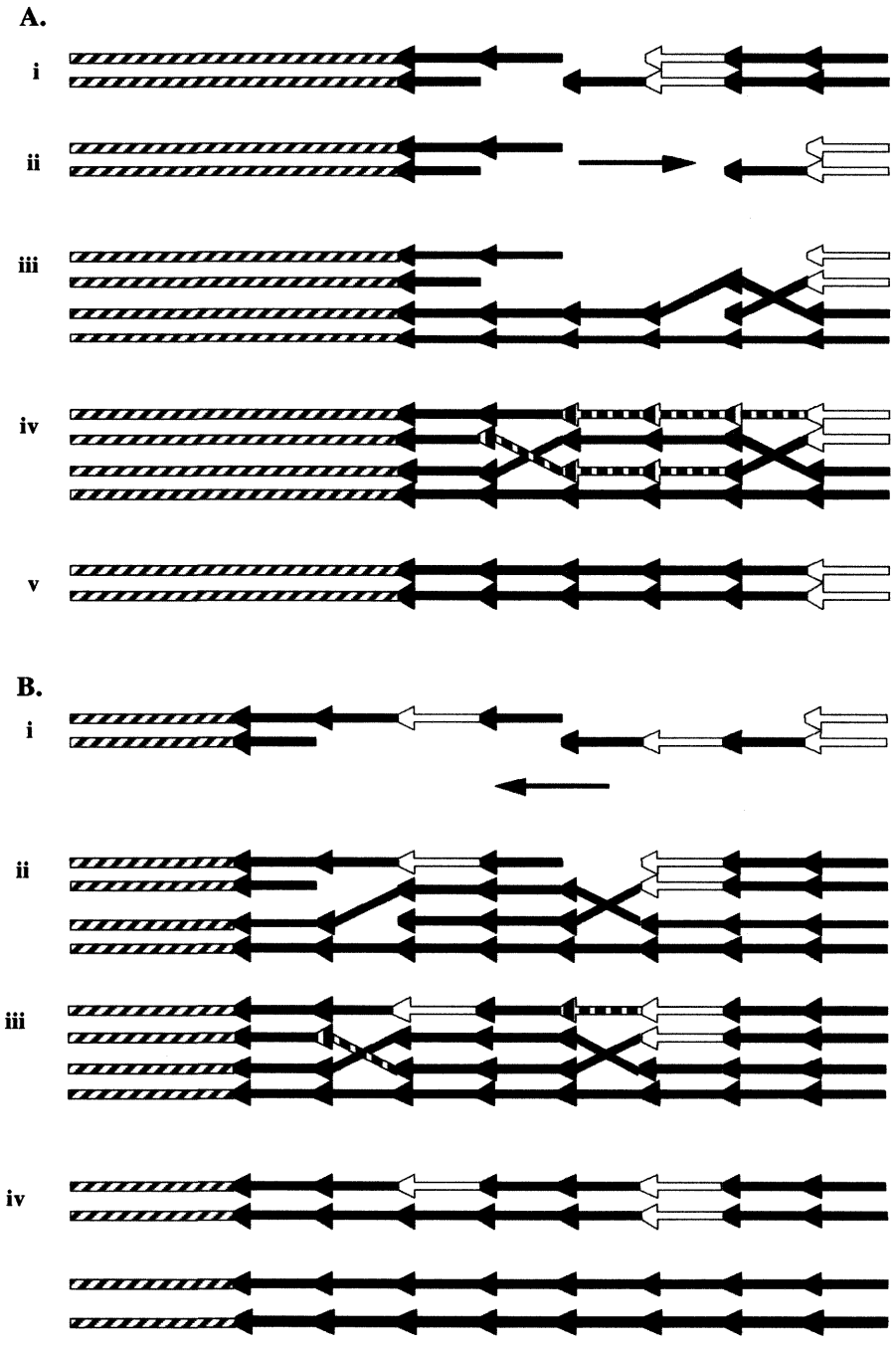
In each case the progenitor allele is depicted by red and white repeat units and the donor allele is shown in black. Diagonal stripes indicate the 5′ flanking DNA and filled arrows the minisatellite repeats. The donated segment in interallelic exchanges is coloured green. Vertical stripes show newly synthesised DNA.

A, B, C i. Mutation is initiated by a DSB in the progenitor allele, 5′ exonuclease activity degrades the 5′ strand to generate 3′ single stranded overhangs on both sides of the break.

A, ii. Gap expansion from a DSB close to the 5′ end of the progenitor allele by lateral diffusion of the minisatellite free 3′ end. **iii.** Out of alignment strand invasion of the minisatellite free 3′ end into the donor allele. **iv.** D-loop expansion promoted by DNA synthesis and annealing of the second free 3′ end, in register. Repair synthesis and ligation form a double Holliday junction structure. **v.** Resolution yields recombinant conversion products, only the converted recipient product is shown (including regions of heteroduplex DNA). This mechanism results in the 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.

B, ii. Gap collapse of a DSB close to the 5′ end of the progenitor allele by lateral diffusion of the minisatellite free 3′ end. **iii.** Out of alignment strand invasion of the minisatellite free 3′ end into the homologous chromosome. **iv.** D-loop expansion promoted by DNA synthesis and annealing of the second free 3′ end, in register. Repair synthesis and ligation form a double Holliday junction structure. **v.** Resolution yields conversion products, only the converted recipient product is shown (including regions of heteroduplex DNA). This mechanism produces a deletion, in this case two repeat units, in the recipient, whereas the other allele does not change length.

This figure was adapted from Monckton, (1993).



Chapter 6 Figure 1

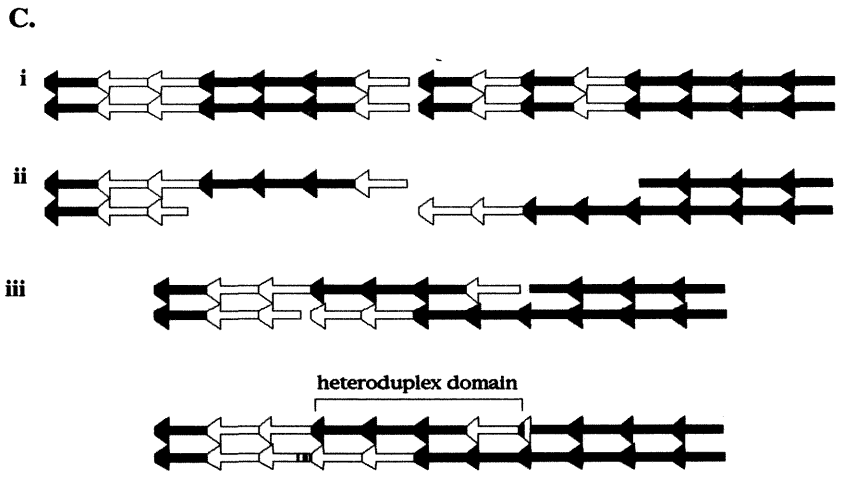


Figure 6.1 Aberrant repair of DSBs occuring within minisatellite tandem repeat arrays.

C. The single stranded annealing model applied to minisatellites.

ii. Tandem repeats anneal. **iii.** DNA synthesis and/or ligation lead to regeneration of duplex tandem repeat DNA accompanied by loss of repeats and generation of a central heteroduplex domain.

These models, although speculative, fit the observed physical consequences of minisatellite mutation at MS32 very well. For example, all of the MS32 mutations observed in pedigrees (Chapter 3, Table 3.2) could be accounted for by gap expansion (Monckton, 1993f), although the presence of extraneous repeats still has no simple explanation. Two events (mutants c and d, Table 3.2) would require interstitial DSBs, but the remaining five events could be accounted for with an initiating DSB located in the first few repeat units. In fact mutants a, b, e and f have obligate first repeat unit breakpoints. Terminal initiating DSBs repaired according to the GEM would be expected to give rise to the observed terminal polarity and size increase bias. Furthermore, both interallelic and intraallelic size gain mutations are accommodated in this model, since it assumes that both proceed through the same pathway, and differ merely in the repair template used. This allows for the insertion of either terminal or interstitial material from the homologous chromosome or sister-chromatid into the terminal position of the recipient, with no loss of information in the recipient and without exchange of flanking markers. In some mutations, sequence analysis across breakpoints in donor, recipient and new mutant alleles may allow more detailed inferences about the mechanisms involved, particularly with respect to the GEM.

The GEM assumes that DSBs occur preferentially at or towards the 5′ end of the tandem repeat array without explaining how this comes about. An attractive hypothesis postulates the existence of a specific initiation sequence, located in the flanking DNA, that directs DSB formation to the beginning of the tandem array of a nearby minisatellite. Comparisons of a few hundred bp of DNA immediately flanking several hypervariable minisatellites have not revealed any obvious sequence elements conserved between them (A.J. Jeffreys, personal communication); however, it is possible that such interactions may occur over much larger distances. It is tempting to speculate that the high rates of mutation at such loci are a result of the proximity of recombination hotspots active in the same regions of the chromosome, and that minisatellites therefore derive from recombination hotspots, rather than as being recombinators in their own right.

This Work. The similarities between MS31A and MS32 in terms of allelic structure and diversity suggested that similar repeat unit turnover processes may be operating at both of these loci. The study of allelic diversity at MS31A (Chapter 5) had already provided circumstantial evidence for a mutation hotspot at one end of MS31A alleles. To see if this was the case I used MVR-PCR to analyse several MS31A *de novo* mutation events and determine possible mechanisms by which they may have occured. The haplotypes of flanking polymorphisms of mutant and progenitor alleles were also determined to find out whether there was evidence for the involvement of simple unequal recombination. Once these mutations had been characterised, they were compared with those seen at MS32, to assess the extent to which the GEM, which had been proposed to account for the generation of most allelic diversity at MS32, also fitted mutation events found at MS31A. Blood and DNA samples from children scored as having mutant MS31A alleles and their parents were kindly provided by Prof. Jean Dausset (CEPH), Dr. Mike Webb (Cellmark Diagnostics, Abingdon, UK) and Dr. Lotte Henke (Instit für Blutgruppenforschung, Dusseldorf, Germany). Analysis of the somatic mutation mosaic described was carried out by Ila Patel in collaboration with Dr. Lotte Henke and myself. Some of the work presented in this chapter has been published (Jeffreys *et al.*, 1994). A further comparison of mutation events at these, and other minisatellite loci, is made in Chapter 7, where tandem repeat mutation processes are discussed further and their implications for these loci are considered.

Results

1. Germline mutation at MS31

Detection of *de novo* length change mutations at MS31. Southern blot allele length analysis of MS31 alleles segregating in pedigrees was performed on *Hinf*I digests of DNA from 40 of the CEPH panel of large families (Jeffreys *et al.*, 1988), as well a from large numbers of families in immigration cases and from mother, father and child trios in paternity cases, in laboratories which routinely use MS31 as a probe (M. Webb & L. Henke, personal communications). In some of these families mutant MS31 alleles, with different length than either parental allele, were inherited along with an apparently non-mutant allele from the other parent (data not shown, see Fig. 6.2 for examples of PCR confirmation of mutants). Correct parentage was established beyond doubt in all these families, by typing with other markers (data not shown). A total of thirty four *de novo* mutation events were detected in this way; their characteristics are summarised in Table 6.1. Eight of these occurred in CEPH families, five in paternity cases conducted by Cellmark Diagnostics and twenty one in paternity cases from the Institut für Blutgruppenforschung in Germany.

MS31 germline mutation rates detectable by Southern blot analysis. Cellmark have no records of MS31 mutations found in paternity cases prior to the start of this study in 1992, nor of exact numbers of paternity tests conducted; therefore the 5 mutations they found were not used in calculations of mutation rate and this analysis was restricted to the remaining 29 mutation events. The children of the 40 CEPH families and of the trios from Germany represent 3869 offspring scored for mutation (684 and 3185 respectively). This gives a Southern blot detectable mutation rate of 29/3869 = 0.8% per gamete for MS31. However, this figure is misleading because the distribution of MS31 mutations is markedly biased toward the male germline (Henke *et al.*, 1993). There was no information on the parental origin of two of the 29 mutants, but of the remaining 27, 21 were present in 1805 paternally derived alleles, giving male germline mutation rate of 1.2% per gamete. All 5 of the Cellmark mutants not included in this calculation were also of paternal origin. The 6 female germline mutants were sampled from 2064 maternal alleles giving a lower female germline mutation rate of 0.3% per gamete. Although a few (mostly maternal) mutation events appeared to involve large (>1kb) deletions, most (generally paternal) were small size changes making it difficult to deduce whether they involved gain or loss of repeat units, since they fell between upper and lower parental alleles.

Southern blot analysis of mutation events. The size and nature of mutation events in the CEPH families, with respect to gain or loss of repeat units, were more accurately determined from Southern blots of *Hae*III digests of DNA from members of these families. *Hae*III cuts occasional MS31A repeats with a particular sequence variant, to generate a diagnostic set of fragments from alleles containing such repeats (data not shown). The progenitor of each mutant allele was identified by comparing these profiles, and RFLPs in the *Hae*III profile of a mutant allele were used to estimate the difference in size from its progenitor, and hence the number of repeat units involved in the mutation event (data not shown). One of the 8 mutant CEPH alleles, a paternal deletion of ~2 repeat units (mutant 31, Table 6.1), was fortuitously discovered as a result of this analysis. It had not been detected in the original CEPH *Hinf*I screen (Jeffreys *et al.*, 1988) because the size change involved was too small to resolve; however, it was present in the same pedigree as a mutant which had been scored and was revealed as a variant in the *Hae*III profile from this family (data not shown).

Figure 6.2. Confirmation of MS31A mutant alleles by PCR amplification and Southern blot detection.

MS31A alleles were PCR amplified between flanking primers 31A and 31B for 20 cycles with annealing at 68°C and extension for 5 minutes. PCR products were resolved on a 0.7% agarose gel, which was then Southern blotted. The filter was hybridised with MS31 and autoradiographed. Lanes 1, 2 and 3, and 4, 5 and 6, are mother, child, father trios. In both examples the maternal allele is the same size in mother and child, but the paternally inherited allele is of a different size to either of those in the father. True paternity was established by hybridisation with other hypervariable probes (not shown), therefore the paternally derived alleles are mutant versions of a paternal allele. Only one allele is seen in lane 1, presumably because the other allele was too large (>8kb) to be efficiently amplified.

MVR-PCR analysis of mutation events. In order to characterise mutation events in more detail, and gain an insight into the mechanisms by which they may have occured, I compared the MVR structures of mutant alleles with those of their progenitors, and also those of the non-mutant allele present in the parent from which these mutants were derived. Single-allele MVR codes were generated using either size-separated alleles or, in individuals heterozygous for flanking polymorphisms, allele-specific MVR-PCR. Initially I used primers corresponding to all 4 combinations of the 2 polymorphic positions in the MS31A repeat unit, to maximise the amount of MVR information obtainable from these alleles and hence increase the precision with which mutation events could be defined (see Fig. 4.1 for repeat unit sequence; Table 4.1 for primer sequences). MVR reactions using each of the primers 31-TAG-AC, 31-TAG-GC, 31-TAG-AT and 31-TAG-GT were loaded in 4 adjacent lanes on an MVR gel, giving a 4-state, rather than 2-state, single allele code following hybridisation and autoradiography (Fig. 6.3A). To score these codes the a-type and t-type repeat units scored in two state mapping were divided into two subclasses and renamed (for ease of reading in figures), to give; e, E, y and Y-type repeats (see Tamaki *et al.*, 1993 for description of MS32 4-state mapping). e-type repeat units are equivalent to a-type repeats in 2-state mapping, they are detected by 31-TAG-AC and start with GT; y-type repeats are equivalent to t-type repeats in 2-state mapping, they are detected by 31-TAG-GC and start with GC; E-type repeat units are detected by 31-TAG-AT and start AT, and Y-type repeat units are detected by 31-TAG-GT and start AC. The first 4-state mapping experiments showed that Y-type repeats were rare, only one was detected in 4-state MVR maps from the first 13 alleles mapped (~900 repeat units). For this reason only e, E, and y-type repeat units were mapped in subsequent experiments (Fig. 6.3B), thus increasing the throughput of samples on MVR gels without significant loss of MVR code information. As with 2-state mapping, 3-state MVR maps contained infrequent O-type positions, presumably due to the presence of additional sequence variants and perhaps the occasional Y-type repeat. The MVR codes of 16 mutant alleles, along with their progenitors and the non-mutant allele from the same parent were determined in this way (see Table 6.1. & 12 examples in Fig. 6.4). This analysis showed that there are distinct classes of mutation events, which may be due to the operation of different mutational mechanisms at this locus.

A. Male Germline. 8 out of 9 paternal mutations investigated by MVR mapping (mutant alleles 1-8, Table 6.1; Fig. 6.4) involved the incorporation of a small number of repeat units at or near the 5′ end of the tandem repeat array. In the cases where Southern blot length analysis had already defined the direction of mutation and the number of repeats involved, MVR mapping confirmed that this was due to a single event at the mapped end of the minisatellite (data not shown). This indicated that the observed mutational polarity was not an artifact of MVR mapping, which is directed to this end of the array and therefore biased to the detection of mutation events here. In the remaining paternal germline mutant (mutant 9, Fig. 6.4) MVR mapping detected no difference between progenitor and mutant allele codes (data not shown). It is possible that this mutation was misscored, for example because of a bandshift on the original gel. However, judging by the relatively large size change involved (+~60 or -~10 repeats, depending on which paternal allele was the progenitor), it is more likely that this mutation either involved a small deletion of the larger allele, too far into the allele to MVR map, or was due to a reduplication of the beginning of the smaller allele that extended beyond the MVR mapped region. These possibilities might be distinguished by *Hae*III mapping of these alleles, or perhaps, extended or reverse MVR mapping.

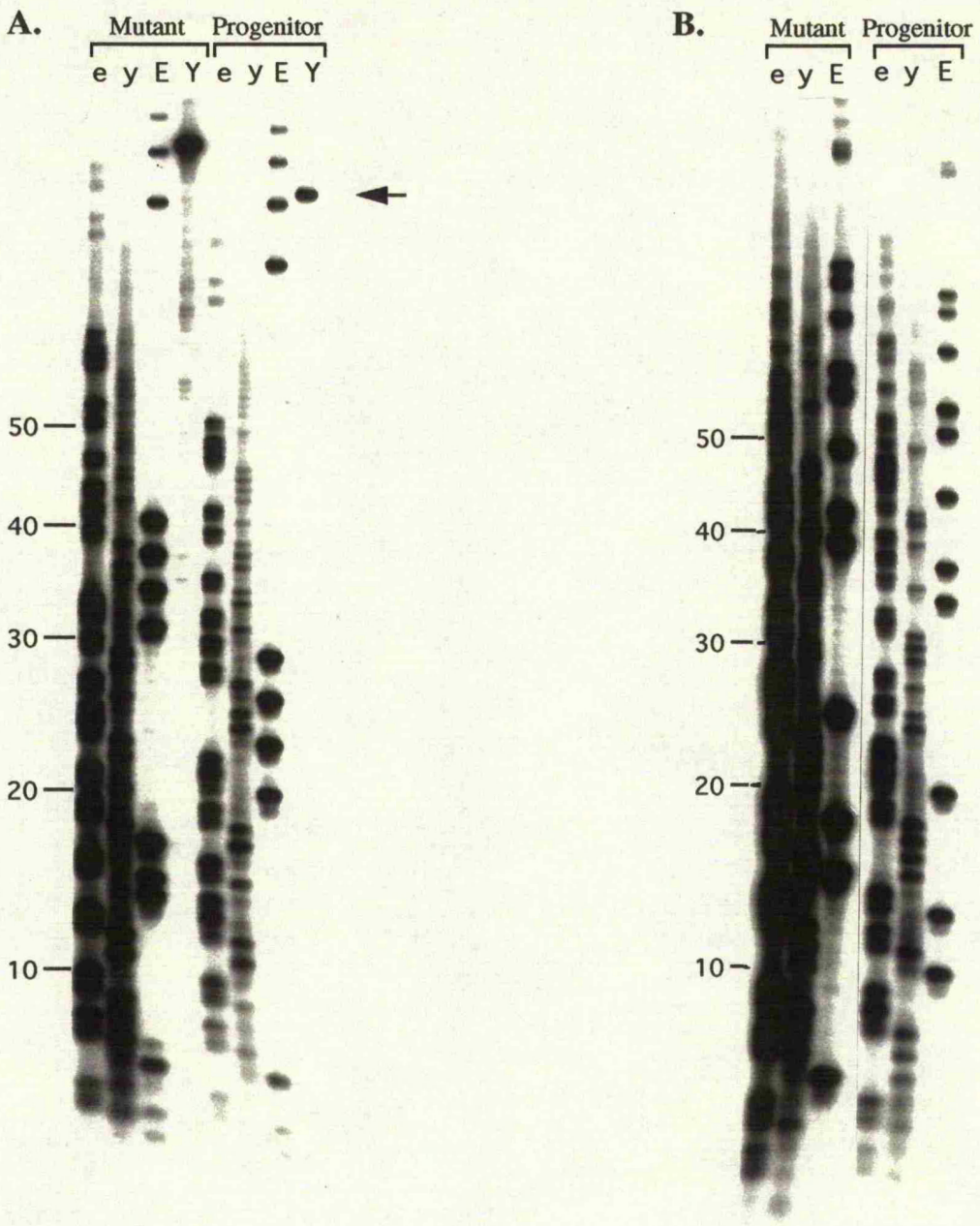


Figure 6.3. MVR maps of mutant MS31A alleles and their progenitors.

100ng genomic DNA from children and the parents from whom they had inherited a mutant allele, as identified by Southern blot analysis, was subjected to MVR-PCR (see Materials and Methods). Allele-specific flanking primer Hga1- was used together with MVR-specific primers 31-TAG-AC (e), 31-TAG-GC (y), 31-TAG-AT (E) or 31-TAG-GT (Y) (see Table 4.1 for sequences).

A. 4-state mapping of mutant 7 (Table 6.1) and its progenitor. Arrow indicates the position of a rare **Y**-type repeat unit in the progenitor. The mutation is a gain of 9 repeat units. MVR codes of progenitor and mutant alleles are shown in Fig. 6.4

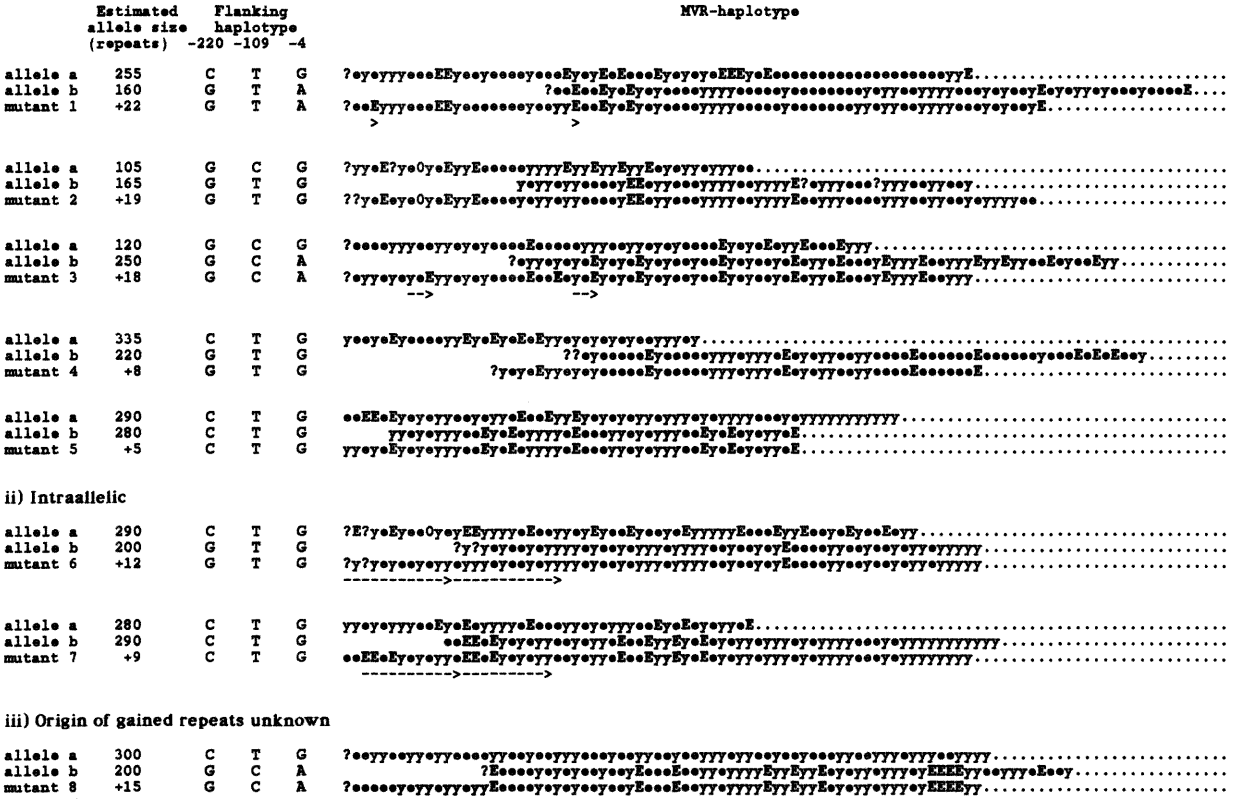
B. 3-state mapping of mutant 5 (Table 6.1) and its progenitor. This mutation is a gain of 5 repeat units, as shown in Fig. 6.4.

Figure 6.4. MVR analysis of mutant MS31A alleles detected in pedigrees.

Allele-specific, 3-state MVR-PCR maps (e-,y- and Y-type repeats plus O-type repeats, (see Materials and Methods & Chapter 6) were generated from each parental allele (a and b) and the mutant allele (numbered according to Table 6.1), mutant 34 was 2-state mapped (a- and t-type repeats) from size-separated alleles. Repeat copy number (rounded to the nearest 5 repeats), estimated from allele length analysis of *Hinf*1 digests of genomic DNA, or PCR amplification of whole alleles, and haplotypes of 5′ flanking polymorphic positions (-4A/G, -109C/T and -221G/C), determined as described in Chapter 5 are shown. For each mutation event (**A.** paternal, **B.** maternal, **C.** somatic) the progenitor allele (b, defined as that contributing the major proportion of the mutant) and derived mutant are aligned and shown in red. The non-progenitor allele (a, contributing a minor proportion of the mutant allele, or apparently not involved in the mutation event) is shown in black. Regions donated from non-progenitor to progenitor alleles in interallelic exchanges are shown in green, reduplicated repeats are arrowed below, and deleted regions are indicated by a "-". Repeats of unknown origin are shown in blue. (.....) Indicates allele continues beyond mapped region. Parental alleles were not mapped from mutant 34, making it impossible to distinguish between duplication and deletion.

A. Paternal

i) Interallelic



ii) Intraallelic

iii) Origin of gained repeats unknown

Continued over

Figure 6.4. MVR analysis of mutant MS31A alleles detected in pedigrees

B. Maternal

i) Intraallelic

C. Somatic

i) Intraallelic

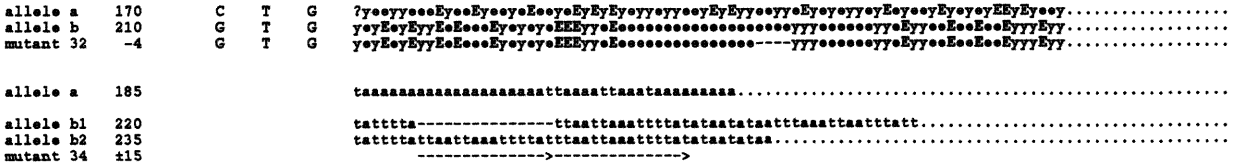


Table 6.1 Summary of MS31A mutants detected and characterised to date.

Mutant	Parental origin (progenitor)	Tissue of origin (progenitor)	Size change (repeat units)	MVR mapped ?	Distance of 5´exchange point from 1st repeat (repeat units)		Putative mutation mechanism
					progenitor	donor	
1	paternal	germline	+22	3-state	4*	4	complex interallelic conversion
2	paternal	germline	+19	3-state	1	1	interallelic conversion
3	paternal	germline	+18	3-state	10*	10	complex interallelic conversion
4	paternal	germline	+8	3-state	4	20	interallelic conversion
5	paternal	germline	+5	3-state	5	5	interallelic conversion
6	paternal	germline	+12	3-state	1	1	intraallelic conversion/slippage
7	paternal	germline	+9	4-state	2	2	intraallelic conversion/slippage
8	paternal	germline	+15	3-state	1	?	unknown/terminal
9	paternal	germline	+~60/-~10†	3-state	?	?	?
10-25	paternal	germline	nd	nd	?	?	?
26	maternal	germline	-52	3-state	2	na	terminal deletion
27	maternal	germline	-~100	2-state	>70	na	internal deletion
28-30	maternal	germline	nd	nd	?	?	?
31	paternal	germline/somatic	-2#	nd	?	?	?
32	maternal	somatic	-4	3-state	29-48	na	internal deletion
33	unknown	germline/somatic	+/-~150	2-state	>70	na	internal deletion/ terminal reduplication
34	unknown	somatic (mosaic)	+15	2-state	7	na	intraallelic conversion/slippage

nd = not done; ? = not known; na = not applicable

* these alleles show reduplications at the ends of the exchanged fragment (see Figure 6.3)

† depending on which is the progenitor allele

mutation in CEPH lymphoblastoid cell line may be somatic or germline.

The paternal germline repeat unit gain mutations fall into two categories; in five mutants (Fig. 6.4. mutants 1-5.) there was clear evidence of interallelic conversion, with mutation involving the insertion of a donor segment at, or near, the beginning of the recipient allele with no loss of information from the recipient. The insertion site was at most 10 repeats into the allele (mutant 3) and in three cases (mutants 2, 6, & 8) was at the first repeat unit. In most interallelic exchanges where the donor segment could be identified (mutants 1, 2, 3 & 5), this was also at, or close to, the beginning of the donor allele. In fact in five cases (mutants 1, 2, 3, 5 & 6) there was perfect alignment between the beginning of donor and recipient alleles with respect to the position of the donated segment of repeats, with only one example (mutant 4) of exchange between misaligned alleles. In two mutants, (1 & 3) anomalous repeats were present in the donated segment and there was evidence of insertion site reduplication. Mutant 1 also showed an internal switch in internal repeat unit type accompanying the mutation event. Examples of intraallelic mutation were also found (mutants 6 & 7). These had arisen by terminal duplication of a segment of repeats at or near the beginning of the mutating allele. The remaining paternal mutation (mutants 8) showed a 5′ extension of repeats of unknown origin at the start of the progenitor allele.

B. Female germline. By contrast, the three of the six maternal mutations which were analysed (26, 27 and 32, Table 6.1) were all shown to be deletions by Southern blot length analysis. Two of these involved the loss of large numbers of repeat units (>1kb, 50 repeats), while the third was predicted from *Hae*III profiles to be a 4 repeat unit deletion. Comparison of MVR maps from these alleles and their progenitors showed that one of the two large deletions (Fig. 6.4B, mutant 26) was of 52 repeats from the 5′ end of the allele, the second (Table 6.1, mutant 27) showed no difference between progenitor and mutant alleles was therefore presumed to be internal to the MVR mapped region.

2. Somatic mutation at MS31

Identification of somatic mutation events by MVR analysis. Allele-specific MVR mapping showed that the four repeat deletion (Table 6.1, mutant 32) was actually a somatic event. Instead of a giving a single-allele code as expected, the DNA from the child gave a diploid code when amplified with an allele-specific primer, revealing the presence of both the maternal progenitor allele and the mutant allele that had been identified by Southern blot analysis (Fig. 6.5). The progenitor was present at a lower level than the mutant which, combined with the small size change involved, explains why it was not detected as a third band on Southern blot hybridisation. This reduces the estimated female germline mutation rate to 0.2% per gamete.

Identification of somatic mutation events by Southern blot analysis. Two further putative examples of somatic mutation were identified by the presence of 3 minisatellite hybridising bands on Southern blots of *Hinf*I digested genomic DNA derived from blood cells. However, there are several alternative explanations for such a 3 band pattern. These are; contamination, the presence of a *Hinf*I site in an internal minisatellite repeat unit (eg. caused by a rare sequence variant), partial blood cell clonality for cells carrying the mutant allele (eg. resulting from a haematological tumor), chimaerism (the fusion of 2 separate fertilised eggs to form 1 embryo), or genuine somatic mosaicism caused by a mutation event in the early embryo. DNA from both of the "three-band" individuals was analysed further to distinguish these possibilities.

A.

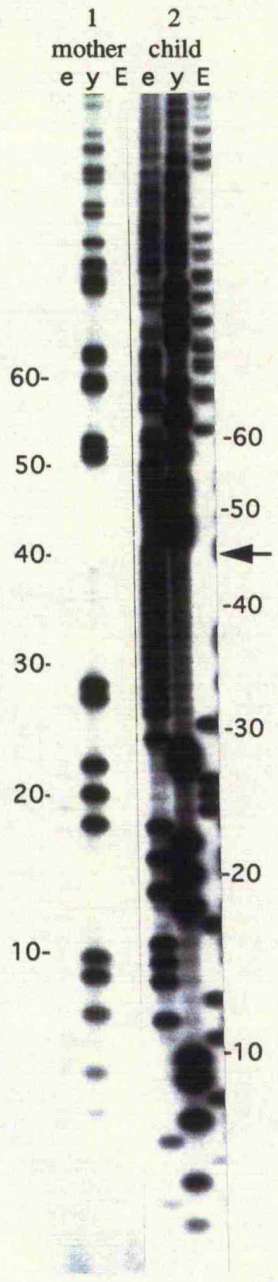


Figure 6.5. MVR mapping of a presumed germline mutant and its progenitor.

A. Allele-specific MVR-PCR was performed on genomic DNA from a mother (1) and her child (2) using primers 31HgaI+ and 31Psp1406I+ respectively, to generate 3-state single allele MVR maps of mutant 32 (2) and its progenitor (1). Arrow indicates position of transition from single allele to diploid code.

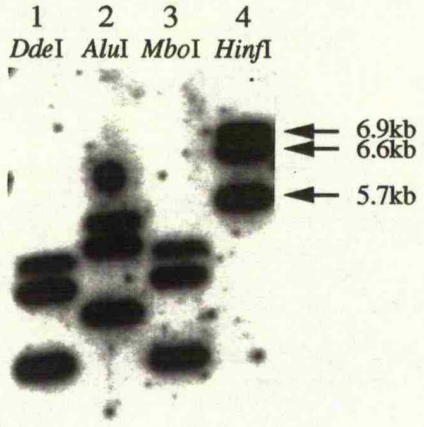
B. Allele codes derived from MVR maps of 1 and 2. The single allele code from the progenitor (red) was compared to the partial single allele code (red) and the diploid code (red and black, from above the arrowed position) to deduce the structure of the mutant allele shown below.

```
                    10        20        30        40        50        60        70
                    .         .         .         .         .         .         .
allele 1  yeyEeyEyyEeEeeeEyeyeyeEEEyyeEeeeeeeeeeeeeeeeeeeeeeyyyeeeeeeeyyeEyyeeEeeEeeE...
allele 2  yeyEeyEyyEeEeeeEyeyeyeEEEyyeEeeeeeeeeeeeeeeeyyyeeeeeeeyyeEyyeeEeeEeeEyyyE...
                                                        eeeeyyyeeeeeeeyyeEyyeeEeeEeeE...

mutant 32 yeyEeyEyyEeEeeeEyeyeyeEEEyyeEeeeeeeeeeeeeeeee----yyyeeeeeeeyyeEyyeeEeeEeeE...
```

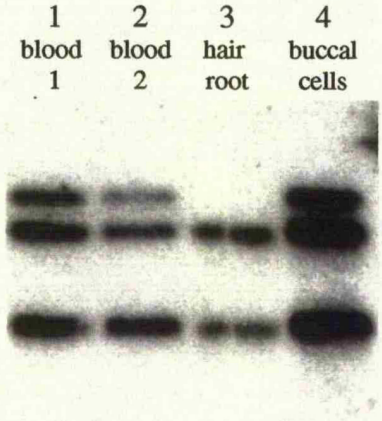
Figure 6.6. Southern blot analysis of DNA from somatic mutant 34.

A.



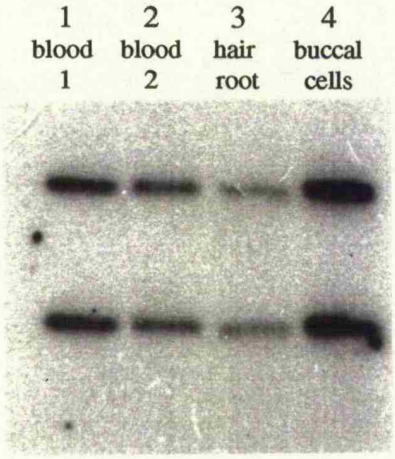
A. DNA from this individual was extracted from leukocytes and digested with the enzymes shown. Digestion products were resolved by agarose gel electrophoresis. Southern blot hybridisation with MS31 followed by autoradiography reveals the presence of three alleles in all digests.

B.



B. DNA from two different blood samples (1 and 2), hair roots, (3) and buccal cells (4) was digested with *AluI*. Digestion products were resolved by agarose gel electrophoresis. Southern blot hybridisation with MS31 followed by autoradiography reveals the three band pattern in DNA from blood and buccal cells, but not hair roots.

C.



C. The filter from B. was stripped and rehybridised with MS32, all tissues give a two band pattern, as would be expected from a normal heterozygote.

Data kindly provided by Mrs I. Patel.

A.

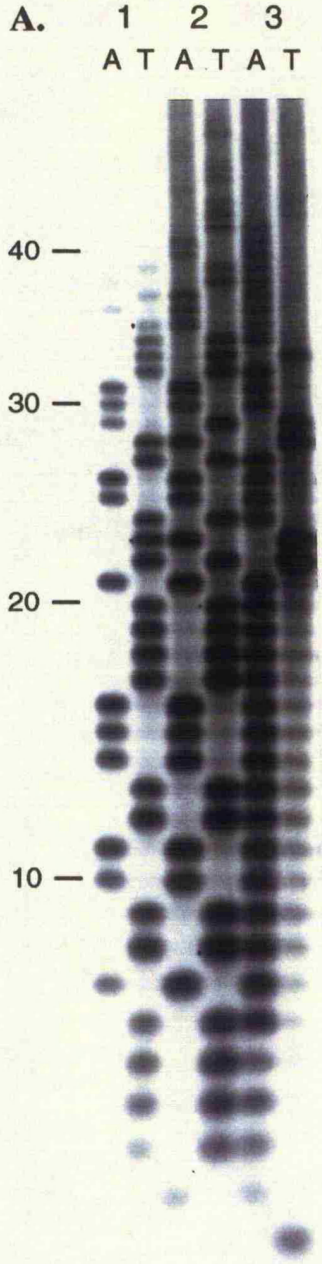


Figure 6.7. MVR maps of 3 alleles from mutant 34.

A. The three alleles 1, 2 and 3 from somatic mutant 34, (Table 6.1) were size separated by preparative gel electrophoresis of size fractions from an *Mbo*I digest of genomic DNA and 2-state MVR mapped, as described previously (Materials and Methods & Chapter 4).

Data kindly provided by Mrs I. Patel.

B.

```
allele 1a tatttta---------------ttaattaaattttatataatataa.....

allele 1b tattttattaattaaattttatttaattaaattttatataatataa.....
                  -------------->-------------->

allele 2 taaaaaaaaaaaaaaaaaaattaaaattaaataaaaaaaaat.........
```

B. 2-state single allele codes of alleles 1, 2 and 3. Alleles 1 and 2 are versions of the same allele (1a and 1b) related by a reduplication (arrows indicate repeated region) or a deletion (indicated by ---). ..... indicates allele continues beyond mapped region. Related segments of these alleles are highlighted in red. The single allele code of the other allele in this individual (allele 2) is also shown.

A. Mosaicism detected in lymphoblastiod cell line DNA. One of the DNA samples giving three bands on a Southern blot was from a lymphoblastoid cell line, cultured from the father of one of the CEPH pedigrees. Digestion with different enzymes and hybridisation with other probes showed that this was not due to contamination, an internal *Hinf*I site or chimaerism (data not shown). Two of the MS31 alleles from this individual appeared to segregate normally among his children, however the third allele, which was ~2kb larger than either of these, was not found in any of the 6 children from this family. MVR maps from this larger allele and one of those transmitted to the children were identical (data not shown). In the absence of DNA from the parents of this individual it was not possible to determine whether the larger allele represented a somatic reduplication of the beginning its progenitor, that arose during cell culture, or whether this individual was a mosaic for a reduplicated/deleted allele, only the shorter form of which was transmitted through the germline.

B. Mosaicism detected in DNA extracted directly from different tissues. The second of the two three-band individuals was investigated in greater detail. Digestion with different enzymes excluded the possibility an internal *Hinf*I site (Fig. 6.6A). DNAs extracted from two independent blood samples, taken ~1 year apart, both showed the same 3 band pattern on hybridisation with MS31 (Fig. 6.6B), while other probes give the expected two band pattern (Fig. 6.6C) excluding contamination or chimaerism. A clinical blood test showed no abnormalities such as the presence of leukaemia or a tumor (data not shown). As well as the two blood samples, the mutant allele was found in the DNA extracted from buccal cells (Fig. 6.6B), from urine (data not shown) and from hair roots (data not shown). However, Southern blot analysis of DNA from hair roots taken from different parts of the body showed that the the mutant allele was present in some hair roots, but not others (eg. Fig. 6.6B). This provides strong evidence for a somatic MS31 mutation during embryogenesis that has resulted in a mosaic of cells, some with and some without the mutant allele. The three alleles present in DNA prepared from blood samples from this individual were physically separated from each other, by preparative gel electrophoresis, and then MVR mapped using the 2-state mapping system (Fig. 6.7A). The mutation is a 15 repeat reduplication/deletion close to the start of the tandem repeat array (Fig. 6.7B). The estimated sizes of the upper two bands concur with this size change (Fig. 6.6B), suggesting that one is derived from the other. Although DNA from the parents of this individual is not available, it is probable that this mutation involved a gain in size, since a simple intraallelic reduplication is more likely than the independent duplication and subsequent deletion of the same stretch of repeat units. However, it is not possible to unequivocally determine whether this mutation involved the gain or loss of repeat units.

Discussion

Detection of MS31A mutants. By definition, all of the mutation events analysed at MS31A were large enough (≥ 2 repeat units) to be detected by Southern blot analysis. At MS32 comparison of diploid MVR codes from families made it possible to derive single allele codes from the ternary codes of parents and their offspring, revealing mutation events as multiple parental exclusions. Some mutations detected in this manner had not been scored by length analysis because the size change involved was too small to distinguish mutant and progenitor alleles (Jeffreys *et al.*, 1991a; Chapter 3). This approach to the detection of *de novo* mutations was not possible at MS31A, due to the difficulty in accurately scoring diploid codes (Chapter 4). It is therefore possible that small mutations (≤2 repeat units), of the sort revealed in the female germline by MVR analysis at MS32, may have been missed at MS31A, and that as a consequence the MS31A germline mutation rate detectable by length analysis is an

underestimate. MVR-PCR mapping could be used to detect such mutation events in pedigrees with flanking heterozygous positions by using allele-specific MVR-PCR to generate single allele codes from all 4 parental alleles and comparing these with single allele codes from the children. It would also be possible to screen children uninformative at flanking polymorphic positions, provided the segregation of informative parental alleles was known, by comparing predicted diploid codes assembled from the parental single-allele codes with actual diploid codes from these children, to look for discrepancies caused by small gains or losses of repeats. Although these approaches would be rather laborious, the data from MS32, where several additional mutants were revealed by MVR mapping, suggest that such an investigation might be rewarding.

Comparison with MS32. As with MS32, MVR-PCR has enabled a much more detailed analysis of mutation at MS31A than it was possible to make using Southern blot length analysis. The most significant features of the characterised mutation events at MS31A were the bias towards mutation in the male germline and the qualitative difference between *de novo* mutations in the male and female germline. All but one of the paternal mutants involved gains of small numbers of repeat units and showed extreme polarity with respect to the progenitor allele and usually the donor allele. The observed 5′ polarity of *de novo* mutation at this locus confirms the hypothesis that the gradient of interallelic variation within groups of aligned alleles at MS31A is caused by the action of a mutation hotspot localised to this end of the tandem repeat array. The similarities between MS31A and MS32 in this respect are extremely striking. Both have a mutation hotspot located at one end of the minisatellite and at both loci mutations in the paternal germline usually involve the gain of small numbers of repeat units at, or near, this end (Jeffreys *et al.*, 1994). The mechanisms of these repeat unit gain mutations also appear to be the same with examples of intraallelic reduplications, reminiscent of USCE or replication slippage, and interallelic exchanges, which are sometimes complex. These frequently involve only short regions of the donor allele, inserted into a highly active region in the recipient, by localised conversion processes. The sizes of the alleles involved show that either the smaller or the larger of a pair of alleles can act as donor or recipient.

There are no examples of exchange of flanking markers accompanying any of interallelic exchanges at either locus, strongly suggesting that these are confined to the repeats, although displacement of a D-loop by the invading 3′ end beyond the minisatellite repeats and into the flanking DNA may be the cause of the occasional switches in flanking haplotype observed among groups of aligned alleles at both loci (Monckton, 1993c; Chapter 5). There are several examples of apparently simple conversion at MS31A (Fig. 6.4, mutants 2, 4 & 5.) although sometimes, as at MS32, interallelic exchange is accompanied by target site reduplication and the appearance of a short region of anomalous repeats, not corresponding to the cognate position in either the donor or recipient alleles, at, or close to, the junction between the contributions of donor and recipient (Jeffreys *et al.*, 1994; Fig. 6.4, mutants 1 & 3). These unattributable repeats may reflect the involvement of mismatch repair of heteroduplexes within the conversion complex during the generation of the mutant structure, but in most cases are complex and have no obvious origin. Most simple conversion events at MS32 and MS31A involve the transfer of repeat units from positions at, or near, the beginning of the donor allele into the equivalent position with respect to the beginning of the mutating recipient allele, for example, mutant 1 (Fig. 6.4) has acquired donor repeats 5-23 inserted immediately downstream of repeat 4 in the recipient allele. This registration suggests that the alleles are usually aligned over the 5′ flanking region during strand invasion and suggests that interallelic exchanges may occur between the alleles of synapsed chromosomes during meiosis. For sister chromatid conversion, this will result in duplication of a terminal repeat

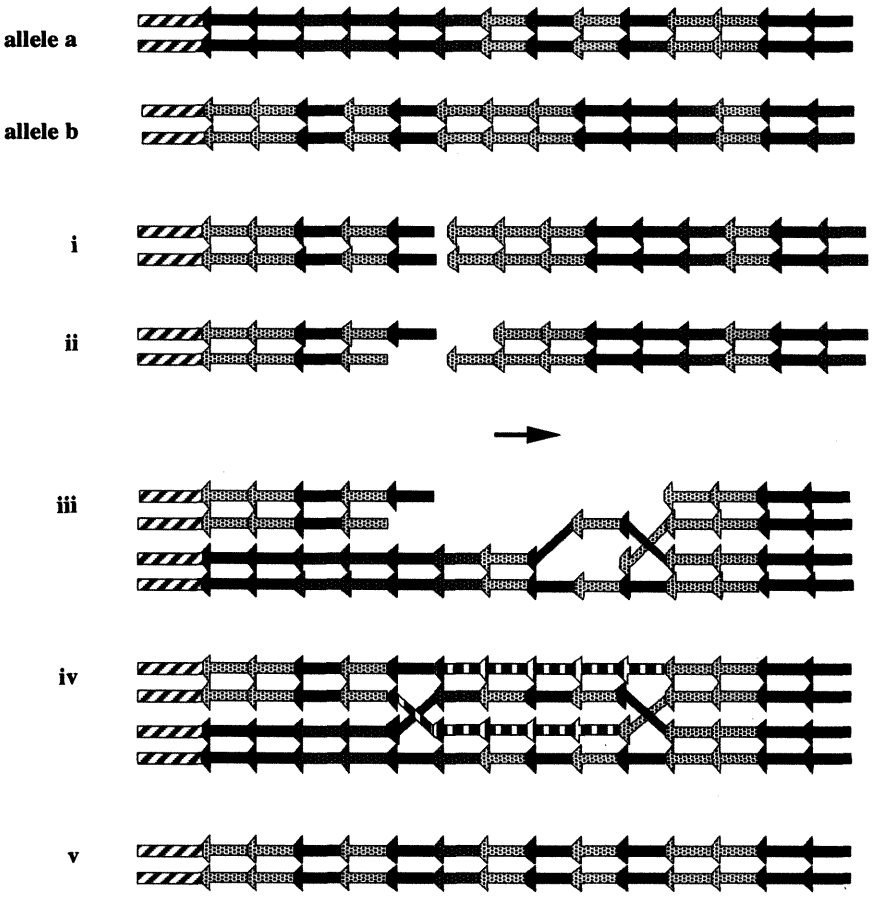


Figure 6.8. Application of the gap expansion model to MS31A mutant 5.

The progenitor allele is shown in red, the donor in black and exchanged section of repeats in green, according to Figure 6.3. Diagonal stripes = 5′ flanking DNA. Arrows represent minisatellite repeats. Filled = e-type repeats; light shading = y-type repeats; dark shading = E-type repeats. Vertical stripes indicate newly synthesised DNA.

i. Initiation by a DSB in the fifth repeat of allele b. **ii.** Generation of free 3′ overhangs by exonuclease activity. **iii.** Gap expansion and out of alignment strand invasion of the minisatellite free 3′ end into the homologous chromosome. **iv.** 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. **v.** Resolution to yield the observed conversion mutant.

segment in the mutating allele, for example mutant 6 (Fig. 6.4). However, it is clear at both loci, particularly MS32 that mutations are sometimes more complex and may simultaneously involve different processes, as shown by the presence of insertion site reduplications and anomalous repeats in converted segments (eg. mutants 1 & 3, Fig. 6.4). The origin of the repeat units gained by mutant 8 was not apparent, suggesting either repeat scrambling during mutation or alternatively the interallelic or intraallelic acquisition of repeats from a site distal to the region mapped by MVR-PCR, which would suggest large misalignments with respect to the 5′ end of the tandem repeat array.

Compatability of size gain mutations with the GEM. The mutation data obtained from MS31A can be most simply explained by invoking the same model of gap expansion that was proposed to explain length gain mutations at MS32. Some of the mutations seen at MS31A fit this model, if anything, even better than those at MS32, (Fig. 6.8) which serves both to strengthen the evidence supporting this hypothesis and also to suggest that the process may be a general mutational mechanism by which hypervariability is generated at different human minisatellite loci. Applying the gap expansion model to mutant 5 (Fig. 6.8) shows that four repeat units at the free 3′ end are aligned with repeat units of the same type in the donor allele, suggesting that, in this case at least, the size of the gap expansion may have been determined by alignment of the free 3′ end of the DSB with the nearest downstream region of homology in the other allele. However, this may just be a coincidence since none of the other interallelic exchanges show evidence of a similar homology search. The extreme polarity of length gain mutations seen at both loci implies that mutation is in some way modulated by element(s) outside the array and the transfer of repeat units implies that the recipient allele has in some way to be activated for mutation prior to exchange. This may due to the presence of a locally acting element(s) in *cis*, for example an initiation site for recombination, that acts as a mutation initiator element by activating an allele for mutation, or be a reflection of longer range polarity, such as orientation within the chromosome as a whole. Both of these theories suggest that minisatellites are coincidental with, rather than equivalent to, recombination hotspots.

Bias toward mutation in the male germline. The marked bias towards mutation in the male germline shown by MS31A (Henke *et al.*, 1993) has also been observed at several other hypervariable minisatellites in standard use in paternity testing; these are MS205 (Jeffreys *et al.*, 1994), MS43A (D12S11) and pλg3 (D7S22) (L. Henke, personal communication). This phenomenon has also been observed at the human loci, MS1 (Olaisen *et al.*, 1993) and CEB1 (D2S90), which shows an extreme bias (Vergnaud *et al.*, 1991), and at the mouse minisatellite locus Ms6-hm (Kelly *et al.*, 1989). This sex-specific pattern of mutation may result from the greater number of germline mitoses in the male, indicating a predominantly mitotic mutation process. On the other hand it could be connected with elevated male-specific rates of recombination in subtelomeric regions, such as have have been inferred from the observed expansion of these regions in the male linkage map (Nakamura *et al.*, 1988b, 1989; Weber *et al.*, 1993). The numbers of mutation events so far detected in pedigrees at MS32 is relatively small, since this locus is not used in paternity testing. It would be intriguing to know whether there is a similar sex bias to the loci mentioned above at MS32, or whether the small, but roughly equal, numbers of male and female germline mutation events observed are a true reflection of the respective sex-specific mutation rates at this locus. The fact that the maternal mutations observed were all small size increases suggests that the latter may be the case, although it is also possible that such mutations are somatic in origin, due to instability of this locus in lymphoblastiod cell lines.

Deletion mutants. Both of the two (out of five) female germline *de novo* MS31A mutants characterised were deletions involving large numbers (>50) of repeat units. These appeared to be more similar in nature to the size selected MS32 deletion mutants PCR amplified from single molecules of blood and sperm DNA (Jeffreys *et al.*, 1990), than to MS32 mutations detected in pedigrees by MVR-PCR. At MS32 deletions tend to have an endpoint located towards the beginning of the tandem repeat array, but this polarity is not as pronounced as that observed for size increase mutations (Jeffreys *et al.*, 1990, 1994). Although one of the MS31A deletions was polar (mutant 25, Fig. 6.4), with only two examples of such events at MS31A it is not possible attach any significance to their location. No paternal deletion mutants were found at MS31A, however this may be a reflection of the small number of mutants analysed and a lower level of deletions, rather than their absence in the male germline. Investigation of large numbers of mutation events detected in sperm DNA at MS32 suggests that deletions in the male germline occur at lower frequency (26%) than size gain mutations (74%) (Jeffreys *et al.*, 1994). The development of single molecule PCR techniques at MS31A should enable a more thorough survey of male germline mutation with respect to deletions to be made.

Somatic mutation. Somatic mutants of human minisatellite alleles have previously been detected by single molecule amplification of size fractionated blood DNA (Jeffreys *et al.*, 1990) and by Southern blot length analysis of DNA recovered from lymphoblastoid cell lines and clonal tumor cells (Armour *et al.*, 1989b). The three MS31A somatic mutants identified in this study were all intraallelic in origin and two (mutants 32 & 34, Table 6.1) were displaced toward the 5′ end, as shown by MVR mapping, suggesting that these may be generated by similar deletion/reduplication mechanisms to those already known to occur at MS32 in blood DNA (Jeffreys *et al.*, 1990, 1994). The two somatic MS31A mutants (mutants 32 and 33, Table 6.1) detected in DNA extracted from lymphoblastoid cell lines derived from the CEPH panel of families were both atypical. One was a large deletion/duplication in a male and the other was a small deletion in a female. Such mutants may arise at any stage from early development to late in the propagation of the lines *in vitro* and although these possibilities cannot be distinguished in the somatic mutants seen at MS31A, it is possible that these mutants may reflect cell line instability at this locus. At microsatellite loci, a high proportion (upto 40%) of presumptive germline mutants detected in lymphoblastoid cell line DNA were found to be somatic in origin (Weber & Wong, 1993; Banchs *et al.*, 1994). Mutant 31 (Table 6.1) was also detected in lymphoblastoid cell line DNA. It was not MVR mapped and could therefore represent either a genuine germline event, or a somatic event occuring in this individual or in the cell line.

The other somatic mutant (mutant 34, Table 6.1) was present in DNA extracted directly from the blood of the individual concerned, and therefore represents a case of *bona fide* somatic mutation occuring early enough in development to be present at Southern blot detectable level in the adult. Such events are very rare, mutant 34 was the only genuine somatic mutant seen in 3088 people tested with probes for four loci; MS31, MS1, MS43 and pλg3, giving a frequency of somatic mutational mosaicism (detectable by Southern blot analysis) of 0.008% per locus, per individual. This accords with previous calculations which estimated that the incidence of minisatellite somatic mutation is very low, ($<10^{-5}$ per mitosis, Armour *et al.*, 1989b). Mutant 34 is the only known example of somatic mosaicism at a human hypervariable minisatellite locus, and was only identified by virtue of the use of MS31 as a probe in thousands of paternity cases. It seems likely that this mutation involved a size increase, since a simple duplication is more easy to envisage than the precise deletion of an already duplicated segment, and also

implicates only one mutation event. However, the DNA from the parents of this individual was not available, making it impossible to determine absolutely whether the mutation was a deletion or a reduplication. If this mutation event was a gain in size, it would be similar to the small intraallelic reduplications already seen in the male germline and may therefore have arisen in the same way, suggesting that such male germline mutations may also occur during mitosis. The tissue distribution of the mutant allele suggests that this mutation event occured early in embryogenesis, prior to the partitioning of blood (embryonic mesoderm) and epidermal (embryonic ectoderm) cell lineages, which occurs during the third week of development. Intriguingly, some of the hair root cells examined only had two alleles; unfortunately, in the absence of parental DNA, it was not possible to determine whether the version of the somatic mutant allele in hair roots is the mutant or the progenitor. If the mutation was indeed a size increase, then these hair roots contain the progenitor allele only. The simplest explanation for this observation is that this individual is not only a mosaic of two types of cells, but that his skin is also divided into clonal patches, each containing cells of a particular type, such that hair root DNA from different sites was derived from both of the two classes of cells present in his individual. If this were the case it would be possible to map the size of such clonal patches on the skin of this individual, by taking hairs from different sites and determining which contained either two or three alleles. Somatic mosaicism has also been observed, albeit at much higher frequency, at two mouse hypervariable minisatellite loci, Ms6-hm (Kelly *et al.*, 1989) and Hm-2 (Gibbs *et al.*, 1993). These have frequencies of offspring showing 3 or more bands on Southern blots of 3% and 20% respectively. In mice, analysis of allele dosage and tissue distribution of somatic mutants suggest that somatic mutation events preferentially occur very early in embryogenesis probably during the first two cell divisions post-fertilisation (Gibbs *et al.*, 1993). Again these events show no exchange of flanking markers and fall within the size spectrum of mutations seen in the germline, suggesting that they may be caused by similar processes. The observation of high somatic mutation rates of the two mouse minisatellite loci compared with relative somatic stability at MS31A, suggests that somatic mutation mechanisms are locus, and possibly species, dependent.

Somatic mutation at MS31A could in principle lead to problems in the use of this locus for linkage analysis and forensic medicine. Using Southern blot hybridisation it is only possible to detect clonal mutant cells if they make up a significant proportion (1-10%) of the cell population (Armour *et al.*, 1989b). In polyclonal tissues, for example blood, such mutants will be heterogenous in size and hence not detectable by Southern blot analysis of bulk tissue DNA, unless early stem cell mutation has occurred. Only one such mutant (mutant 33, Table 6.1) was detected in lymphoblastoid cell line DNA, therefore somatic mutation does not present a significant problem for linkage analysis using MS31A with such cell lines. PCR amplification is much more sensitive to the presence of additional mutant alleles in a DNA sample from somatic cells, as evidenced by the detection of deletion mutants in somatic and germline samples using single molecule PCR (Jeffreys *et al.*, 1990). If forensic analysis by minisatellite amplification using PCR was used on very small samples approaching the single molecule level (Jeffreys *et al.*, 1988b), somatic mutation could lead to the erroneous exclusion of a true association between a suspect and a forensic specimen. However, forensic analysis using PCR does not usually proceed unless considerably larger DNA samples than this are available. Although MVR-PCR is also more sensitive to somatic mutations, as shown by the detection of a somatic mutant not identified by Southern blot analysis, the presence of a third mutant allele is easily diagnosed by the generation of a diploid code from an allele-specific flanking primer. If the single allele MVR code of the parental progenitor allele from which the somatic mutant was derived is known, the allele code of such a somatic mutant can then be deduced from this diploid code, as shown in Fig. 6.4.

Chapter 7

DISCUSSION

Summary

The development of MVR-PCR is a good example of how the application of a simple idea can have profound ramifications in different areas of science. The ability to efficiently characterise internal variation at hypervariable minisatellite loci has not only provided new insights concerning the dynamics of particularly volatile regions of the human genome, but also has potentially very important applications in forensic analysis. We have shown that the technique is applicable to any suitable minisatellite locus. This may allow the typing of additional loci which will be necessary to achieve the statistical power that would be required for successful forensic application of the technique and has already enabled detailed comparative analysis of different hypervariable loci. Preliminary investigations of allelic diversity, relationships between alleles and the mutation processes that give rise to hypervariability have been made using MVR-PCR at three minisatellite loci in this laboratory. These are: MS32, MS31A and MS205 (Jeffreys *et al.*, 1991a, 1994; Neil & Jeffreys, 1993; Armour *et al.*, 1993). MVR-PCR revealed for the first time the astonishing levels of allelic diversity that can be exhibited by human hypervariable minisatellite loci, extending estimates of heterozygosity far beyond those calculated from Southern blot length analysis (Jeffreys *et al.*, 1991; Armour *et al.*, 1993; Neil & Jeffreys, 1993). While these discoveries clearly have considerable implications for forensic science and population analysis, it is the ability to analyse *de novo* mutation events at hypervariable minisatellites in detail that is of most importance. Recently a novel technique for the detection and isolation of single MS32 mutant alleles from small pools of sperm DNA has allowed the characterisation of large numbers of mutant alleles (Jeffreys *et al.*, 1994). Combined with characterisation of mutant alleles detected by pedigree analysis at the other loci, this study has shown that mutation at these loci appears to operate by a generalised mechanism that tends to increase allele size by introducing small numbers of repeats preferentially into one end of the tandem repeat array. The detection of variations in the MS32 mutation rate between individuals and alleles and the discovery that these can be associated with particular variant sequences in the flanking DNA has provided further compelling evidence that *cis*-acting flanking elements may be responsible for the initiation of mutation at these loci. In this final discussion I will briefly review the implications of the work described in this thesis, relate it to similar projects being carried out in this laboratory, and attempt to put put it into the wider context of tandem repeat biology. Finally I will indicate where some of these findings may lead us in the immediate future.

Mapping internal variation at minisatellite loci

MVR mapped loci. The three loci which have been characterised by MVR-PCR in most detail in this laboratory were all selected for different reasons and therefore provide a good basis for comparison of the features of different minisatellite loci. MS32 is located interstitially on chromosome 1, while MS31A and MS205 are located near the ends of chromosome arms 7p and 16p respectively (Royle *et al.*, 1988; Royle *et al.*, 1992). These three loci are among the most variable isolated in this laboratory, with heterozygosities >97% and mutation rates to new length alleles of at least 0.4% per gamete (MS205), based on Southern blot length analyses (Wong *et al.*, 1987; Jeffreys *et al.*, 1988a; Armour *et al.*, 1989b; Royle *et al.*, 1992). As with several other human hypervariable minisatellites these loci show associations with other tandemly repeated and dispersed repeat elements. There is evidence that MS32 has expanded from within a retroviral LTR-like repeat located close to a truncated member of the L1 sequence family (Wong *et al.*, 1987; Armour *et al.*, 1989a). MS205 is also associated with a truncated L1 element (Armour *et al.*, 1993) as well as showing linkage to the 3′ and 5′ hypervariable tandemly repeated sequences of the α-globin locus and the variable D16S83 tandem repeat locus. Although no dispersed repeat elements have yet been identified close to MS31A, it provides an extreme example of association between distinct tandemly repeated sequences, with an adjacent minisatellite, MS31B, only 15bp away (Armour *et al.*, 1989a). Interestingly only MS31A shows high levels of variability at the MS31 locus, with only two alleles detectable at MS31B. Both MS31A and MS32 have a wide range of allele sizes, ~2-30kb and 3.5-13kb estimated from Southern blot length analysis of *Alu*I and *Hinf*I digested DNA respectively (Wong *et al.*, 1987), while MS205 has much shorter *Hinf*I digested alleles of 1.5-4.5kb (Royle *et al.*, 1992).

MS32 was chosen for MVR analysis because of the presence of repeat units with sequence variants that create convenient restriction sites, thus allowing discrimination between different repeat types depending on whether or not they are cut with *Hae*III (Chapter 3). However, mapping using this approach was restricted to alleles small enough to amplify by PCR and also required size separation of alleles prior to analysis, making it extremely laborious. The development of MVR-PCR allowed the rapid mapping of alleles of any length making it a much more efficient MVR mapping strategy. MS32 single allele codes were obtained either by comparison of diploid codes generated from genomic DNA, since most repeat units were of the same size and diploid codes therefore stayed in register (Chapter 3), or by allele-specific MVR-PCR (Monckton *et al.*, 1993). Although MS31A also has convenient sequence variants for distinguishing between variant repeats, most of its alleles are too long to map using enzymatic mapping and initial attempts at mapping those alleles small enough to amplify proved extremely difficult (Armour, 1990d). Since there was no apparent repeat unit length variation at this locus it was therefore considered to be a prime candidate for investigation using MVR-PCR (Neil & Jeffreys, 1993). MS205 was not chosen for MVR-PCR by the same criteria as MS31A and MS32. This minisatellite has repeat unit length variants, which precludes the interpretation of diploid codes derived from genomic DNA and also make the design of MVR-specific primers more difficult. However, the restricted size of alleles at this locus makes it possible to amplify most alleles directly from genomic DNA to levels detectable by ethidium bromide staining, making it much easier to separate them prior to analysis and, importantly, allowing the determination of the MVR structure of entire alleles (Armour *et al.*, 1993).

An MVR-PCR system for another highly variable human minisatellite, pλg3 (D7S22) is also under development (T. Guram, unpublished results) and promises to extend such analyses to a fourth locus. MVR-PCR is also being applied to a Y-chromosome specific tandem repeat (M. Jobling, unpublished results) and has been used to investigate other tandemly repeated loci, for example, to analyse variation in the Apolipoprotein B gene 3′ HVR (Desmarais *et al.*, 1993), to characterise mutation events at the CEB1 locus (Buard & Vergnaud, 1994) and to obtain DNA profiles from *Plasmodium falciparum* isolates (Arnot *et al.*, 1993). The extension of the principle to shorter tandem repeats is also being tested, for example in telomere variant repeat mapping (D. Baird, unpublished results) and detection of triplet repeat variants (D. Monckton, personal communication).

Allelic diversity of hypervariable human minisatellites. MVR mapping has the ability to distinguish between alleles of the same, or similar length; in fact individuals who appear to be homozygous are often heterozygous for alleles with different MVR maps (Monckton & Jeffreys., 1991, 1994; Tamaki *et al.*, 1993; Chapter 3). Poisson analysis of the sampling frequency distributions of MS32 and MS205, based on the assumption that all alleles are equally rare, indicates that at least 6800 and 265 alleles respectively must be present at these loci in Caucasians (calculations not shown). It was not possible to obtain a minimum estimate for MS31A allele numbers, since all Caucasian alleles mapped to date are different and the maximum MS31A allele frequency for Caucasians is therefore not known (Chapter 5). However, estimates of allelic diversity indicate that this locus may show even more variation. As has already been mentioned (Chapters 3 & 5) consideration of the current world population size and the mutation rates seen at these loci suggests that the actual numbers of alleles at these loci may be orders of magnitude in excess of such estimates, with perhaps $>10^8$ alleles at MS31A and MS32 (Jeffreys *et al.*,1991a; Chapter 3, Chapter 5). Unlike Southern blot analysis, MVR-PCR has the theoretical capacity to distinguish such huge numbers of alleles. 3-state (a, t & 0-type repeats) MVR-PCR is capable of unambiguously defining 3^{50} ($7x10^{23}$) different allelic states mapped over the first 50 repeat units.

Forensic applications of MVR-PCR. At MS31A and MS32 diploid codes comprised of the superimposed contribution from each allele in genomic DNA can be generated. The underlying allelic diversity at these means that such codes are highly individual specific and thus potentially provide an ideal system for personal identification and can also be used for parentage analysis (Chapter 4). Allele-specific MVR-PCR may also have forensic applications, particularly where mixed DNA samples are concerned (Chapter 5). However, before MVR-PCR can become widely adopted as a standard forensic technique there are technical, as well as commercial and legal obstacles, that need to be overcome. The high germline mutation rates at these loci mean that parental exclusions in the diploid codes from children could be due to mutation rather than non-parentage, a potential drawback in the application of the technique to paternity testing or immigration cases. It is also possible that highly related individuals, for example siblings, will share the same alleles, and therefore diploid code, at a given locus, making them indistinguishable. The most obvious way to overcome these limitations is to extend MVR-PCR analysis to further unlinked loci and use a battery of such systems in forensic analysis. We estimate that one additional locus as variable as MS31A and MS32 would be sufficient to achieve the additional power necessary for widespread forensic application of the technique to be viable (A.J. Jeffreys, personal communication). Although there are few minisatellites that meet the harsh criteria for diploid MVR-PCR (Chapter 4), one possible candidate locus (pλg3) has been identified and its suitability for diploid MVR-PCR is currently being assessed (T.Guram personal communication). An interesting recent development has been the discovery of a variable Y-chromosome-specific minisatellite (M. Jobling, unpublished

results). Since this locus has no homologue, MVR-PCR would generate single allele codes directly from genomic DNA. Such an effectively haploid system would not only be very useful in the analysis of male genealogies but would also allow analysis of mutation in the absence of interallelic interactions. Preliminary trials of MVR-PCR at this locus are being conducted (M. Jobling, personal communication).

If the forensic community are to recognise the advantages of MVR-PCR over tried and tested systems which are already backed by considerable investments of time and money, the technique must be demonstrably more efficient than those currently available. To this end there are several technical improvements that would make forensic analysis using MVR-PCR a more attractive proposition. I have already shown that diploid MVR-PCR can be simultaneously applied to MS31A and MS32 thus increasing sample throughput and the amount of information that can be obtained from a given quantity of input DNA; the ability to perform multiplex MVR-PCR including the anticipated third locus would obviously be advantageous for the same reasons. The use of PCR is becoming more widely accepted in forensic science, most notably in the analysis of microsatellite variation (Hagelberg *et al.*, 1991; Jeffreys *et al.*, 1992; Gill *et al.*, 1994). The current MVR-PCR system limits the number of PCR cycles (15-25 cycles for an input of 100ng genomic DNA) to prevent over-amplification and collapse of minisatellite PCR products (Jeffreys *et al.*, 1988b). As a result, low levels of PCR products have to be detected by Southern blot hybridisation (which incidentally reduces the risk of carry-over contamination and provides a further level of locus-specificity to the analysis of PCR products). However, the use of non-standard sized gels and Southern blot detection using radiolabelled probes in MVR-PCR may reduce its appeal, especially as some forensic laboratories are prohibited from using radioactive isotopes and others are moving toward to non-isotopic detection methods. Both of these potential difficulties could be addressed by the development of in-gel detection of MVR-PCR products, for example by using fluorescent-tagged MVR-specific primers. Besides avoiding Southern blot hybridisation this could increase the throughput of forensic samples by allowing both a- and t-tracks to be loaded in one lane, and also enable repeated use of the same gel; it could also allow standardisation between forensic laboratories using the same apparatus. Preliminary trials of such systems have given encouraging results (R. Fourney, J. Brombaugh, personal communications).

Allelic diversity and population analysis. The ability to rapidly map large numbers of allelic structures from different hypervariable loci and identify relationships between them may prove to be a useful tool in the analysis of recent population divergence. Several groups of closely related alleles at MS32 and MS31A appear to be broadly population specific. However, the high mutation rates at these loci, coupled with a large degree of admixture in modern human populations and the location of some of them in regions of high recombination, may restrict the informativeness of such analyses, except for small well defined populations where there are specific questions of descent that can be addressed. MS205 has a lower mutation rate and also shows a very limited repertoire of MVR-maps at the less variable end of the locus (Armour *et al.*, 1993). This locus may therefore be more useful in population analysis (J.A.L. Armour, personal communication).

Comparative analysis of MVR variation within and between loci. A comparative study of allelic structure both within and between loci is highly informative when considering the mutation mechanisms that give rise to such massive allelic variation at some hypervariable minisatellite loci. Detailed analyses of allelic variation at MS32, MS31A and MS205 have revealed a remarkably similar picture, showing that these tandemly repeated loci

have much more in common than just a high mutation rate that gives rise to a large number of different length alleles. Most significantly all three of these MVR-mapped loci show a marked polarity in allelic variation. Differences within groups of closely related alleles at all of these loci are concentrated at an "ultravariable" end that is different in almost all alleles within such a group and generally constitutes a small proportion of an allele's total length; adjacent to this are regions of map similarity which extend into the minisatellite array. At MS32 and MS205, where several groups of alleles have been mapped in their entirety, these internal regions of map similarity extend to the other end of the alleles. The observed polarity in variation is therefore unidirectional, with a limited range of allelic haplotypes at the relatively invariant end of the tandem repeat array. (Jeffreys *et al.*, 1990, 1991a; Monckton 1993c; Armour *et al.*, 1993). Most information from MS31A is derived from the 5′ end of the minisatellite only, so little is known about variation at the 3′ end. However, it is clear that MS31A has at least one hypervariable end, and preliminary evidence suggests that the other end is less variable (Neil & Jeffreys, 1993, Chapter 5). There is also some evidence for this phenomenon at other tandemly repeated loci, for example the circumsporozoite gene of *Plasmodium falciparum* (Arnot *et al.*, 1993), the human hypervariable minisatellite pλg3 (T. Guram, personal communication) and the A/T rich minisatellite 3′ to the Apolipoprotein B gene (Desmarais *et al.*, 1993).

Minisatellite mutation

Detection of mutation events by Southern blot length analysis. The highly polar pattern of variation in allelic structure observed at these tandem repeat loci implies that during the turnover of repeat units leading to the structures seen in contemporary populations, *de novo* mutations have rearranged repeat units predominantly at one end of the tandem repeat array. However, comparative studies which reveal phenomena such as polarised variability at tandem repeat loci give little information concerning mutation processes; these can only be dissected by direct analysis of new mutant alleles. Southern blot analysis of the CEPH pedigrees of large families revealed several *de novo* mutation events at MS32, MS31A and MS205, allowing the quantification of mutation rates (Jeffreys *et al.*, 1988a; Armour *et al.*, 1989b; Royle *et al.*, 1992); additional MS31A and MS205 mutant alleles were also observed during paternity testing where these loci are routinely used as SLPs. Southern blot length analysis of mutation in pedigrees is limited by the small number of mutants that can be identified, particularly if the locus is not used in paternity testing, for example MS32, or has small repeat units that make it difficult to score subtle mutation events, (length changes of one or two repeat unit occuring in large alleles), for example MS31A. At MS32 such mutants were detected by comparisons of diploid codes in pedigrees (Jeffreys *et al.*, 1991a, Chapter 3), unfortunately this approach is not possible at MS31A (Chapter 4).

Detection of mutant alleles by SP-PCR. The detection and of mutants in pedigrees is laborious and gives no information on mutation rates per individual. In order to overcome these limitations and obtain large numbers of mutant minisatellite alleles for MVR analysis, a simple method for the detection and quantitative recovery of mutant alleles in individual gametes, direct from sperm DNA, has been developed in this laboratory and applied to the MS32 locus (Jeffreys *et al.*, 1994). This approach uses PCR amplification of minisatellite alleles from multiple dilute aliquots of germline (sperm) DNA ("small pool PCR", SP-PCR), so that the products amplified from individual mutant molecules constitute a detectable fraction of the total signal from each pool. SP-PCR, allows 10,000 or more sperm or other cells to be screened in a single experiment for abnormal length minisatellite mutants

sufficiently different in size from the progenitor allele to be resolved by gel electrophoresis. For example, if a minisatellite has a germline mutation rate of 1% per gamete, then on average one mutant molecule will be present per 0.3ng sperm DNA, equivalent to 100 haploid genomes. If all 100 minisatellite molecules (50 per allele) are amplified from such a small pool of DNA, then PCR products will be generated from each of the progenitor allele molecules and from any mutant molecules present. Each mutant PCR product will represent approximately 1% of the total product and should be detectable by Southern blot hybridisation following resolution by electrophoresis. SP-PCR can therefore provide a large number of *de novo* mutant alleles from any sample of germline or somatic DNA, without the need for pedigree analysis. Screening many thousand gametes from the same individual in a single experiment, not only allows recovery and analysis of mutant alleles, but also enables the rates of different types of mutation event to be assessed in separate tissues and for individual alleles. PCR artefacts in SP-PCR are surprisingly rare, suggesting that mutation rates as low as 10^{-4} per gamete should be measurable using this system. Furthermore, like previous single molecule PCR studies (Jeffreys *et al.*, 1990; Monckton & Jeffreys, 1991), SP-PCR does not introduce a significant level of MVR-map artefacts into PCR products recovered from single MS32 molecules, as shown by the constancy of the 3′ end of the MVR maps of MS32 mutants recovered by SP-PCR (Jeffreys *et al.*, 1994). It is now clear that the MS32 mutants recovered by the physical selection of mutant alleles much shorter than the progenitor allele (loss of >30 repeats), followed by recovery of individual mutant molecules by single molecule PCR (Jeffreys *et al.*, 1990), while authentic and similar in frequency to large-deletion mutants detected by SP-PCR, are rare and highly atypical of the bulk of MS32 sperm mutants.

Polarity of variation is caused by a mutation hot-spot. MVR-PCR was used to characterise length change germline mutations identified by Southern blot analysis of pedigrees at MS31A, MS32 and MS205 and also the additional MS32 mutants detected in families by diploid MVR-PCR ternary code comparisons and those recovered by SP-PCR (Jeffreys *et al.*, 1991a; Jeffreys *et al.*, 1994; Chapter 3; Chapter 6). Comparison of the structures of mutant alleles in children with progenitor and non-mutant alleles in the appropriate parent enabled the molecular dissection of several mutation events at each of these loci. This analysis confirmed the prediction of mutational polarity, by showing that most of the mutations detected at all three loci were confined to the first few repeats at the end of the minisatellite already shown to be most variable from the alignment of related alleles. This observation was not an artifact of the MVR mapping procedure; mutants from pedigrees showed changes in repeat unit number compatible with Southern blot allele length estimates, and MS32 mutant alleles small enough to MVR map in their entirety showed repeat copy number changes restricted to the ultravariable end. The direct correlation between observed allelic variability and *de novo* mutation events indicates the presence of a localised mutational hotspot at the ultravariable end of each of these minisatellite loci.

Size gain bias in minisatellite mutation. Most of the characterised mutation events at these loci involved gains of a small number of repeat units at, or close to, the ultravariable end of the tandem repeat array, with evidence for both interallelic and intraallelic unequal exchanges. A similar study of the D2S90 locus, (CEB1) showed that although mutation here was generally more complex, all the interallelic exchanges observed were also size increases displaced towards one end of the locus (Buard & Vergnaud, 1994). The bias towards size gains was demonstrably significant at MS32, where 74% of 761 mutant alleles detected in sperm DNA involved gains rather than losses of repeat units (Jeffreys *et al.*, 1994) and at CEB1 where gains outnumbered losses 2:1 (Buard & Vergnaud., 1994). At MS31A, where paternal mutations are significantly more frequent than maternal ones (Henke *et al.*, 1993), all male

germline mutations characterised were also found to be size increases. Although a relatively large survey at D1S8 has failed to detect structural rearrangements which do not alter allele length (Jeffreys *et al.*, 1991a), it also remains possible that reciprocal exchanges are contributing to the evolution of minisatellite loci.

The bias toward size gain mutations shown by these loci invalidates previous computer modelling of minisatellite evolution that was based on the assumption that repeats are gained and lost with equal frequency (Gray & Jeffreys, 1991; Harding *et al.*, 1992). Such a bias will greatly accelerate the evolutionary expansion of tandem repeat arrays. For example, the bias observed at MS32, would result in alleles growing deterministically at a mean rate of one repeat unit per 43 generations (~1,000 years) (Jeffreys *et al.*, 1994), implying that long arrays could evolve extremely rapidly. The 7 million years that have been estimated to have elapsed since the human/great ape divergence (Koop *et al.*, 1986) is long enough for MS32 alleles 7000 repeats long to have been generated at this rate. However, the lengths of such loci cannot increase indefinitely; Southern blot length estimates suggest that the upper limits for allele length at MS32 and MS31A, are ~800 repeat units, implying that expansion is balanced by some other process that acts to reduce allele lengths. It is possible to envisage several mechanisms that may act to counter the perpetual expansion of minisatellite alleles. The occurrence of rare, but large, deletions, that increased in frequency with increased array size, would serve to counteract more frequent but smaller size increases. Such deletions could be caused by occasional random DSB (distinct from polar DSB) induced collapse, which would be expected to occur more frequently in longer tandem repeat arrays, because the probability of a random DSB should presumably increase directly with the length of DNA tract involved. Another possibility is that small and relatively frequent size gains in the male germline are balanced by less frequent, but larger, deletions in the female germline. The data from MS31A shows that such deletions do occur in the female germline and, although all four maternal mutations so far detected at MS32 show size increases, large deletions have been detected at low frequency in the male germline (Jeffreys *et al.*, 1990). An alternative, or perhaps additional, mechanism that may serve to limit array length could be the operation of truncating selection against gene or chromosomal dysfunction induced by very long arrays (Caskey *et al.*, 1992; Orr *et al.*, 1993; Huntingdon's Disease Collaborative Research Group, 1993). It has been suggested that the discontinous allele length distributions shown by many microsatellites and some of the less variable minisatellite loci are a reflection of selective constraints imposed on the structure of functional chromatin regions (Desmarais *et al.*, 1993).

Evidence for interallelic conversion in germline length gain mutations. In approximately 50% of sperm length gain mutants, at the three loci investigated in this laboratory, and 25% of those characterised at CEB1, there is evidence for interallelic transfer of repeat segments during the mutation process (Jeffreys *et al.*, 1994; Buard & Vergnaud, 1994). Analysis of interallelic exchanges at MS32 suggest that these transfers are non-reciprocal, in that both alleles in an individual appear to be capable of acquiring repeats from the other allele, rather than one allele gaining repeats at the expense of the other. No exchange of informative flanking markers within a few hundred basepairs of any of these minisatellites was ever seen to accompany interallelic mutation events, suggesting that, consistent with earlier studies, unequal homologous exchange is not the dominant mechanism of mutation that gives rise to the hypervariability exhibited by these loci. Rather, these features provide evidence for the hypothesis that interallelic events frequently involve a small "patch" of exchange between alleles, in which a small number of repeat units from the donor allele are inserted into a highly active region close to the beginning of the recipient, as the result of a gene conversion-like process. In most of the simple conversion events at MS32 and MS31A and all

of those observed at CEB1, the beginning of the donated segment in the donor allele and the insertion point in the progenitor allele are in perfect alignment with respect to the beginning of the tandem repeat array (Jeffreys *et al.*, 1994; Buard & Vergnaud, 1994). This registration suggests alignment, and possibly synapsis, of the 5′ flanking regions of pairs of different alleles at each of these loci during these mutation events, strongly suggesting the involvement of a meiotic process. In this case a terminal mismatched region at the boundary of 5′ synapsed DNA and the minisatellite, caused by the different length of minisatellite alleles on homologous chromosomes, may contribute to the instability exhibited by these loci.

The first repeat unit of MS31A alleles shows significant conservation with respect to repeat unit type compared to the second repeat unit, suggesting that the 5′ boundary of the conversion hotspot at MS31A lies between the first and second repeat units (Chapter 5). This may also be the case at MS32 where there is a diverged repeat adjacent to the first repeat unit (Monckton *et al.*, 1994). It is not known whether this diverged repeat is exchanged during interallelic conversions, but its altered sequence suggests that it may be a repeat unit that has accumulated additional sequence variants because it is not rapidly turned over by this process. Despite these observations, the presence of different flanking haplotypes for nearby substitutional polymorphisms occasionally seen in some groups of apparently closely related alleles at MS31A, MS32 and preferentially at the ultravariable end of MS205 (Monkton, 1993c; Neil & Jeffreys, 1993; Armour *et al.*, 1993), suggests that conversion may sometimes include regions of flanking DNA, or that rare interallelic recombinations do indeed occur. It remains to be seen whether the extreme conversion hotspots seen at the ends of minisatellites (approximately 1/250 sperm carry interallelic MS32 conversion products) also serve as true recombination hotspots of a kind which would be detected in linkage analysis. The mutation rate at some minisatellite loci is so high that even if only a very small proportion of mutations were due to "simple" unequal recombination, this would represent a greatly enhanced local rate for meiotic recombination, possibly sufficient to account for the generally increased rates of recombination in subtelomeric regions.

A possible role for *cis*-acting elements in minisatellite mutation. The evidence for a conversion hotspot, confined to one end of the array, not only explains the polarity in MVR map variation observed at these loci but also implies that localised mutation is in some way modulated by element(s) outside the array; for example it may reflect the presence of a local *cis*-acting mutation initiator element that activates an allele for mutation. However, the biological significance of such conversion hotspots is unclear. One possibility is that conversion patches may be the remnants of homology searches between chromosomes required for homologue recognition and the initiation of synapsis at meiosis (Carpenter, 1987). In this case it may be significant that minisatellites are not randomly distributed in the human genome but are clustered near telomeres over regions within which synapsis/recombination is initiated (Royle *et al.*, 1988). The presence of different hypervariable human minisatellites at these locations may reflect the coincidental positioning of a common element that functions as a promoter of chromosome synapsis/recombination and consequently acts as a flanking initiator of minisatellite mutation. However, possible orientation effects relative to chromosome ends and/or replication origins are also possible. Comparisons of the placement and orientation of minisatellites within large scale physical maps may help in distinguishing these possibilities.

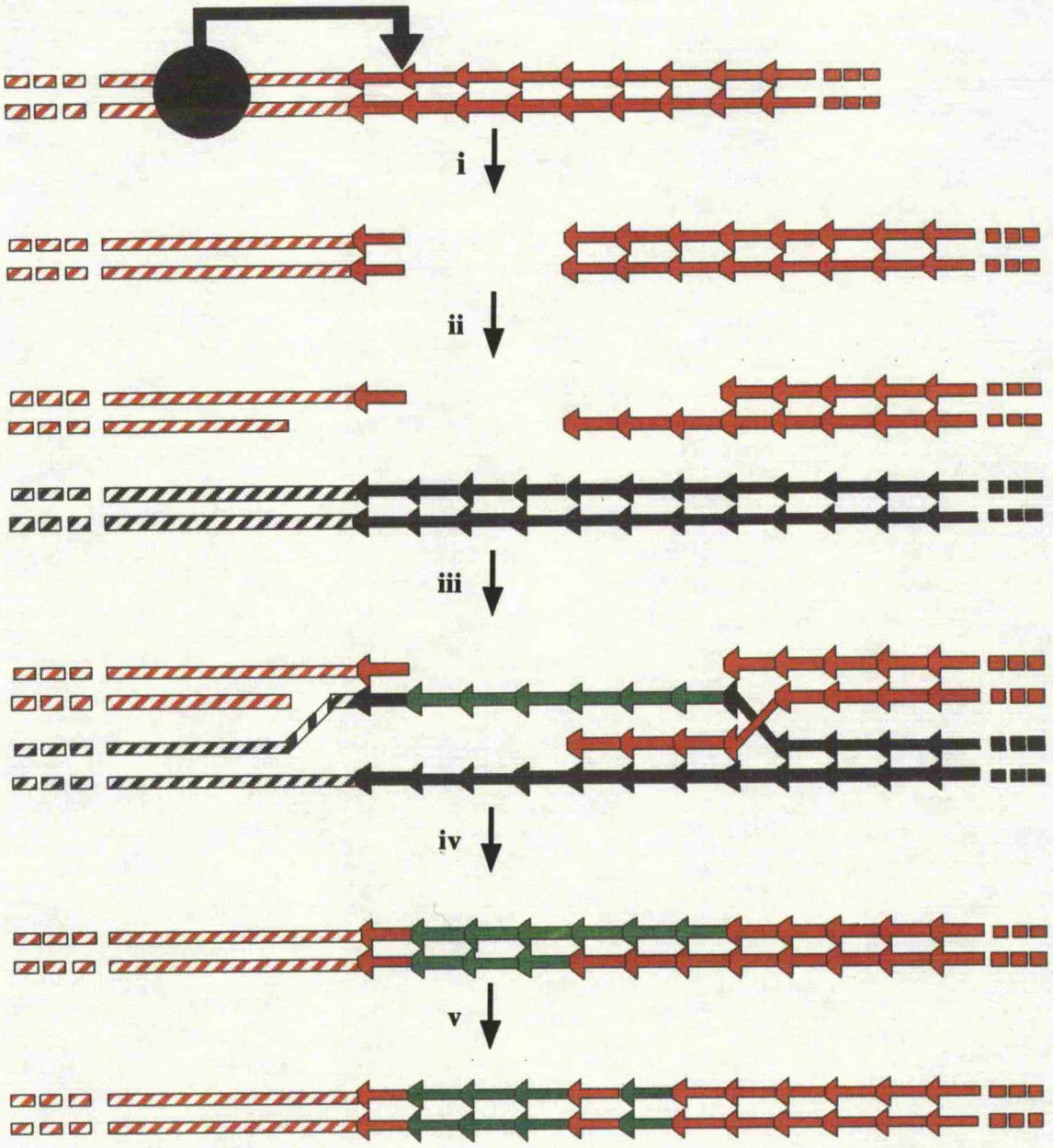


Figure 7.1. A speculative model for mutation at minisatellites.

i. A protein binds to a mutation initiator element in the 5′flanking DNA (diagonal stripes) resulting in the introduction of a double-strand break near the beginning of the recipient allele (red). **ii.** The break expands to a gap and exonuclease creates single-strand overhangs flanking the gap. The recipient allele pairs with the donor allele (black), with the 5′ends of the two alleles in register. **iii.** The gap is bridged by strand invasion. **iv.** The resulting single-strand gap is filled by repair synthesis, (green) and the conversion complex resolved, for example by extrusion of the donor strand from the recipient allele. **v.** Short domains of mismatch repair over the heteroduplex created near the filled gap result in the scrambling of donor and recipient repeat units near the donor segment.

This figure was adapted from Jeffreys *et al.*, (1994)

A model for the generation of allelic variability at some human hypervariable minisatellites. The remarkable similarities between the profiles of allelic structural variability and the mutation events which give rise to them are highly suggestive of a generalised mutation process operating at these different loci. The simplest, though highly speculative, explanation is the GEM (Monckton, 1993e). This proposes that minisatellite mutations arise through the aberrant repair of DSBs that are introduced near the beginning of the tandem repeat array under the influence of a 5′ initiator element (Fig. 7.1). Following DSB initiation, such gaps may open and then be bridged by strand invasion from either the sister chromatid or the homologous allele, to provide a single-strand template for gap repair. Resolution of such a conversion complex would result in the recipient allele gaining a segment of repeats from the donor allele; consistent opening of the initial gap before strand invasion could explain the observed bias towards gaining repeat units. Depending on whether the donated repeats come from the sister chromatid or the homologous allele this model will result in intraallelic reduplications or interallelic conversions such as those seen at the three loci, in particular MS31A, where mutation has been examined in this laboratory (Jeffreys *et al.*, 1994; Chapter 6, Fig. 6.8). The presence, in several cases, of target site duplications in the recipient allele suggest that DSBs may frequently be staggered (Jeffreys *et al.*, 1994; Buard & Vergnaud, 1994). There is also circumstantial evidence for the involvement of DSB gap repair. Gap repair appears to play an important role in the initiation of meiotic recombination in yeast (see Sun *et al.*, 1991b), most notably at the *ARG4* meiotic recombination hotspot at which a *cis*-acting DNA sequence is required to activate the locus for recombination or conversion by the introduction of a DSB (Schultes & Szostak 1991; Massey & Nicholas 1993). The analogy between yeast recombination and minisatellite mutation supports the possibility that male germline minisatellite mutation might also arise predominantly by meiotic conversion processes initiated by a *cis*-acting element (Fig. 7.1).

Evidence for the initiation of terminal interallelic conversions by a *cis*-acting element. The hypothesis of *cis*-acting mutational initiators at minisatellite loci gives rise to certain predictions that can be tested experimentally. Firstly mutation rate would be expected to be independent of allelic repeat copy number, since only those repeats adjacent to the initiator would be vulnerable to this mutation process, irrespective of the size of the tandem repeat array. Evidence obtained from mutations detected by Southern blot length analysis has revealed no obvious correlation between allele length and mutation rates (Jeffreys *et al.*, 1988a; Vergnaud *et al.*, 1991) and direct measurement of allelic mutation rates in sperm at MS32 by SP-PCR supported these findings, showing that mutation rate was independent of allele size (Jeffreys *et al.*, 1994), at least over the range tested (22-164 repeats). A further prediction is that gain or loss of function of such an initiator by mutation of its sequence, sequences with which it interacts, or sequences coding for protein components of any part of this system, will result in changes in mutation rate. Variations in mutation rate between alleles have been detected at MS32 by SP-PCR. A significant reduction in the mutation rate of both alleles was found in one individual, suggesting loss of function of a general component of the mutation machinery (Jeffreys *et al.*, 1994). It will be very interesting to compare the mutation rates of other minisatellite loci in this man with those in the population, assuming that we can extend SP-PCR to these loci and the alleles are small enough to analyse using this technique.

Loss of function of a *cis*-acting initiator of mutation linked to a particular allele would be expected to have particular population-genetic consequences. Alleles with a lower mutation rate than their more unstable counterparts could drift to relatively high population frequencies, and spread through the population as a "dead" allele, hence reducing allelic diversity and the overall heterozygosity of the locus in that population. The recent discovery of alleles at MS32 that

show reduced variability in human populations and are associated with a G to C transversion (O1C) upstream of the array seems to be a prime example of such a phenomenon (Monckton, 1994). The O1C variant is rare in Caucasians (frequency ~0.004) and Japanese (frequency <0.02), but is common in both Zimbabweans and Afro-Caribbeans (frequency 0.19 and 0.13, respectively). Like Caucasians, Japanese show extreme allelic diversity with a very low estimated homozygosity. In sharp contrast, African diversity is substantially reduced, with a ~70-fold increase in homozygosity. This was found to be largely due to the presence of a single allele at high frequency (0.103 cf. most common Caucasian allele frequency of 0.009) in both Africans and Afro-Caribbeans. This allele contains 38 repeats of the 29bp MS32 repeat unit and is associated with the O1C variant. There are several other African and Afro-Caribbean alleles closely related to this allele, these are also present at elevated population frequency and all share the same flanking haplotype containing the O1C variant. Comparison of the allelic variability of O1C and O1G linked African and Afro-Caribbean alleles by MVR mapping suggested that alleles in this group had not risen to high population frequency by chance, due to recent genetic drift of alleles showing a normal mutation rate. Rather it appeared that the O1C variant was specifically associated with reduced variability and polarity, irrespective of MVR code or more distal 5′ flanking haplotype.

Mutation suppression at MS32 by a 5′ flanking variant. The O1C variant was also found associated with other MS32 alleles, including some in individuals of apparently Caucasian descent. Comparison of mutation rates of O1G- and O1C-linked alleles, including the 38 repeat allele, by SP-PCR demonstrated a frequently dramatic reduction in O1C mutation rate, irrespective of allele length, MVR code or flanking haplotype at other sites. Sequence analysis of the immediate 324bp of 5′ flanking DNA, first full MS32 repeat and 181bp of immediate 3′ flanking DNA of 16 different O1C and O1G alleles, of various 5′ haplotypes, analysed for mutation failed to detect any additional variants. It is therefore likely that mutation suppression is not only correlated with the O1C variant but is caused by this single base transversion 48bp upstream of the MS32 array.

Beside having a reduced mutation rate the size distribution of mutants at O1C alleles was also abnormal. Mutations involving repeat unit gains were reduced by ~80 fold (~0.007% per sperm in O1C-linked alleles compared with 0.55% in O1G alleles). Small deletions were more common than gains, though occurred less frequently than in O1G-linked alleles. The rate of appearance of larger deletions (>5 repeats) was similar to that seen in O1G-linked alleles. Interestingly one O1C allele with a less marked reduction in mutation rate than the others examined showed a similar spectrum of mutations to that of O1G alleles. This allele is closely related to non-mutating alleles with which it shares a common 5′ flanking haplotype. Unfortunately, the other O1G-linked allele in this individual was too large (>800 repeats) for mutation analysis, and it is not known whether elevated mutation at the O1C allele occurs in *cis*, or instead affects both alleles through some mechanism acting in *trans* which causes a more widespread increase in minisatellite mutation rate. Sequence analysis of DNA flanking this allele failed to reveal additional sites of variation which might alleviate O1C-associated suppression of mutation.

Analysis of individuals heterozygous for the 38 repeat O1C allele and short O1G alleles showed that there were almost no gain mutations (rate ~0.01% per sperm) in the former but a normal rate (0.46% gain mutations per sperm) and distribution of mutant allele sizes in the latter, indicating that mutation suppression in O1C-linked alleles occurs in *cis*, not *trans*. No gain mutations were seen in 11,000 sperm tested from a homozygote for the common 38-repeat O1C allele, suggesting that mutation suppression in O1C/G heterozygotes occurs strictly in *cis*,

and does not result from competition between O1C and O1G alleles for some rate-limiting factor required for mutation. Comparison of the internal structures of mutant alleles from an O1C/G heterozygote showed the familiar pattern of mutation in the O1G allele. Almost all gain mutants were polar and restricted to the first few repeats with many showing clear evidence for transfer of repeat units from the corresponding position of the O1C allele, however with no exchange of flanking markers. Deletion mutants were less obviously polar but again in several cases showed evidence of interallelic transfer. By contrast mutations recovered from the O1C allele almost exclusively involved simple, and not obviously polar, deletions of one or more repeats, with no evidence for interallelic transfer of repeat units. These data showed that O1C alleles can act as donors of sequence to mutating O1G-linked alleles, but not recipients of information in interallelic exchange events with these alleles.

The results of this study explain the elevated population frequency of the common 38 repeat African allele and also provide strong support for the view that instability is not necessarily intrinsic to the minisatellite. They also suggest that the initiating chromosome is the recipient of information during mutation events, as has been proposed previously (see Szostak *et al.*, 1983). The implications that mutation is effected by elements outside the tandem repeat array and that donor and recipient allele function during mutation are distinct, are fully consistent with the mutation initiator model, in which mutational activation of the recipient allele precedes synapsis with the donor and information transfer from donor to recipient (Jeffreys *et al.*, 1994; Chapter 6, Fig. 7.1).

Although co-transfer of the flanking O1C variant was not observed during the largely unidirectional transfer from O1C to O1G alleles in O1C/G heterozygotes, the data do not exclude the possibility of this occuring at a low rate, perhaps by the same process that occasionally switches flanking markers within groups of aligned alleles. If such a biased conversion process existed it would constitute a form of meiotic drive at the MS32 locus which would tend to sweep through populations, progressively immunising MS32 alleles to length gain mutation as further O1G alleles were converted to the O1C state and potentially reducing mutation rate and locus variability. Such a scenario may explain the existence of monomorphic minisatellites and the presence of short, common alleles at other minisatellite loci, such as have been identified in Caucasians at pλg3 by MVR mapping (T. Guram personal communication) and by Southern blot analysis at the D17S79 locus (Waye *et al.*, 1994). To test this hypothesis, and distinguish it from the possibility that these may be representatives of the ancestral state prior to repeat unit expansion that have risen to appreciable population frequency by genetic drift, it will be necessary to develop SP-PCR for the accurate measurement of mutation rate per allele at these loci. We can then compare the mutation rates of short, common alleles with those of similar size, but lower population frequency. The short Japanese MS31A alleles (group 19, Fig. 5.4), which show high population frequency and do not exhibit polar variation, provide good candidates for such an analysis. MVR-PCR can be used to look at the population distribution of these alleles and sequence analysis of the MS31A 5′ flanking DNA of Japanese alleles can be used to reveal whether those in group 19 are also associated with a rare flanking variant close to the minisatellite array. It is a priority to develop SP-PCR at this locus, so that the relative mutation rates of these alleles can be investigated.

The possible biological function of minisatellite flanking sequences. The most likely explanation of a flanking variant associated with loss of ability to initiate mutation is that the O1G/C transversion inactivates some component of the mutational machinery. For example it may destroy a binding site for an activator of mutation, in which case at least part of the initiator lies very close to the mutation hotspot, or fortuitously create a binding site for a protein which blocks mutation. The sequence around O1C gives few clues as to how it may

participate in the initiation of mutation. The O1C variant disrupts an 18bp same-strand mirror image sequence (TTGGTGGGA/AGGGTGGTT) of unknown significance immediately upstream of the diverged repeat next the MS32 minisatellite array. If the hypothesis of a common *cis*-acting mutation mechanism operating at the different loci we have examined is correct, we might expect to find conservation of flanking sequences between them. However, no conserved sequences, including that containing the O1C variant, have been identified in the few hundred basepairs of 5′ flanking DNA immediately adjacent the human minisatellites showing polar mutation (A.J. Jeffreys, personal communication). The O1 site is located within an element homologous to the LTR of the retroviral-like RTVL-1 family (Maeda, 1985), from which the MS32 tandem repeat array has amplified (Armour *et al.*, 1989a). Other examples exist of human and mouse minisatellites which have evolved from within retroviral LTRs (Kelly *et al.*, 1991; Mermer *et al.*, 1987), in particular from members of the mammalian apparent LTR retrotransposon (MaLR) superfamily (Smit, 1993). There is abundant evidence that MaLR LTRs are recombinationally active and promote genome instability, and it is possible that an LTR-associated recombination hotspot exists at the O1 site which serves to initiate MS32 mutation (Monckton *et al.*, 1994). However, the 30 bp region spanning O1 is substantially diverged from the RTVL-1 LTR sequence and does not detect similar sequences in DNA sequence database searches (Monckton *et al.*, 1994).

It is possible that interactions with more distal elements common to several minisatellites are mediated by proteins that bind different sequences immediately flanking each locus. However, there are more likely alternative explanations. One possibility is that the O1 site is not a component of the mutation initiator, but that the O1C transversion creates a binding site for a protein that represses mutation initiation. Another possibility is that the O1C variant is in strong linkage disequilibrium with other variant(s) further 5′ or 3′ to the tandem repeat array which suppress mutation, although this seems unlikely given the diversity of O1C-linked MVR codes and 5′ haplotypes and the lack of immediate 3′ flanking variants.

A role for minisatellite binding proteins? Proteins that bind the minisatellite repeat array itself might influence mutation, either directly, or perhaps through long range interactions with other proteins bound to the flanking DNA. Proteins that specifically bind G-rich (Collick & Jeffreys, 1990; Wahls *et al.*, 1991; Collick *et al.*, 1991) and C-rich (Yamazaki *et al.*, 1992) tandem repeat sequences have been identified. One of these proteins Msbp-1 was isolated from mice by using synthetic binding substrates with homology to the minisatellite "core" sequence and is apparently ubiquitous among eukaryotes. It binds multiple single-stranded G-rich repeats in a sequence-specific manner, suggesting that the core sequence may indeed have some functional relevance (Collick *et al.*, 1991). Other well characterised single stranded binding proteins, for example the *E. coli* proteins ssb and RecA and the gene32 protein of bacteriophage T4, have been shown to play important roles in recombination (reviewed by Dressler & Potter 1982; Chase & Williams, 1986). It is tempting to speculate that Msbp-1, or related proteins may bind single stranded DNA generated by exonuclease activity following a DSB, perhaps stabilising either the G-rich or C-rich strand and thus activating it for strand-exchange during repair in a manner similar to RecA (Dressler & Potter, 1982). Since 5′-3′ direction of the G-rich strand of the minisatellites analysed in this laboratory is not constant with respect to the 5′-3′ orientation of the minisatellite from the ultravariable to the less variable end, such mechanisms would have to be independent of strand base composition. Since none of these proteins have been sequenced and their functions remain unknown, extreme caution in according them any relevance to minisatellite biology is necessary, until a biological link is formally established.

Evidence for more complex length gain mutation processes. There is a wide range of structures seen in mutant minisatellite alleles at all loci investigated and simple unequal interallelic or intraallelic conversions according to Fig. 7.1 cannot account for many of the more complex mutants. For example, some MS32 mutants appear to result from allele reduplication plus interallelic conversion (Jeffreys *et al.*, 1994) and the majority of mutations at CEB1 consist of non-polar intraallelic deletions and reduplications, which are often complex. Both interallelic and intraallelic mutation events at CEB1 often involve further duplications or deletions within the exchanged repeats (Buard & Vergnaud, 1994), suggesting that most mutations arise by a mechanism distinct from terminal conversion by DSBR; the authors propose a mechanism of staggered single-stranded nick repair by SCE. These observations suggest that minisatellite mutation can be a multistep process, for example resulting from additional breaks being introduced into the conversion complex itself. At MS31A, and more frequently at MS32, mutant alleles containing repeat unit blocks of no obvious origin were seen, for example mutants 1 and 3, Fig. 6.4. In a few cases at MS32, these anomalous repeats could have resulted from short domains of mismatch repair occurring on alternate strands of heteroduplex DNA, formed adjacent to the repaired gap, leading to microconversions next to the conversion domain (Jeffreys *et al.*, 1994). In most cases though, the origin of these anomalous repeats remains completely mysterious, and again points to a complex multistep mutation process that can scramble the order of repeats near the beginning of the array. The complexity of mutations at some loci may not only reflect a single multistep mutation process, but also the simultaneous or sequential operation of different mutational mechanisms.

Germline deletions are predominantly simple intraallelic events. Interestingly, deletion mutants show very little evidence for interallelic exchange and contain relatively few patches of anomalous repeats. A few of the deletions at MS32 and CEB1 showed evidence for interallelic exchange, but most involved the simple loss of a contiguous run of repeats (Jeffreys *et al.*, 1994; Buard & Vergnaud, 1994). The few deletion mutants observed at MS205 and MS31A also involved straightforward loss of a varying number of repeat units, consistent with single, entirely intramolecular events (Jeffreys *et al.*, 1994; Chapter 6). The lower frequency of deletions compared to gains of repeat units observed at all loci examined suggests that they may arise through a different mechanism. Nonetheless, the deletions characterised at MS32 do still display a significant polarity in the location of their breakpoints (Jeffreys *et al.*, 1990, 1994) and one deletion at MS31A was terminal (Chapter 6; Jeffreys *et al.*, 1994). It is therefore possible that these mutants are initiated in the same way as the small gains (eg. by a DSB), but are then processed via an alternative intramolecular repair pathway, for example SSA, that operates at lower frequency than that responsible for size increases.

Differences between germline and somatic mutation processes. By definition somatic mutants must be mitotic in origin and it is therefore instructional to compare them with germline mutations to see if there are any similarities which indicate whether the latter occur during either mitosis or meiosis, or differences which indicate that different mutational processes operate in different tissues. SP-PCR was used to measure the mutation rates of MS32 directly in blood and sperm cells. The mean sperm mutation rate for MS32 established by SP-PCR is 0.8% per gamete (Jeffreys *et al.*, 1994), very similar to the approximate rates of 1.0% and 1.4% determined by pedigree analysis for paternal and maternal mutation respectively (Jeffreys *et al.*, 1991a). However, the somatic mutation rate was found to be much lower (0.06% per gamete) (Jeffreys *et al.*, 1994), consistent with earlier studies which indicated that somatic mutation is rare ($<4x10^{-5}$, per allele per mitosis, Armour *et al.*, 1989b). If the same low rate

of mitotic mutation also applies to germ cell lineages, then the majority of germline mutation events cannot be coupled to cell division, since they occur at much higher frequencies. This would be consistent with previous suggestions that most germline mutation events arise at one stage of gametogenesis, possibly meiosis (Jeffreys *et al.*, 1988a). Furthermore, almost all MS32 sperm mutants detected by SP-PCR were different, as predicted for meiotic products. One example of mutational mosaicism which must have arisen before meiosis, discovered in the sperm of another individual, showed no polarity and was more reminiscent of putative mutants seen in blood DNA (Jeffreys *et al.*, 1994). There was no evidence for size gain bias in the MS32 blood mutants (Jeffreys *et al.*, 1994) and, in common with most of the germline deletion events analysed at these loci, all of the somatic mutants identified at MS32 and the one confirmed at MS31A had relatively simple structures, with no anomalous repeats, and appeared to be entirely intraallelic in origin. Mapping of the break-points of *de novo* MS32 deletion mutants recovered from blood *in vitro* had previously showed a bias towards deletion nearer the ultravariable end of the locus (Jeffreys *et al.*, 1990). The five MS32 somatic mutants characterised following identification by SP-PCR were present in a short allele (63 repeats) making it difficult to attach any significance to their location, however, four were present in the first half of the allele. The somatic mutant identified at MS31A was also located close to the beginning of the tandem repeat array. The substantial differences between the frequency, size distribution and, apparently, internal structure of MS32 mutant alleles between sperm and blood DNA suggest that the processes resulting in sperm mutation are largely germline specific, and possibly meiotic.

Differences between male and female germline mutation. Comparisons between male and female germline mutation rates may also give some indication as to whether mutations are occuring by meiotic, or mitotic processes. A fundamental prediction arising from the proposal of a predominantly mitotic mutation process, is that the mutation rate will be higher in the male germline than in the female germline, since there are far more mitoses in male gametogenesis (Vogel & Rathenberg 1975). Recent evidence has shown that a number of hypervariable minisatellites do indeed have significantly higher mutation rates in the male germline than the female germline, consistent with mutation during mitosis. This has been observed by pedigree analysis at MS205 (Jeffreys *et al.*, 1994), MS31A (Henke *et al.*, 1993), MS43A, pλg3, and at MS1 (Olaisen *et al.*, 1993; L. Henke personal communication), where size gain increases seem to occur preferentially in the male germline (L. Henke personal communication). At CEB1 the bias is extreme, with a mutation rate of 15% per male gamete, compared to 0.3% per maternal gamete (Vergnaud *et al.*, 1991). MS32 is not routinely used in paternity testing and the number of pedigrees examined was unfortunately too small to detect such a bias. The paucity of polymorphic minisatellites in the sex-specific region of the X-chromosome (Donis-Keller *et al.*, 1987; Fraser *et al.*, 1989; Armour *et al.*, 1990; Consalez *et al.*, 1991) compared to the abundance found in the the pseudoautosomal region (Cooke *et al.*, 1985; Page *et al.*, 1987) which has a homologue in male meiosis, is also suggestive of a specific role for the male germline in minisatellite evolution.

Meiosis or mitosis? Although minisatellite mutation has been characterised in considerable detail, the question of whether mutation processes operate during meiosis or mitiosis is still a tantalising one. The rather speculative deductions from the mutational analyses performed to date give conflicting indications concerning the timing of minisatellite mutation in the cell cycle, suggesting that the assumptions on which they are based are either too simple, or do not apply. Qualitative comparison of somatic mutation with male and female germline mutation events may help to resolve this apparent contradiction. Unfortunately female germline material is much more

difficult to obtain than that from males and our observations, based on the few mutation events detected in pedigrees, give a very limited picture of female germline mutation processes at these loci, compared to the wealth of data derived from sperm mutants. However, using the information available it is possible to conceive some tentative explanations for the observed data. All somatic and female germline mutation events analysed so far have been simple intraallelic duplications or deletions, while male germline mutations include this type of mutation as well as more complex events involving interallelic exchange. It is therefore possible that at least two separate mutation processes are governing the evolution of these loci. A simple mitotic intraallelic mutation process, for example replication slippage, could give rise to deletions and duplications of repeat units with equal frequency and operate at low level in all tissues and in both males and females. The relatively higher male germline mutation rate could be explained by overlaying this with a male germline specific mechanism, operating at higher frequency to produce polar size gain mutations by inter or intraallelic conversion (see Dover, 1989). The exchange of information between homologous chromosomes, increased rate compared to somatic mitotic mutation and evidence for 5′ alignment of the minisatellite flanking DNA during conversion, suggests that this may occur during male meiosis, although mitotic mechanisms cannot be absolutely ruled out.

Evidence for the involvement of replication slippage in minisatellite mutation. There is circumstantial evidence that replication slippage as well as unequal conversion may contribute to minisatellite allelic variability. Simple intraallelic reduplications or deletions of repeat units, as seen at all of the MVR mapped loci, would be compatible with such a mechanism, but in fact cannot be formally distinguished from intraallelic USCE. It has been suggested that at di- and trinucleotide repeats small events may be caused by slippage and larger ones by USCE (see Nelson, 1993), this may also apply to minisatellites. At MS32 and MS205 many intraallelic mutation events involved only one, or a few, repeat units. At MS31A there is no information concerning mutations involving less than two repeat units, due to the difficulty in resolving small differences in size between large alleles, meaning that such events could not be scored. However, larger intraallelic reduplications were seen (Chapter 6; Jeffreys *et al.*, 1994) and small internal differences in allele structure between the alleles in aligned groups of MS31A alleles suggested that slippage-like mechanisms may be operating internally at this locus. (Chapter 6. Fig. 6.4). These variants, which included both small deletions and duplications of repeat units, were seen more frequently at MS31A than MS32 (compare Figs. 3.5 and 6.4; see also Monckton, 1993c) and MS205, where very few were seen (Armour *et al.*, 1993). MS31A has the shortest repeat unit (20bp) of these three loci, MS32 has longer repeats (28 or 29bp) and the repeat units of MS205 are longer still (45-54bp). Studies have shown that the rate of slippage, for small repeats (<5bp) at least, increases with decreased repeat unit length (see Dover, 1989; Schlotterer & Tautz, 1992), increased array length (Levinson & Gutman, 1987; Murphy *et al.*, 1989) and repeat unit homogeneity (Weber, 1990). The increased level of repeat unit length and sequence variation seen at MS205, in comparison to MS31A and MS32, would be expected to make it more refractory to replication slippage, reducing the number of small internal variants seen. Furthermore, intragroup comparisons of O1C linked MS32 alleles, which have been shown not to participate in interallelic conversion, suggest that polar mutations are less prevalent in these alleles, and that the residual mode of mutation may involve simpler processes, such as replication slippage, which result in small and non-polar duplications and deletions further into the array (Monckton *et al.*, 1994). Unfortunately it is difficult to confirm this by direct analysis of mutants derived from O1C alleles. Size gain mutations are too rare (~10^{-4} per gamete) to be isolated in bulk by SP-PCR of sperm DNA, and although deletions in the 38 repeat O1C allele appear to conform to this prediction, they may include PCR artefacts (Jeffreys *et al.*, 1994). These

observations are consistent with a possible role for replication slippage operating in tandem with the conversion mechanism responsible for most length gain mutation, to cause small, non-polar, changes in allele structure at all of these loci.

It is more difficult to explain how some of the other features of the mutation process we have observed, such as, size gain bias, terminal polarity, and interallelic exchanges would result from simple replication slippage. Although interallelic strand switches could occur at a replication fork, these would be expected to result in the inversion of repeat units (Wells & Sinden, 1993). An explanation for polarity might be that the direction of DNA replication in the region close to a minisatellite favours the accumulation of mismatch repair processes (and therefore mutation events) at one end of the tandem repeat array (Richards & Sutherland, 1994). However, if this were the case mutation rate would be expected to increase the further a repeat was from the replication origin, resulting in an increased mutation rate with allele repeat copy number, contrary to our observations (Jeffreys *et al.*, 1994). Furthermore, *in vitro* studies suggest that slippage rates increase with A/T richness and decrease with repeat unit length (Schlotterer & Tautz, 1992). Since microsatellite loci generally have mutation rates of <0.05% per gamete (Kwiatkowski *et al.*, 1992, Weissenbach *et al.*, 1992; Bowcock *et al.*, 1993; Weber & Wong 1993; Banchs *et al.*, 1994) it seems unlikely that slippage alone could be responsible for the much higher mutation rates observed at G/C rich minisatellites with longer repeats. Although these considerations suggest that slippage is not the predominant mode of mutation at the minisatellites we have examined they do not eliminate its possible involvement in some aspect of allele diversification at these, or other loci.

MVR analysis of the AT rich minisatellite of the Apolipoprotein B gene (ApoB) shows also revealed polarity in variation like that observed at MS32, MS31A and MS205. However, there is no evidence for interallelic exchange at ApoB, and length variation between related alleles at this locus is found at the end of the locus with least MVR variation, making it seem likely that the mechanism of mutation is sequence dependent replication slippage (Desmarais *et al.*, 1993). The 5′ end of this minisatellite, which shows length variation between related alleles, has a number of alternating 15bp repeats of two related AT-rich sequences (X and Y) that have the potential to pair, forming hairpin structures. If such a structure were formed between non-consecutive X and Y repeats DNA polymerase error would be expected to result in the gain or loss of an even number of repeats. The 3′ end of alleles at this locus have variant repeats containing C and G nucleotides which would be expected to break perfect matches between repeats, thus reducing the potential for secondary structure formation, and hence mutation by slippage, explaining the reduced variation observed at this end of the locus. Similar factors have been proposed to influence trinucleotide expansion mutations at the CGG triplet repeat of the FMR1 locus, which are responsible for fragile-X syndrome, and by inference some of the other triplet repeat disease loci (Kunst & Warren, 1994).

Two mouse hypervariable minisatellite loci, Ms6-hm and Hm-2, have long tandem repeat arrays (>1000 repeat units) of a short GC-rich repeat unit (5bp and 4bp respectively), with no observed repeat unit variants (Kelly *et al.*, 1991; Gibbs *et al.*, 1993). Ms6-hm and Hm-2 show roughly even numbers of gains and losses of repeat units, with a significant bias toward mutation in the male germline at Ms6-hm. They both have higher germline mutation rates (2.5% and 3.6% per gamete, respectively) and a much higher frequency of somatic mutation (Ms6-hm, 3%; Hm-2, 20%) than has been observed at any human minisatellite locus. It is possible that the higher somatic and germline mutation rates at these loci are the result of a different mutation process acting on these repeats, or perhaps a process

that mutates arrays of shorter repeat units at higher frequency. The large size of these mouse loci might be presumed to increase their instability, since has been proposed that long homogeneous tandem repeat arrays of a short repeat unit would be expected to show high rates of length change mutation due to slippage or unequal-crossover (Stephan, 1989). The observations at Apo(B), Hm-2 and Ms6-hm are therefore consistent with the hypothesis that slippage-like mechanisms may be the dominant mode of mutation at minisatellite loci with shorter, and/or more AT-rich repeats in contrast to the conversion-like processes that typify mutation at GC-rich loci with longer repeats.

Interestingly one of the most variable human minisatellites, MS1, (mutation rate 5% gamete) appears to be similar in some respects to these mouse loci. It also has long tandem arrays (140-2500 repeats) of a comparatively short, GC-rich repeat unit (9bp), but has several repeat unit variants and does not exhibit somatic instability (Wong *et al.*, 1987; Jeffreys *et al.*, 1988a). This locus has a high germline mutation rate compared to minisatellites with longer repeats, for example, MS31A (20bp) and MS32 (29bp), and, unlike these loci, undergoes gains and losses of repeat units with roughly equal frequency (Jeffreys *et al.*, 1988a). However, MS1 does not appear to exhibit the same polarity in allelic variation as MS32, MS31A and MS205 since unrelated MS1 alleles have relatively conserved 5′ and 3′ ends, but show considerable difference in internal structure (Gray & Jeffreys, 1991). Although initial studies of MS1 showed equal frequencies of mutation in the male and female germline (Jeffreys *et al.*, 1988a), in contrast to the male bias found at other hypervariable human minisatellite loci, more recent data have shown a 2:1 bias toward mutation in the male germline at this locus (Olaisen, *et al.*, 1993) and that size gains occur preferentially in the male germline (L. Henke, personal communication). This mutational profile suggests that the predominant mode of mutation at MS1 is by mechanisms distinct from those generating new length alleles at the MVR mapped loci, but may be similar to the mouse hypervariable loci, perhaps involving replication slippage.

A complex picture of minisatellite mutation. Mutation and allelic variability have now been investigated in a number of hypervariable minisatellite loci and it appears that the processes involved in minisatellite mutation and the factors influencing them are many, varied and more complex than was initially anticipated. Although these observations give some predictive power as to the types of mutation mechanism we may expect to be responsible for the generation of allelic diversity at a new hypervariable minisatellite locus, it seems likely that the specific factors affecting the evolution of any one minisatellite will vary between loci. Hypervariable minisatellites all share the features of tandemly repeated structure and high heterozygosity, generated by a high mutation rate to new length alleles, but they vary considerably in repeat sequence, length and copy number and in allele size ranges. Such differences may explain the observation of qualitative and quantitative variation in the mutation processes operating at different loci and in different tissues. For example, some loci may mutate predominantly by replication slippage while conversion-like processes might contribute to most allelic variability at others. It also seems likely that the balance between different turnover processes may be influenced by the immediate flanking sequence and/or genomic location of a minisatellite locus. Minisatellites are known to cluster in subtelomeric DNA (Royle *et al.*, 1988) and are often expanded from within, or associated with, dispersed repeats (Armour *et al.*, 1989a, 1993), but the relevance of these observations remains unclear. Other factors, such as location in late or early replication regions of the genome, the proximity of active genes or other functional components of the chromosome and methylation status, may also influence mutation rate.

Relevance to shorter tandem repeats?

Our analysis of mutation at some of the most hypervariable human minisatellite loci has provided detailed information concerning mechanisms of tandem repeat turnover. It is of considerable interest to discover whether any of these findings are relevant to the generation of variation at loci with tandem repeats of shorter sequences, (microsatellites, simple tandem repeats, STRs) particularly since instability at such loci is being increasingly associated with somatic and genetic human diseases.

STRs and cancer. Recent studies have demonstrated the association of unstable STRs with multiple forms of cancer (reviewed by Richards & Sutherland, 1994). In one of these diseases, hereditary non-polyposis colon cancer (HNPCC), the genome wide instability of mono-, di-, and trinucleotide repeats has been shown to result from defects in a protein, hMSH2, responsible for mismatch repair of heteroduplex DNA (reviewed by Bodmer *et al.*, 1994). This protein was identified by homology to bacterial and yeast proteins that also bind to and repair mismatched sequences of DNA, and in these species there is evidence that such mismatches are generated by replication slippage (Strand *et al.*, 1993; Lustig & Petes, 1993). It has therefore been suggested that malfunction of trans-acting factors may also cause the STR instability manifested in other cancers, and perhaps other heritable diseases, by the aberrant repair of DNA mismatches that arise during replication slippage (Kunkel, 1993; Richards & Sutherland, 1994). Single strand breaks occuring within the repeated region during replication have been proposed as the initiation events of replication slippage (Richards & Sutherland, 1994) and strand displacements, which are known to occur during replication from such a nick, followed by repeated rounds of strand displacement/slippage and ligation could explain the expansions seen in triplet repeat arrays (reviewed by Wells & Sinden, 1993). By contrast, our studies have indicated that minisatellite mutation is influenced by *cis*-acting elements outside the tandem repeat array and may be initiated by DSBs, making a predominant role for replication slippage seem unlikely. Although it is still possible that some minisatellites, for example MS1, which has the shortest repeat units of any minisatellite, do mutate by such mechanisms, there is no evidence of instability at this or several other human hypervariable minisatellite loci examined in HNPCC cell lines (M. Allen, personal communication).

Triplet repeat disease loci. STR instability was first implicated in human genetic disease by the discovery that fragile-X syndrome (FraX) was caused by expansion of a CGG triplet repeat (FRAXA) (Kremer *et al.*, 1991; Fu *et al.*, 1991), located in the 5′ untranslated DNA of the FMR1 gene. Since then an increasing number of predominantly neurological disorders have been associated with large increases in repeat copy number at GC-rich triplet repeat loci, which have consequently been the subject of much recent research. These include myotonic dystrophy (DM), spinobulbar muscular atrophy (SBMA), Huntingtons disease (HD), spinocerebellar ataxia type 1 (SCA1), FRAXE linked mental retardation and dentatorubral pallidoluysian atrophy (reviewed by: Caskey *et al.*, 1992; Richards & Sutherland, 1992, 1994; Kuhl & Caskey 1993; Nelson, 1993; Mandel, 1994). Several features of these repeat unit expansions are superficially similar to the minisatellite mutations we have observed. The most obvious of these is a bias towards size increases in all of these diseases, although the expansions seen at FRAXA, FRAXE and DM can be massive compared the other triplet repeats and minisatellite repeat unit gains. There is also evidence for differential sex-specific effects on mutation at some of these loci, although these are not consistent; in some diseases eg FraX large expansions occur preferentially in the female germline, while in others, eg. HD, SCA1 and SBMA, increased instability is found in the male germline. There are also contrasts between mutation at these

triplet repeat loci and the human hypervariable minisatellites we have investigated. The FMR1 (FRAXA), DM and FRAXE triplet repeats exhibit high levels of somatic instability, detected as mosaicism in Southern blots of genomic DNA isolated from blood and other tissues of single individuals. In FRAXA there is evidence that this is limited to a brief period during early embryonic development, in a manner analogous to the murine minisatellite loci, Ms6-hm and Hm-2 (Kunst & Warren, 1994). However, the human hypervariable minisatellite loci we have studied show levels of somatic mutation which are generally too low to detect in this manner.

An early observation was the increase of mutation rates at triplet repeat disease loci with allele length; this was termed "dynamic mutation" and explains the phenomenon of anticipation exhibited by these loci (Richards & Sutherland, 1992). There is evidence from some loci that this is a multi-step process, with alleles first undergoing modest size increases from the normal size range to create a pool of "premutation" founder alleles that are much more likely to undergo much larger further expansions (reviewed by Richards & Sutherland, 1994; Mandel, 1994). Our studies have shown that minisatellites do not appear to show a correlation between allele length and mutation rate (Jeffreys *et al.*, 1994), but rather that flanking DNA can directly influence tandem repeat instability (Monckton *et al.*, 1994). If an analogous phenomenon exists at some of the triplet repeat disease loci it may therefore be possible to identify specific and localised flanking DNA variants that can influence trinucleotide repeat instability in these diseases. Analysis of flanking markers in FraX, HD and DM has revealed evidence that these loci show a greater tendency to expansion in particular chromosomal lineages. These data have been interpreted as being the result of a limited pool of ancestral founder haplotypes, in linkage disequilibrium with relatively high copy number and therefore potentially unstable trinucleotide repeat arrays, from which the expanded repeats found in modern populations arise. (Richards *et al.*, 1992; Richards & Sutherland, 1992; Oudet *et al.*, 1993; MacDonald *et al.*, 1992; The Huntington's Disease Collaborative Research Group, 1993; Harley *et al.*, 1992; Imbert *et al.*, 1993). However, recent evidence suggests that additional haplotype-specific influences, other than repeat length alone, caused by *cis*-acting properties of the triplet repeat tandem array may also lead to variable mutation rates.

Studies of the FMR1 triplet repeat have shown that regions of alleles with perfect runs of a CGG triplet repeats, uninterrupted by AGG repeat unit variants and longer than a certain threshold, are much more likely to undergo expansion than regions with shorter uninterrupted runs of CGG repeats. (Kunst & Warren, 1994). This phenomenon results in polarity of variation at the FMR1 triplet repeat. In contrast to the MVR-mapped minisatellites, most length variation is found at the end of these alleles with least internal repeat unit type variation and longer uninterrupted runs of perfect CGG repeats; much less length variation is seen over the remainder of these alleles, where the CGG repeat array contains interspersed AGG variants. Significantly some alleles associated with triplet repeat expansions show loss or lack of AGG repeats, implying that interspersed repeat unit variants help to maintain stability as has been suggested previously (Richards *et al.*, 1992; Richards & Sutherland, 1992). This polarity of variation is similar to that seen at the ApoB minisatellite, where replication slippage has been proposed as the major mutational mechanism (Desmarais *et al.*, 1993), but the opposite to that seen at MS31A, MS32 and MS205 where different mutation processes predominate. *Cis*-acting influences at other triplet repeats, for example SCA1, may also be explained by the same general principle (Richards & Sutherland, 1994; Kunst & Warren, 1994) and common features of repeat tract length and purity may play a role in all trinucleotide repeat diseases (Kunst & Warren, 1994; Mandel, 1994). However, others have observed rare haplotypes at increased risk of expansion despite having triplet repeats within the normal size range, suggesting jumps from normal sized alleles (MacPherson *et al.*, 1994). This

may indicate that haplotypes exist where flanking DNA has an influence on repeat instability and disease risk (Richards *et al.*, 1992;.Richards & Sutherland, 1992; Oudet *et al.*, 1993; Zhong *et al.*, 1993; MacPherson *et al.*, 1994), or that that triplet repeats are inherently less stable in such individuals due to other genetic variation, for example at the loci coding for mismatch repair enzymes (Mandel, 1994).

As with minisatellites there is no evidence of simple unequal recombination at STR loci. Examples of *de novo* mutation at dinucleotide repeats and the FRAXA and DM loci have been observed directly, and have not been accompanied by the exchange of (relatively distant) flanking markers (Kwiatkowski *et al.*, 1992; Weissenbach *et al.*, 1992; Fu *et al.*, 1991; Yu *et al.*, 1992; Shelbourne *et al.*, 1992; Harley *et al.*, 1992; Richards *et al.*, 1992). Similarly, studies of linkage disequilibrium around dinucleotide arrays suggest that, at least on a large scale, most allelic diversification occurs on fixed haplotypic frameworks (Sherrington *et al.*, 1991; Morral *et al.*, 1991). Flanking markers much closer (~80bp) to a microsatellite with a very high germline mutation rate (~0.75% per gamete) also showed no evidence of exchange, again suggesting slippage, USCE, or exchange over short conversion tracts as possible modes of mutation (Mahtani & Willard, 1993). This microsatellite is located on the proximal long arm of the X-chromosome, absolutely ruling out unequal meiotic recombination in half of these mutations which were paternally transmitted, since this region X-chromosome has no partner in male meiosis.

These observations, along with many other features of di- and trinucleotide instability, are consistent with mutation by replication slippage, and a number of models have been proposed to explain how this may occur (Kunkel, 1993; Lustig & Petes, 1993; Wells & Sinden, 1993; Richards & Sutherland, 1994). Trinucleotide repeats at several disease loci become unstable above a similar size threshold (40-50 repeats), leading to the suggestion that above this size, single-stranded nicks are likely to occur in the repeat during replication resulting in the loss or gain of a few repeats by slippage. Triplets which expand to a size where two single-strand breaks can occur in the same Okazaki fragment (~80 repeats) may be liable to massive expansion, because the repeats between these breaks are not anchored at either end by unique sequence and can therefore slip and/or slide during polymerisation, resulting in the addition of many more copies than were present in the original sequence (Richards & Sutherland, 1994).

It may be premature to conclude from the evidence outlined above that shorter arrays do not mutate by unequal exchange in the germline. Meiotic USCE or gene conversion, which can be difficult to distinguish from slippage, cannot be easily ruled out. Furthermore, we have shown that interallelic recombination mutations at minisatellites appear to leave little or no evidence of the exchanges even in very close flanking markers. The demonstration of a complex event, involving at least two small patches of interallelic exchange, near the myotonic dystrophy repeats suggests that recombinational mechanisms may indeed be involved in length change mutations at shorter repeat arrays (O' Hoy *et al.*,1993). Interestingly a DSBR model, similar to that we have proposed for length gain mutations at MS31A, MS32 and MS205, has been suggested as a possible mechanism which would be able to account for modest triplet repeat expansions (Wells & Sinden 1993). In this model triplet repeats would have to be the site of frequent chromosome breaks, providing an initiation event for recombination or repair. The association of the FRAXA locus with an *in vitro* fragile site may be significant in this respect.

Future directions

As with many important scientific developments MVR-PCR and SP-PCR have opened the door to an almost bewildering number of possible new experiments. Our investigations of minisatellite allelic variation and mutation are still in the preliminary stages, with each new result throwing up more new questions than it answers. It is therefore necessary to carefully define priorities and directions for new research, in order to answer these questions efficiently, and hopefully establish fundamental principles that give rise to predictions which can then be tested by future experiments. There are two broad avenues of investigation that arise from the work described in this thesis. These are the further characterisation of MS32, the locus for which we have most detailed information at present, and the application of the same techniques to apparently similar minisatellite loci, for example MS31A. The extension of MVR-PCR to additional loci will not only benefit potential forensic application of the technique but also enable comparative analysis between loci, hopefully revealing common properties that are the result of general mutational mechanisms.

Mutation analysis. The detailed analysis of minisatellite evolution outlined in this thesis derives from studies of allelic structure in human populations, and of changes in structure occurring during *de novo* mutations. Even at the relatively high rates found at minisatellite loci, the incidence of naturally occurring sporadic mutations detected in families is too low to allow a number of important questions relating to mutational load to be addressed. In addition too many potentially important variables will differ between cases to allow valid comparisons to be made. SP-PCR provides for the first time a precision tool which can be used to assess the relative importance of different factors and determine their influence on mutation. Initial studies in the male germline have already answered questions concerning the relationship between mutation rate and allele length, the involvement of flanking sequences and the existence of variation in instability between both alleles and individuals. This technique may also be valuable in the analysis of factors important in what now appear to be different processes involved in somatic mutation. Somatic length change mutants from normal somatic tissues, tumour DNA and cultured cell lines can all be investigated. For example, experiments with cell lines are currently in progress to assess the effects of various environmental agents that may be active in the production of minisatellite mutations (C. May, personal communication). A low resolution DNA fingerprinting study of somatic DNA from irradiated mice has already shown an increase in murine minisatellite length change mutation with radiation dose (Dubrova *et al.*, 1993). Although the direction of the size change events could not be determined, these results do demonstrate a probable mutational role for random DSBs in chromosomal DNA, since radiation is assumed cause such damage. SP-PCR should enable a more detailed analysis of these mutants. A collection of blood and semen DNA samples from individuals accidentally exposed to large doses of radiation by the Chernobyl nuclear reactor disaster has recently been obtained (Y. Dubrova, personal communication); multilocus analysis of these samples followed by SP-PCR will hopefully provide a system for the investigation of the effects of radiation on human minisatellites *in vivo*. Although SP-PCR is a very powerful technique, it brings us no closer to being able to analyse female germline mutation processes; the only practical way of obtaining minisatellite alleles undergoing mutation in the female germline is by pedigree analysis. Too few female germline mutants have been characterised in our initial surveys to draw any firm conclusions concerning female germline mutation processes. The huge numbers of families typed by SLP analysis in paternity cases provides a valuable potential source of such alleles. MS32 is unfortunately not used as a probe in paternity testing,

meaning that the ability to MVR map MS31, which is in widespread use as an SLP in forensic laboratories worldwide, will be of great importance in a comparative study of female and male germline mutation processes. To this end I have contacted several laboratories in order to obtain DNA samples from more families showing MS31 germline mutations.

MS31B 3′ flanking sequence. The immediate priority for MS31A is the further characterisation of flanking sequences, in particular at the 3′ end of the locus distal to MS31B, about which little is currently known. To this end preliminary trials of vectorette PCR amplification (Riley *et al.*, 1990) between the MS31A locus and unknown sequences in the 3′ flanking DNA have already been conducted. If 3′ flanking sequence can successfully be obtained, this will not only enable the characterisation of MS31B, but will also allow determination of 3′ MVR variability at MS31A, by reverse MVR mapping, the addition of 3′ flanking markers to MS31A allele haplotypes, and the development of SP-PCR at this locus. With a means of quantifying the mutation rates of individual alleles and identifying large numbers of *de novo* mutants for investigation, we will be able to obtain information concerning the mutation mechanisms operating at this locus much more efficiently, and in more detail, than by pedigree analysis. At the same time continued mapping of alleles from different populations will increase our view of the extent of allelic diversity at this locus and may also reveal alleles with unusual characteristics, for example the group of short Japanese alleles. If it transpires that the alleles in this group are analagous to the non-mutating 32 repeat African MS32 allele, they will provide a powerful tool for MS31A mutational analysis.

3′ sequence information may also allow the efficient PCR amplification of MS31A from primates, as was done with the primate homologues of other hypervariable human minisatellite loci (Gray & Jeffreys, 1991). These loci are presumably the ancestors of their human counterparts and can therefore provide useful information concerning minisatellite evolution. Sequencing showed that MS32 and MS1 are short, and largely monomorphic in the great apes. Preliminary results suggest that this is also the case at the ancestral MS31A locus (data not shown). SP-PCR has shown that the MS32 locus from the gorilla, has a mutation rate at least 100 fold less than man, (Jeffreys *et al.*, 1994) suggesting either that the necessary factors for tandem repeat expansion are not present in this species, or that mutation rate constancy does not extend to the shortest MS32 alleles. It would be of interest to perform a similar analysis of the gorilla MS31A locus.

The discovery of mutation rate polymorphism at MS32 by SP-PCR has already highlighted the significant involvement of flanking *cis*-acting sequences in influencing mutation at this locus and has provided an ideal, and importantly natural, experimental resource for the investigation of factors influencing mutation at this locus. The identification of these flanking DNA elements and further characterisation of the nature and extent of the mutation initiator and the effect of the O1C variant on initiator function will require detailed analysis of the flanking DNA. Attempts to isolate more distal flanking sequences at this locus is already underway (A.J. Jeffreys, personal communication), combining the techniques of a vectorette PCR walk (Riley *et al.*, 1990) out from the known sequence and long range PCR (Cheng *et al.*, 1994). This work will help to set MS32 in a wider genomic framework and may help in the identification of further flanking sequence elements that have a role in mutation.

Protein characterisation. The identification and characterisation of the protein components involved in minisatellite mutation is an important objective, and there are several strategies that we can use to achieve it. The further characterisation of minisatellite-specific DNA binding proteins that have already been described is the logical place to start. Another priority is the search for germinal proteins that interact with the O1 region, and perhaps more distal flanking sequences, to regulate mutation/conversion. This may be best approached in a manner similar to that used to isolate tandem repeat specific binding proteins, namely affinity binding of eukaryotic proteins to target sequences followed by gel retardation analysis (Collick *et al.*, 1990, 1991). Once again the study of naturally occurring variants may also provide a launching pad for further investigations. Several human diseases thought to arise from defects in DNA recombination/repair have been identified and the genes for some of these have been cloned (see Hoeijmakers & Bootsma, 1992; Bodmer *et al.*, 1994). If such pathologies are responsible for minisatellite instability, analysis of affected individuals may therefore provide a means of identifying the enzymes involved.

Modelling minisatellite mutation. An alternative approach to analysing minisatellites in the human genome is the construction of simpler model systems. Yeast is a very well characterised eukaryotic organism that is eminently suited to genetic analysis. This has already been used to obtain evidence for specific mechanisms by the isolation of predicted intermediates such as DSBs and 3′ overhangs (Cao *et al.*, 1990; Sun *et al.*, 1989, 1991a) and end-products of mutation (Strand *et al.*, 1993). Human minisatellites have been successfully integrated into yeast chromosomes (Cederberg *et al.*, 1993) and this may provide a suitable system for studying minisatellite mutation, particularly with respect to analysing meiotic events by tetrad analysis. However, although it may be easier to determine the mechanistic basis for mutation events in such a simple system the results obtained may not be reflective of the processes operating in man.

Although the systematic mutagenesis of DNA flanking the tandem repeat array, followed by sperm mutation analysis to identify specific flanking mutants that influence minisatellite mutation rate is not feasible in man, it could in principle be carried out in mice transgenic for a human minisatellite, provided that the transgene mutated by the appropriate pathway in the germline. For this reason a number of transgenic mice, into which the human MS32 locus has been incorporated, were generated to provide a tool for the analysis of *de novo* mutation, *cis*-acting sequences, genomic location and environmental factors, for example radiation, on mutation rate. Preliminary analyses of these mice have indicated that the minisatellite inserts are not randomly distributed in the mouse genome, are frequently multicopy and do not appear to mutate by the same mechanisms as in humans (Allen *et al.*, 1994, and unpublished results). SP-PCR of sperm DNA obtained following irradiation of the transgenic mice was also performed, but unfortunately gave inconclusive results as far as an effect on minisatellite mutation was concerned (M. Allen, unpublished results). However, further studies of these mice offer many potential insights into the behaviour of tandemly repeated sequences in different genetic and physical environments and constitute an ongoing area of research in this laboratory. The creation of mark II transgenic mice, containing larger minisatellite flanking sequences, and the development of a system for the culture of human spermatogenic cells are also being considered as a means of augmenting these studies (A. Collick, personal communication).

Concluding Remarks

Our present view of hypervariable human minisatellite loci is a frozen snapshot of a dynamic evolutionary process that takes place over a far longer timescale. Hopefully, further detailed analysis of allelic variability and mutation will indicate whether we are watching the birth or death of some of these loci and better enable us to probe their past histories, to identify influences shaping them at present and to predict their possible futures. A number of studies have indicated that the phenotypic variety of hypervariable minisatellites is, at least in part, a result of differences in their mutation rates and profiles. Comparisons of allelic variability and mutation at several loci have shown that the differential operation of a number of mutation processes may uniquely shape each locus. These may act either individually, or in concert with others and operate to varying extents at different loci and in different tissues, with the exact contribution of each being governed by numerous factors. Different loci will therefore display similar or differing patterns of mutation and allelic variation, depending on the relative contributions of, and levels of interplay between, these mechanisms and other influences, many of which remain unknown. As loci under investigation are further characterised and new ones discovered we will be better able to assess and evaluate the contributions of the multiple factors governing minisatellite evolution. A comparison of flanking sequences at these and other tandem repeat loci may reveal conserved elements responsible for similarities in allelic variability and mutational mechanisms such as those that we have observed. Tantalising circumstantial links between minisatellites and homologous recombination have already been noted but virtually nothing is known about the mechanistic basis of this process in the human genome or whether minisatellites are functionally involved. It would be a great reward, not to mention justification of basic scientific research, if our study of minisatellites were to provide the key that opened the door to our understanding of such a fundamental process. Although the the data provided by the investigation of different tandemly repeated loci may initially appear have no obvious direct relevance to each other, it seems that there may be areas of overlap between the diverse dynamic processes operating at different tandem repeat loci. Taken together these studies provide a wider and therefore potentially unifying view of the full spectrum of tandem repeat instability.

REFERENCES

Ali, S., Muller, C.R. and Epplen, J.T. (1986). DNA fingerprinting by oligonucleotide probes specific for simple repeats. *Hum. Genet.* **74**: 239-243.

Allen, M.J., Jeffreys, A.J., Azim-Surani, M., Barton, S., Norris, M. and Collick, A. (1994). Tandemly repeated transgenes of the human minisatellite MS32 (D1S8), with novel mouse gamma satellite integration. *Nucleic Acids Res.* **22**: 2976-2981.

Allshire, R.C., Gosden, G.R., Cross, S.H., Cranston, G., Rout, D., Sugawara, N., Szostack, J.W., Fantes, P.A. and Hastie, N.D. (1989). Telomeric repeat from *T. thermophila* cross-hybridises with human telomeres. *Nature* **332**: 656-659.

Anderson C. (1993) The genome project goes commercial. *Science* **259**: 300-302.

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

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

Armour, J.A.L. (1990). The isolation and characterization of human minisatellites. *PhD thesis, University of Leicester*: a, 134; b, 77; c, 88; d, 89; e, 76.

Armour, J.A.L., Povey, S., Jeremiah, S. and Jeffreys, A.J. (1990). Systematic cloning of human minisatellites from ordered array Charomid libraries. *Genomics* **8**: 501-512.

Armour, J.A.L. and Jeffreys, A.J. (1991). STS for minisatellite MS607 (D22S163). *Nucleic Acids Res.* **19**: 3158.

Armour, J.A.L., Vergnaud, G., Crosier, M. and Jeffreys, A.J. (1992a). Isolation of human minisatellite loci by synthetic tandem repeat probes: direct comparison with cloned DNA fingerprinting probes. *Hum. Mol. Genet.* **1**: 319-323.

Armour, J.A.L., Crosier, M. and Jeffreys, A.J. (1992b) Human minisatellite alleles detectable only after PCR amplification. *Genomics* **12**: 116-124.

Armour, J.A.L., Harris, P.C. and Jeffrreys, A.J. (1993). Allelic diversity at minisatellite MS205 (D16S309): evidence for polarised variability. *Hum. Mol. Genet.* **2**: 1137-1154.

Armour, J.A.L., Crosier, M., Malcolm, S., Chan, J. and Jeffreys, A.J. (1994). Human triplet repeat minisatellites (manuscript submitted).

Arnheim, N., Krystal, M., Schmikiel, R., Wilson, G., Ryder, O. and Zimmer, E. (1980). Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. *Proc. Natl. Acad. Sci. USA.* **77**: 7323-7327.

Arnot, D.E., Roper, C. and Bayoumi, R.A.L. (1993). Digital codes from hypervariable tandemly repeated DNA sequences in the *Plasmodium falciparum* circumsporozoite gene can genetically barcode isolates. *Mol. Biochem. Parasitol.* **61**: 15-24.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K. (eds). (1994). *Current protocols in molecular biology*. Wiley Interscience.

Azen, E., Lyons, K.M., McGonigal, T., Barrett, N.L., Clements, L.S., Maeda, N., Vanin, E.F., Carlson, D.M. and Smithies, O. (1984). Clones from the human gene complex coding for salivary proline rich proteins. *Proc. Natl. Acad. Sci. USA.* **81**: 5561-5565.

Balazs, I. (1993). Population genetics of 14 ethnic groups using phenotypic data from VNTR loci. In. *DNA fingerprinting: State of the science*. Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. (eds). Birkhaüser Verlag, Basel: 193-210.

Balding, D.J and Donnelly. (1994). How convincing is DNA evidence? *Nature* **368**: 285-286.

Baltimore, D. (1981). Gene conversion: Some implications for immunoglobulin genes. *Cell* **24**: 592-594.

Banchs, I., Bosch, A., Guimera, J., Lazaro, C., Puig, A. and Estivill, X. (1994). New alleles at microsatellite loci in CEPH families mainly arise from somatic mutations in the lymphoblastoid cell-lines. *Hum. Mut.* **3**: 365-372.

Beckman, J.S. and Weber, J.L. (1992). Survey of human and rat microsatellites. *Genomics* **12**: 627-631.

Bell, G.I., Selby, M.J. and Rutter, W.I. (1982). The highly polymorphic region near the human insulin gene is composed of simple tandemly repeated sequences. *Nature* **295**: 31-35.

Bellamy, R.J., Inglehaerm, C.F., Jalili, I.K., Jeffreys, A.J. and Bhattacharya, S.S. (1991). Increased band sharing in DNA fingerprints of an inbred human population. *Hum.Genet.* **87**: 341-347.

Berg, E.S., Puers, C., Skowasch, K., Wiegand, P., Budowle, B. and Brinkmann, B. (1993). Characterization of the COL2A1 VNTR polymorphism. *Genomics* **16**: 350-354.

Blackburn, E.H. (1991). Structure and function of telomeres. *Nature* **350**: 569-573.

Blake, E., Mihalovich, J., Higuchi, R., Walsh, S. and Erlich, H.A. (1992). Polymerase chain reaction (PCR) amplification and human leukocyte antigen (HLA)-DQα oligonucleotide typing on biological samples: casework experience. *J. For. Sci.* **37**: 700-726.

Bodmer, W.F. (1981). HLA structure and function: a contempory view. *Tissue antigens* **17**: 9-20.

Bodmer, W.F., Bishop, T. and Karren, P. (1994). Genetic steps in colorectal cancer. *Nature Genet.* **6**: 217-219.

Boerwinkle, E., Xiong, W., Fourest, E. and Chan, L. (1989). Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: Application to the Apolipoprotein B 3′ hypervariable region. *Proc. Natl. Acad. Sci. USA.* **86**: 212-216.

Bowcock, A., Osborne-Lawrence, S., Barnes, R., Chakravarti, A., Washington, S. and Dunn, C. (1993). Microsatellite polymorphism linkage map of human chromosome 13q. *Genomics* **15**: 376-386.

Boylan, K.B., Ayres, T.M., Popko, B., Takahashi, N., Hood, L.E. and Prusiner, S.B. (1990) Repetitive DNA $(TGGA)_n$ 5′ to the myelin basic protein gene: A new form of oligonucleotide repetitive sequence showing polymorphism. *Genomics* **6**: 16-22.

Braman, J., Barker, D., Schumm, J., Knowlton, R. and Donis-Keller, H. (1985). Characterization of very highly polymorphic RFLP probes. *Cytogenet. Cell Genet.* **40**: 589.

Britten, R.J. and Kohne, D.E. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161**: 529-540.

Britten, R.J., Baron, W.F., Stout, D.B. and Davidson, E.H. (1988). Sources and evolution of human Alu repeated sequences. *Proc. Natl. Acad. Sci. USA.* **85**: 4770-4774.

Brook, J.D., McCurragh, M.E., Harley, H.G., Buckler, A.J., Church, D., Aburatani, H., Hunter, K., Stanton, V.P., Thirion, J-P., Hudson, T., Sohn, R., Zemelman, B., Snell, R.G., Rundle, S.A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P.S., Shaw, D.J. and Housman, D. (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell* **68**: 799-808.

Brookfield, J. (1992). Law and probabilities. *Nature* **355**: 207-208.

Brutlag, D., Fry, K., Nelson, T. and Hung, P. (1977). Synthesis of hybrid bacterial plasmids containing highly repeated satellite DNA. *Cell* **10**: 509-519.

Buard, J. and Vergnaud, G. (1994) Complex recombination events at the hypermutable minisatellite CEB1 (DSS90). *EMBO. J.* **13**: 3203-3210.

Budowle, B., Giusti, A.M., Waye, J.S., Baechtel, F.S., Fourney, R.M., Adams, D.E., Presley, L.A., Deadman, H.A. and Monson, K.L. (1991a). Fixed-bin analysis for stastistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. *Am. J. Hum. Genet.* **48**: 841-855.

Budowle, B., Chakraborty, R., Giusti, A.M., Eisenberg, A.J. and Allen, R.C. (1991b). Analysis of the VNTR locus D1S80 by the PCR followed by high resolution PAGE. *Am. J. Hum. Genet.* **48**: 137-144.

Budowle, B. (1993). VNTR population data from various reference groups and the significance of application to identity testing. In. *DNA fingerprinting: State of the science.* Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. (eds). Birkhaüser Verlag, Basel: 177-192.

Buetow, K.H., Weber, J.L., Ludwigsen, S., Scherpbier-Heddema, T., Duyk, G.M., Sheffield, V.C., Wang, Z. and Murray, J. (1994). Integrated human genome-wide maps constructed using the CEPH reference panel. *Nature Genet.* **6**: 391-393.

Burke, T. and Bruford, M.D. (1987). DNA fingerprinting in birds. *Nature* **327**: 149-152.

Burke, T., Hanotte, O., Bruford, M.W. and Cairns, E. (1991). Multilocus and single locus minisatellites analysis in population biological studies. In. *DNA fingerprinting: approaches and applications.* Burke, T., Dolf. G., Jeffreys, A.J. and Wolff, R. (eds). Birkhaüser Verlag, Basel: 154-168.

Cao, L., Alani, E. and Kleckner, N. (1990). A pathway for generation and processing of double-strand breaks during meiotic recombination in *S. cerevisiae*. *Cell* **61**: 1089-1101.

Cann, R.L., Stoneking, M. and Wilson, A.C. (1987). Mitochondrial DNA and human evolution. *Nature* **325**: 31-36.

Capon, D.J., Chen, E.Y., Levinson, A.D., Seeberg, P.H. and Goeddel, D.V. (1983). Complete nucleotide sequence of the T24 human bladder carcinoma oncogene and its normal homologue. *Nature*. **302**: 33-37.

Carpenter, A.T.C. (1987). Gene conversion, recombination nodules, and the initiation of meiotic synapsis. *BioEssays* **6**: 232-236.

Caskey, C.T., Pizzuti, A., Fu, Y-H., Fenwick, R.G. Jr. and Nelson, D.L. (1992). Triplet repeat mutations in human disease. *Science* **256**: 784-788.

Cederberg, H., Agurell, E., Hedenskog, M. and Rannug, U. (1993). Amplification and loss of repeat units of the human minisatellite MS1 integrated in chromosome-III of a haploid yeast-strain. *Mol. Gen. Genet.* **238**: 38-42.

Chakraborty, R. and Jin, L. (1992). Heterozygote defficiency, population substructure and their implications in DNA fingerprinting. *Hum. Genet.* **88**: 267-272.

Chakraborty, R. and Jin, L. (1993) A unified approach to study hypervariable polymorphisms: Statistical considerations of determining relatedness and population distances. In. *DNA fingerprinting: State of the science*. Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. (eds). Birkhaüser Verlag, Basel: 153-176.

Chakraborty, R. and Kidd, K.K. (1991). The utility of DNA typing in forensic work. *Science* **254**: 1735-1739.

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

Chase, J.W. and Williams, K.R. (1986). Single-stranded-DNA binding-proteins required for DNA replication. *Ann. Rev. Biochem.* **55**: 103-136.

Chen, H.M., Kalaitsidaki, M., Warren, A.C., Avramopoulos, D. and Antonarakis, S.E. (1993). A novel zinc-finger cDNA with a polymorphic pentanucleotide repeat $(ATTTT)_n$ maps on human chromosome-19P. *Genomics* **15**: 621-625.

Cheng, S., Fockler, C., Barnes, W.M. and Higuchi, R. (1994). Effective amplification of long targets from cloned inserts and human genomic DNA. *Proc. Natl. Acad. Sci. USA*. **91**: 5695-5699.

Coghlan, A. (1994). Noiseless robot speeds genome project. *New Scientist* **141**: 20.

Cohen, J.E. (1990). DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number variability. *Am. J. Hum. Genet.* **46**: 358-368.

Collick, A. and Jeffreys, A.J. (1990). Detection of a novel minisatellite-specific DNA-binding protein. *Nucleic Acids Res.* **18**: 2256-2266.

Collick, A., Dunn, M.G. and Jeffreys, A.J. (1991). Minisatellite binding protein Msbp-1 is a sequence-specific single-stranded DNA-binding protein. *Nucleic Acids Res.* **19**: 6399-6404.

Collins, F. and Galas, D. (1993) A new five-year plan for the human genome project. *Science* **262**: 43-46.

Comey, T.C., Budowle, B., Adams, D.E., Baumstark, A.L., Lindsey, J.A. and Presley, L.A. (1993). Amplification and typing of the HLA-DQα gene in forensic samples. *J. For. Sci.* **38**: 239-249.

Consalez, G.G., Thomas, N.S.T., Stayton, C.L., Knight, S.J.L., Johnson, M., Hopkins, L.C., Harper, P.S., Elsas, L.J. and Warren, S.T. (1991). Assignment of Emery-Dreyfuss muscular dystrophy to the distal region of Xq28-The results of a collaborative study. *Am. J. Hum. Genet.* **48**: 468-480.

Cooke, H.J., Brown, W.R.A. and Rappold, G.A. (1985). Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. *Nature* **317**: 687-692.

Costa, R., Peixoto, A.A., Barbujani, G. and Kyriacou, C.P. (1992). A latitudinal cline in a Drosophila clock gene. *Proc. R. Soc. Lond. B.* **250**: 43-49.

Dallas, J.F. (1988). Detection of DNA fingerprints of cultivated rice by hybridization with a human minisatellite probe. *Proc. Natl. Acad. Sci. USA.* **85**: 6831-6835.

Desmarais, E., Vigneron, S., Buresi, C., Cambien, F., Cambou, J.P. and Roizes, G. (1993). Variant mapping of the Apo(B) AT rich minisatellite. Dependence on nucleotide sequence of the copy number variations. Instability of the non-canonical alleles. *Nucleic Acids. Res.* **21**: 2179-2184.

Devereux, J., Haeberli, P. and Smithies, O. (1984). A comprehensive set of sequence analysis programmes for the VAX. *Nucleic Acids Res.* **12**: 387-395.

Devlin, B., Risch, N. and Roeder, K. (1990). No excess of homozygosity at loci used for DNA fingerprinting. *Science* **249**: 1416-1420.

Devlin, B. and Risch, N. (1992). A note on Hardy-Weinberg equilibrium of VNTR data by using the Federal Bureau of Investigations fixed-bin method. *Am. J. Hum. Genet.* **51**: 549-553.

Dietrich, W., Katz, H., Lincoln, S.E., Shin, H.S., Friedman, J. and Dracopoli, W.C. (1992). A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* **131**: 421-445.

Dixon, A.F., Anzenberger, G., Monteiro Da Cruz, M.A.O., Patel, I. and Jeffreys, A.J. (1992). DNA fingerprinting of free ranging groups of common marmosets (*Callithrix jacchus jacchus*) in NE Brazil. In: *Paternity in primates: Genetic tests and theories*. Martin, R.D., Dixon, A.F. and Wickings, E.J. (eds). Karger, Basel: 192-202.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., Lander, E.S., Botstien, D., Akots, G., Rediker, K.S., Gravius, T., Brown, V.A., Rising, M.B., Parker, C., Powers, J.A., Watt, D.E., Kaufman, E.K., Briker, A., Phipps, R., Muller-Khale, H., Fulton, T.R., Ng, S., Schumm, J.W., Braman, J.C., Knowlton, R.G., Barker, D.F., Crooks, S.M., Lincoln, S.E., Daly, M.J. and Abrahamson, J. (1987). A genetic linkage map of the human genome. *Cell* **51**: 319-337.

Dover, G. (1982) Molecular drive: a cohesive mode of species evolution. *Nature* **299**: 111-117.

Dover, G. (1989). Victims or perpetrators of DNA turnover? *Nature* **342**: 347-348.

Dressler, D. and Potter, H. (1982). Molecular mechanisms of genetic-recombination. *Ann. Rev. Biochem.* **51**: 727-761.

Dubrova, Y., Jeffreys, A.J. and Malashenko, A.M. (1993). Mouse minisatellite mutations induced by ionizing radiation. *Nature Genet.* **5**: 92-94.

Dunlap, J.C. (1993). Genetic-analysis of circadian clocks. *Ann. Rev. Physiol.* **55**: 683-728.

Economou, E.P., Bergen, A.W., Warren, A.C. and Antonarakis, S.E. (1990). The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. *Proc. Nat. Acad. Sci. USA.* **87**: 2951-2954.

Edwards, A., Civitello, A., Hammond, H.A. and Caskey, C.T. (1991). DNA Typing and Genetic Mapping with Trimeric and Tetrameric Tandem Repeats *Am. J. Hum. Genet.* **49**: 746-756.

Edwards, A., Hammond, H.A., Jin, L., Caskey, C.T. and Chakraborty, R. (1992). Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* **12**: 241-253.

Edwards, M.C., Clemens, P.R., Tristan, M., Pizzuti, A. and Gibbs, R.A. (1991). Pentanucleotide repeat length polymorphism at the human CD4 locus. *Nucleic Acids Res.* **19**: 4791.

Epplen, J.T., McCarrey, J.R., Sutou, S. and Ohno, S. (1982). Base sequence of a cloned snake W-chromosome DNA fragment and identification of a male-specific putative mRNA in the mouse. *Proc. Natl. Acad. USA.* **79**: 3798-3802.

Epplen, J.T., Ammer, H., Epplen, C., Kammerbauer, C., Mitreiter, R., Roewer, L., Schwaiger, W., Steimle, V., Zischler, H., Albert, E., Andreas, A., Beyermann, B., Meyer, W., Buitkamp, J., Nanda, I., Schmid, M., Nurnberg, P., Pena, S.D.J., Poche, H., Sprecher, W., Schartl, M., Weising, K. and Yassouridis, A. (1991). Oligonucleotide fingerprinting using simple repeat motifs: A convenient way to detect hypervariability for multiple purposes. In. *DNA fingerprinting: approaches and applications*. Burke, T., Dolf. G., Jeffreys, A.J. and Wolff, R. (eds). Birkhaüser Verlag, Basel: 50-67.

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**: 87-112.

Faik, P., Walker J.I.H. and Morgan, M.J. (1994). Identification of a novel tandemly repeated sequence present in an intron of the glucose phosphate isomerase (GPI) gene in mouse and man. *Genomics* **21**: 122-127.

Feinberg, A.P. and Vogelstein, B. (1984). A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **137**: 266-267.

Fowler, S.J., Gill, P., Werrett, D.J. and Higgs, D.R. (1988). Individual specific DNA fingerprints from a hypervariable region probe: alpha-globin 3′ HVR. *Hum. Genet.* **79**: 142-146.

Fraser, N.J., Boyd, Y. and Craig, I. (1989). Isolation and characterization of a human variable copy number tandem repeat at Xcen-p11.22. *Genomics* **5**: 144-148.

Fu, Y-H., Kuhl, D.P.A., Pizzuti, A., Pieretti, M., Sutcliffe, J.S., Richards, S., Verkerk, A.J.M.H., Holden, J.J.A., Fenwick, R.G. Jr., Warren, S.T., Oostra, B.A., Nelson, D.L. and Caskey, C.T. (1991). Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* **67**: 1047-1058.

Ge, Y., Wagner, M.J., Siciliano, M. and Wells, D.E. (1992). Sequence, higher order repeat structure, and long range organization of alpha-satellite DNA specific to human chromosome 8. *Genomics* **13**: 585-593.

Georges, M.A.J., Gunawardana, A., Threadgill, D.P., Lathrop, M., Olsaker, I., Mishra, A., Sargeant, L.L., Schoeberlien, A., Steele, M.R., Terry, C., Threadgill, D.S., Zhao, X., Holm, T., Fries, R. and Womack, J. (1991). Characterization of a set of variable number of tandem repeat markers conserved in *bovidae*. *Genomics* **1**: 24-32.

Gendler, S.J., Lancaster, C.A., Taylor-Papadimitrou, J., Duhig, T., Peat, N., Burchell, J., Pemberton, L., Lalani, E-N. and Wilson, D. (1990). Molecular cloning and expression of human tumor-associated polymorphic epithelial mucin. *J. Biol. Chem.* **265**: 15286-15293.

Gibbs, M., Collick, A., Kelly, R.G. and Jeffreys, A.J. (1993). A tetranucleotide repeat mouse minisatellite displaying substantial somatic instability during early preimplantation development. *Genomics* **17**: 121-128.

Gill, P., Jeffreys, A.J. and Werrett, D.J. (1985). Forensic application of DNA fingerprints. *Nature* **318**: 577-579.

Gill, P. and Werret, D.J. (1987). Exclusion of a man charged with murder by DNA fingerprinting. *For. Sci. Int.* **35**: 145-148.

Gill, P., Woodroffe, S., Bar, W., Brinkmann, B., Carracedo, A., Eriksen, B., Jones, S., Kloostermann, A.D., Ludes, B., Mevag, B., Pascali, V.L., Rudler, M., Schmitter, H., Schneider, P.M. and Thompson, J.A. (1992). A report of an international collaborative experiment to demonstrate the uniformity obtainable using DNA profiling techniques. *For. Sci. Int.* **53**: 29-43.

Gill, P., Ivanov, P.L., Kimpton, C., Piercy, R., Benson, N., Tully, G., Evett, I., Hagelberg, E. and Sullivan, K. (1994). Identification of the remains of the Romanov family by DNA analysis. *Nature Genet.* **6**: 130-135.

Ginther, C., Issel-Tarver, L. and King, M-C. (1992). Identifying individuals by sequencing mitochondrial DNA from teeth. *Nature Genet.* **2**: 135-138.

Goodbourne, S.E.Y., Higgs, D.R., Clegg, J.B. and Weatherall, D.J. (1983). Molecular basis of length polymorphism in the human ζ-globin gene complex. *Proc. Natl. Acad. Sci. USA.* **80**: 5022-5026.

Gray, I.C. (1991). Polymorphic tandemly repeated sequences in human DNA. *PhD Thesis, University of Leicester*: a, 80; b, 60; c, 79; d, 75; e, 85.

Gray, I.C. and Jeffreys, A.J. (1991). Evolutionary transience of hypervariable minisatellites in man and the primates. *Proc. R. Soc. Lond. B.* **243**: 241-253.

Greenberg, B.D., Newbold, J.E. and Sugino, A. (1983). Intraspecific nucleotide-sequence variability surrounding the origin of replication in human mitochondrial DNA. *Gene* **21**: 33-49.

Greider, C.W. and Blackburn, E.H. (1989). A telomeric sequence in the RNA of *Tetrahymena* telomerase required for telomere repeat synthesis. *Nature* **337**: 331-337.

Greig, G.M. and Willard, H.F. (1992). β-satellite DNA: characterisation and localisation of two subfamilies from the distal and proximal short arms of the human acrocentric chromosomes. *Genomics* **12**: 573-580.

Gum, J.R., Byrd, J.C., Hicks, J.W., Toribara, N.W., Lambport, D.T.A. and Kim, Y.S. (1989). Molecular cloning of human intestinal mucin cDNAs. *J. Biol. Chem.* **264**: 6480-6487.

Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernadi, G., Lathrop, M. and Wiessenbach, J. (1994). The 1993-1994 Généthon human genetic linkage map. *Nature Genet.* **7**: 246-339.

Haaf, T. and Willard, H.F. (1992). Organization, polymorphism, and molecular cytogenetics of chromosome-specific α-satellite DNA from the centromere of chromosome 2. *Genomics* **13**: 122-128.

Hagelberg, E., Sykes, B. and Hedges, R. (1989). Ancient bone DNA amplified. *Nature* **342**: 485.

Hagelberg, E., Gray, I.C. and Jeffreys, A.J. (1991). Identification of the skeletal remains of a murder victim by DNA analysis. *Nature* **352**: 427-429.

Hamada, H. and Kakunaga, T. (1982). Potential Z-DNA forming sequences are widely dispersed in the human genome. *Nature* **298**: 396-398.

Hamada, H., Petrino, M.G. and Kakunaga, T. (1982). A novel repeated element with Z-DNA forming potential is widely found in evolutionarily diverse eukaryotic genomes. *Proc. Natl. Acad. Sci. USA.* **79**: 6465-6469.

Hammond, K.M., Dobrinski, B., Lurz, R., Docherty, K. and Kilpatrick, M.W. (1992). The human insulin gene linked to a polymorphic region exists in an altered DNA structure. *Nucleic Acids Res.* **20**: 231-236.

Hanote, O., Burke, T., Armour, J.A. L. and Jeffreys, A.J. (1991). Hypervariable minisatellite DNA sequences in the Indian peafowl *Pavo cristatus*. *Genomics* **9**: 587-597.

Harada, S., Nakamura, T. and Misawa, S. (1994). Polymorphism of pentanucleotide repeats in the 5′ flanking region of the glutathione-S-transferase (GST) pi-gene. *Hum. Genet.* **93**: 223-224.

Harding, R.M., Boyce, A.J. and Clegg, J.B. (1992). The evolution of tandemly repetitive DNA: recombination rules. *Genetics* **132**: 847-859.

Harley, C.B., Futcher, A.B. and Greider, C.W. (1990) Telomeres shorten during ageing of human fibroblasts. *Nature* **345**: 458-460.

Harley, H.G., Brook, J.D., Rundle, S.A., Crow, S., Reardon, W., Buckler, A.J., Harper, P.S., Housman, D.E. and Shaw, D.J. (1992). Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. *Nature* **355**: 545-546.

Hastie, N.D., Dempster, M., Dunlop, M.G., Thompson, A.M., Green, D.K. and Allshire, R.C. (1990). Telomere reduction in human colorectal carcinoma and with ageing. *Nature* **346**: 866-868.

Hawkins, J.R., Superti-Furga, A., Steinmann, B. and Dalgleish, R. (1991). A 9-base pair deletion in COL1A1 in a lethal variant of Osteogenesis Imperfecta. *J. Biol. Chem.* **266**: 22370-22374.

Heintz, N., Zernik, M. and Roeder, R.G. (1981). The structure of human histone genes: Clustered but not tandemly repeated. *Cell* **24**: 661-668.

Helmuth, R., Fildes, N., Blake, E., Luce, M.C., Chimera, J., Madej, R., Gorodezky, C., Stoneking, M., Schmill, N., Klitz, W., Higuchi, R. and Erlich, H.A. (1990). HLA-DQα allele and genotype frequencies in various human populations, determined by using enzymatic amplification and oligonucleotide probes. *Am. J. Hum. Genet.* **47**: 515-523.

Henke, J., Fimmers, R., Baur, M.P. and Henke, L. (1993). DNA-minisatellite mutations: recent investigations concerning distribution and impact on parentage testing. *J. Leg. Med.* **105**: 217-222.

Hershey, A.D. and Chase, M. (1952). Independent function of viral protein and nucleic acid on growth of bacteriophage. *J. Gen. Physiol.* **36**: 39-56.

Higgs, D.R., Goodbourne, S.E.Y., Wainscoat, J.S., Clegg, J.B. and Weatherall, D.J. (1981) Highly variable regions of DNA flank the human alpha-globin genes. *Nucleic Acids Res.* **9**: 4213-4224.

Higuchi, R. and Blake, E.T. (1989). Applications of the polymerase chain reaction in forensic science. In. *Banbury Report* **32**: *DNA Technology, Forensic Science*. Ballantyne, J., Sensabaugh, B. and Witkowski, J. (eds). Coldspring Harbour Laboratory Press, New York: 265-281.

Hoeijmakers, J.H.J. and Bootsma, D. (1992). DNA repair: two pieces of the puzzle. *Nature Genet.* **1**: 313-314.

Hopgood, R., Sullivan, K.M. and Gill, P. (1992). Strategies for automated sequencing of human mitochondrial DNA directly from PCR products. *BioTechniques* **13**: 82-92.

Horn, G.T., Richards, B. and Klinger, K.W. (1989). Amplification of a highly polymorphic VNTR segment by the polymerase chain reaction. *Nucleic Acids Res.* **17**: 2140.

Huang, L.S. and Breslow, J.L. (1987). A unique AT-rich hyper-variable minisatellite 3′ to the ApoB gene defines a high information restriction fragment length polymorphism. *J. Biol. Chem.* **262**: 8952-8955.

Hulten, M. 1974. Chiasma distribution at diakinesis in the normal human male. *Hereditas* **76**: 55-78.

Hung, T., Mak, K. and Fong, K. (1990). A specificity enhancer for polymerase chain reaction. *Nucleic Acids Res.* **18**: 4953.

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

Hutchinson, R.M., Pringle, J.H., Potter, L., Patel, I. and Jeffreys, A.J. (1989). Rapid identification of donor and recipient cells after allogenic bone marrow transplantation using specific genetic markers. *Br. J. Haematol.* **72**: 133-140.

Imbert, G., Kretz, C., Johnson, K. and Mandel, J.-L. (1993). Origin of the expansion mutation in myotonic dystrophy. *Nature Genet.* **4**: 72-76.

Ingram, V.M. (1957). Gene mutations in human hemoglobin: The chemical difference between normal and sickle cell hemoglobin. *Nature* **180**: 326-328.

Jarman, A.P. and Higgs, D.R. (1988) A new hypervariable marker for the human alpha-globin gene-cluster. *Am. J. Hum. Genet.* **43**: 249-256.

Jarman, A.P. and Wells, R.A. (1989). Hypervariable minisatellites: recombinators or innocent bystanders? *Trends Genet.* **5**: 367-371.

Jarman, A.P., Nicholls, R.D., Weatherall, D.J., Clegg, J.B. and Higgs, D.R. (1986). Molecular characterization of a hypervariable region downstream of the human α-globin gene cluster. *EMBO. J.* **5**: 1857-1863.

Jeffreys, A.J. and Flavell, R.A. (1977). A physical map of the DNA regions flanking the rabbit β-globin gene. *Cell* **12**: 429-439.

Jeffreys, A.J. (1979). DNA sequence variation in the Gγ, Aγ, δ and β globin genes of man. *Cell* **18**: 1-10.

Jeffreys, A.J., Wilson, V. and Thein, S.L. (1985a). Hypervariable "minisatellite" regions in human DNA. *Nature* **314**: 67-73.

Jeffreys, A.J., Wilson, V. and Thein, S.L. (1985b). Individual-specific "fingerprints" of human DNA. *Nature* **316**: 76-79.

Jeffreys, A.J., Brookfield, J.F.Y. and Semeonoff, R. (1985c). Positive identification of an immigration test case using human DNA fingerprints. *Nature* **317**: 818-819.

Jeffreys, A.J., Wilson, V., Thein, S.L., Weatherall, D.J. and Ponder, B.A.J. (1986). DNA "fingerprints" and segregation of multiple markers in human pedigrees. *Am. J. Hum. Genet.* **39**: 11-24.

Jeffreys, A.J. and Morton, D.B. (1987). DNA fingerprints of dogs and cats. *Anim. Genet.* **18**: 1-15.

Jeffreys, A.J., Wilson, V., Kelly, R., Taylor, B. and Bulfield, G. (1987a). Mouse "DNA fingerprints": Analysis of chromosomal location and germline stability of hypervariable loci in recombinant inbred strains. *Nucleic Acids Res.* **15**: 2823-2836.

Jeffreys, A.J., Wilson, V., Wong, Z., Royle, N., Patel, I., Kelly, R. and Clarkson, R. (1987b). Highly variable minisatellites and DNA fingerprints. *Biochem. Soc. Symp.* **53**: 165-180.

Jeffreys, A.J., Royle, N.J., Wilson, V. and Wong, Z. (1988a). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332**: 278-281.

Jeffreys, A.J., Wilson, V., Neumann, R. and Keyte, J. (1988b). Amplification of human minisatellites by the polymerase chain reaction: towards DNA fingerprinting of single cells. *Nucleic Acids Res.* **16**: 10953-10971.

Jeffreys, A.J., Neumann, R. and Wilson, V. (1990). Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* **60**: 473-485.

Jeffreys, A.J. (1991). Advances in forensic science: applications and implications of DNA testing. Science in Parliament **48**: 2-7

Jeffreys, A.J., MacLeod, A., Tamaki, K., Neil, D.L. and Monckton, D.G. (1991a). Minisatellite repeat coding as a digital approach to DNA typing. *Nature*. **354**: 204-209.

Jeffreys, A.J., Turner, M. and Debenham, P. (1991b). The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships, determined from extensive casework. *Am. J. Hum. Genet.* **48**: 824-840.

Jeffreys, A.J., Royle, N., Patel, I., Armour, J.A.L., Macleod, A., Collick, A., Gray, I., Neumann, R., Gibbs, M., Crosier, M., Hill, M. and Signer, E. (1991c). Principles and recent advances in DNA fingerprinting. In. *DNA fingerprinting: approaches and applications*. Burke, T., Dolf. G., Jeffreys, A.J. and Wolff, R. (eds). Birkhaüser Verlag, Basel: 3-19.

Jeffreys, A.J., Allen, M.J., Hagelberg, E. and Sonnberg, A. (1992). Identification of the skeletal remains of Josef Mengele by DNA analysis. *For. Sci. Int.* **56**: 65-76.

Jeffreys, A.J. and Pena, S.J.D. (1993). Brief introduction to human DNA fingerprinting. In. *DNA fingerprinting: State of the science*. Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. (eds). Birkhaüser Verlag Basel: 1-20.

Jeffreys, A.J., Monckton, D.G., Tamaki, T., Neil, D.L., Armour, J.A.L., MacLeod, A., Collick, A., Allen, M. and Jobling, M. (1993). Minisatellite variant repeat mapping: Application to DNA typing and mutation analysis. In. *DNA fingerprinting: State of the science*. Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. (eds). Birkhaüser Verlag Basel: 125-139.

Jeffreys, A.J., Tamaki, T., MacLeod, A., Monckton, D.G., Neil, D.L. and Armour, J.A.L. (1994). Complex gene conversion events in germline mutation at human minisatellites. *Nature Genet.* **6**: 136-145.

Jones, K.W. and Corneo, G. (1971). Location of satellite and homogenous DNA sequences on human chromosomes. *Nature New Biology* **233**: 268-271.

Jurka, J. (1990). Novel families of interspersed repetitive elements from the human genome. *Nucleic Acids Res.* **18**: 137-141.

Kan, Y.W. and Dozy, A.M. (1978). Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. *Proc. Natl. Acad. Sci. USA.* **75**: 5631-5635.

Kasperczyk, A., Kimartino, N.A. and Krontiris, T.G. (1990). Minisatellite allele diversification: the origin of rare alleles at the HRAS1 locus. *Am. J. Hum. Genet.* **47**: 854-859.

Kay, M.A. and Woo, S.L.C. (1994). Gene-therapy for metabolic disorders. *Trends Genet.* **10**: 253-257.

Kelly, R., Bulfield, G., Collick, A., Gibbs, M. and Jeffreys, A.J. (1989). Characterization of a highly unstable mouse minisatellite locus: evidence for somatic mutation during early development. *Genomics* **5**: 844-856.

Kelly, R., Gibbs, M., Collick, A. and Jeffreys, A.J. (1991). Spontaneous mutation at the hypervariable mouse minisatellite locus Ms6-hm: flanking DNA sequence and analysis of germline and early somatic mutation events. *Proc. R. Soc. Lond. B* **245**: 235-245

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M. and Tsui, L.C. (1989). Identification of the Cystic Fibrosis gene: genetic analysis. *Science* **245**: 1073-1080.

Kit, S. (1961). Equilibrium sedimentations in density gradients of DNA preparations from animal tissues. *J. Mol. Biol.* **3**: 711-716.

Koop, B.F., Goodman, M., Xu, P., Chan, K. and Slighthom, J.L. (1986). Primate η-globin DNA sequences and man's place among the great apes. *Nature* **319**: 234-238.

Kremer, E.J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S.T., Schlessinger, D., Sutherland, G.R. and Richards, R.I. (1991). Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence $p(CCG)_n$. *Science* **252**: 1711-1714.

Krontiris, T.G. and Green, M. (1993). Allelic variation of reporter gene activation by the HRAS1 minisatellite. *Genomics* **17**: 429-434.

Krontiris, T.G., Devlin, B., Karp, D.D., Robert, N.J. and Risch, N. (1993). An association between the risk of cancer and mutations in the HRAS1 minisatellite locus. *New Engl. J. Med.* **329**: 517-5233.

Kunkel, T.A. (1993). Slippery DNA and diseases. *Nature* **365**: 207-208.

Kunst, C.B. and Warren, S.T. (1994). Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. *Cell* **77**: 853-861.

Kwiatkowski, D.J., Henske, E.P., Weimer, K., Ozelius, L., Gusella, J.F. and Haines, J. (1992). Construction of a GT polymorphism map of human 9q. *Genomics* **12**: 229-239.

Labella, T. and Schlessinger, D. (1989). Complete human rDNA repeat units isolated in yeast artificial chromosomes. *Genomics* **5**: 752-760.

Lander, E.S. (1989). DNA fingerprinting on trial. *Nature* **339**: 501-505.

Lander, E.S. (1991). Invited editorial: Research on DNA typing catching up with courtroom application. *Am. J. Hum. Genet.* **48**: 819-823.

Landsteiner, K. (1900). Zurkenntnis der antifermentatien, lytischen und agglutinierenden wirkungen den blutserums und der lymphe. *Zentralbl. Bakteriol.* **27**: 357-362.

Lathrop, M., Nakamura, Y., O'Connell, P., Leppert, M., Woodward, S., Lalouel, J.M. and White, R. (1988). A mapped set of genetic markers for human chromosome-9. *Genomics* **3**: 361-366.

Laurie, D.A. and Hulten, M.A. (1985). Further studies on chiasma distribution and interference in the human male. *Ann. Hum. Genet.* **49**: 203-214.

Levinson, G. and Gutman, G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**: 203-221.

Lewontin, R.C. and Hartl, D.L. (1991). Population genetics in forensic DNA typing. *Science* **254**: 1745-1750.

Li, S-H., McInnis, M.G., Margolis, R.L., Antonarakis, S.E. and Ross, C.A. (1993). Novel triplet repeat containing genes in human brain: Cloning, expression, and length polymorphisms. *Genomics* **16**: 572-579.

Lin, F-L., Sperle, K. and Sternberg, N. (1984). Model for homologous recombination during transfection of DNA into mouse L cells: role for DNA ends in the recombination process. *Mol. Cell. Biol.* **4**: 1020-1034.

Litt, M. and Luty, J.A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a nucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44**: 397-401.

Lucassen, A.M., Julier, C., Beressi, J-P., Boitard, C., Froguel, P., Lathrop, M. and Bell, J.I. (1993). Susceptibility to insulin dependent diabetes mellitus maps to a 4.1kb segment of DNA spanning the insulin gene and associated VNTR. *Nature Genet.* **4**: 305-310.

Lustig, A.J. and Petes, T.D. (1993). Genetic control of simple sequence stability in yeast. In. *Genome Analysis* **7**, *Genome rearrangement and stability*. Davies, K.E. and Warren, S.T. (eds). Coldspring Harbour Laboratory Press, New York: 79-106.

Luty, J.A., Guo, Z., Willard, H.F., Ledbetter, D.H. and Litt, M. (1990). Five polymorphic VNTRs on the human X chromosome. *Am. J. Hum. Genet.* **46**: 776-783.

MacDonald, M.E., Novelletto, A., Lin, C., Tagle, D., Barnes, G., Bates, G., Taylor, S., Allitto, B., Altherr, M., Myers, R., Lehrach, H., Collins, F.S., Wasmuth, J.J., Fronali, M. and Gusella, J.F. (1992). The Huntington's disease candidate region exhibitss many different haplotypes. *Nature Genet.* **1**: 99-103.

MacIlwain, C. and Dickson, D. (1994). Academy under fire over plans for new study of DNA statistics.....as confusion leads to retrial in UK. *Nature* **367**: 101-102.

MacPherson, J.N., Bullman, H., Youings, S.A. and Jacobs, P.A. (1994). Insert size and flanking haplotype in fragile X and normal populations: possible multiple origins for the fragile X mutation. *Hum. Mol. Genet.* **3**: 399-405.

Maeda, N. (1985). Nucleotide sequence of the haptoglobin and haptoglobin-related gene pair. *J. Biol. Chem.* **260**: 6690-6709

Maeda, N. and Smithies, O. (1986). The evolution of multigene families: Human haptoglobin genes. *Ann. Rev. Genet.* **20**: 81-108.

Mahtani, M.M. and Willard, H.F. (1990). Pulsed-field gel analysis of alpha-satellite DNA at the human X-chromosome centromere: High frequency polymorphisms and array size estimate. *Genomics* **7**: 607-613.

Mahtani, M.M. and Willard, H.F. (1993). A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: inplications for mechanisms of mutation at short tandem repeat loci. *Hum. Mol. Genet.* **2**: 431-437.

Malcolm, S., Clayton-Smith, J., Nichols, M., Robb, S., Webb, T., Armour, J.A.L., Jeffreys, A.J. and Pembrey, M.E. (1991). Uniparental paternal isodisomy in Angelman's syndrome. *Lancet* **337**: 694-697.

Mandel, J-L. (1994). Trinucleotide diseases on the rise. *Nature Genet.* **7**: 453-455.

Mant, R., Parfitt, E., Hardy, J. and Owen, M. (1991). Mononucleotide repeat polymorphism in the APP gene. *Nucl. Acids. Res.* **19**: 4572.

Massey, B. and Nicholas, A. (1993). The control in *cis* of the position and the amount of the ARG4 meiotic double-strand break of *Saccharomyces cerevisiae*. *EMBO. J.* **12**: 1459-1466.

Mathew, C.G.P., Smith, B.A., Thorpe, K., Wong, Z., Royle, N.J., Jeffreys, A.J. and Ponder, B.A.J. (1987). Deletion of genes on chromosome 1 in endocrine neoplasia. *Nature* **328**: 524-526.

Matise, T.C., Perlin, M. and Chakravarti, A. (1994). Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. *Nature Genet.* **6**: 384-390.

Meisfeld, R., Krystal, M. and Arnheim, N. (1981). A member of a new repeated sequence family which is conserved throughout eukaryotic evolution is found between the human δ and β-globin genes. *Nucl Acids Res.* **9**: 5931-5947.

Melis R., Bradley, P., Elsner, T., Robertson, M., Lawrence, E., Gerken, S., Albertsen, H. and White, R. (1993). Polymorphic SSR (Simple sequence repeat) markers for chromosome 20. *Genomics* **16**: 56-62.

Mermer, B., Colb, M. and Krontiris, T.G. (1987). A family of short, interspersed repeats is associated with tandemly repetitive DNA in the human genome. *Proc. Nat. Acad. Sci. USA.* **84**: 3320-3324.

Miklos, G.L.G. and John, B. (1979). Heterochromatin and satellite DNA in man: properties and prospects. *Am. J. Hum. Genet.* **31**: 264-280.

Mirsky, A.E. and Ris, H. (1951). The deoxyribonucleic acid content of animal cells and its evolutionary significance. *J. Gen. Physiol.* **34**: 451-462.

Monaco, A., Neve, R., Colletti-Feener, C., Bertelson, C., Kurnit, D. and Kunkel, L. (1986). Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature* **323**: 646-650.

Monckton, D.G. and Jeffreys, A.J. (1991). Minisatellite "isoallele" discrimination in pseudohomozygotes by single molecule PCR and variant repeat mapping. *Genomics* **11**: 465-467.

Monckton. (1993). DNA sequence variation within and around human minisatellites. *PhD thesis, University of Leicester*: a, Chapter 6 Page 3; b, Chapter 6, Page 6; c, Chapter 6 Page 4a; d, Chapter 5 Page 3a; e, Chapter 7 Page 10; f, Chapter 7 Page 11.

Monckton, D.G. and Jeffreys, A.J. (1993). DNA profiling. *Curr. Opin. Biotech.* **4**: 660-664.

Monckton, D.G. and Jeffreys, A.J. (1994). Minisatellite isoalleles can be distinguished by single-stranded conformational polymorphism analysis in agarose gels. *Nucleic Acids Res.* **22**: 2155-2157.

Monckton, D.G., Tamaki, K., Macleod, A., Neil, D.L. and Jeffreys, A.J. (1993). Allele-specific MVR-PCR analysis at minisatellite D1S8. *Hum. Mol. Genet.* **2**: 513-519.

Monckton, D.G., Neumann, R., Guram, T., Fretwell, N., Tamaki, K., MacLeod, A. and Jeffreys, A.J. (1994). Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. *Nature Genet.* **8**: 162-170.

Morgan, T.H., Sturtevant, A.H. Muller, H.J., and Bridges, C.B. (1915) *The mechanism of mendelian heredity*. Rinehart and Winston (eds). Holt, New York.

Morral, N., Nunes, V., Casals, T. and Estevill, X. (1991). CA/GT microsatellite alleles within the cystic fibrosis transmembrane conductance regulator (CFTR) gene are not generated by unequal crossingover. *Genomics* **10**: 692-968.

Morsy, M.A., Mitani, K., Clemens, P. and Caskey, C.T. (1993). Progress toward human gene-therapy. *J. Am. Med. Assn.* **270**: 2338-2345.

Moyzis, R.K., Buckingham, J.M., Cram, L.S., Dani, M., Deaven, L.L., Jones, M.D., Meyne, J., Ratliff, R.L. and Wu, J.R. (1988). A highly conserved repetitive DNA sequence, $(TTAGGG)_n$ at the telomeres of human chromosomes. *Proc. Natl. Acad. Sci. USA.* **85**: 6622-6626.

Murphy, G.L., Connell, T.D., Barritt, D.S., Koomey, M. and Canon, J.G. (1989). Phase variation of gonococcal protein II. Regulation of gene expression by slipped-strand mispairing of a repetitive DNA sequence. *Cell* **56**: 539-547.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. and White, R. (1987a). Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**: 1616-1622.

Nakamura, Y., Julier, C., Wolff, R., Holm, T., O'Connell,P., Leppert, M. and White, R. (1987b). Characterization of a human 'midisatellite' sequence. *Nucleic Acids Res.* **15**: 2537-2547.

Nakamura, Y., Carlson, M., Krapcho, K., Kanamori, M. and White, R. (1988a). New approach for the isolation of VNTR markers. *Am. J. Hum. Genet.* **43**: 854-859.

Nakamura, Y., Lathrop, M., O'Connell, P., Leppert, M., Barker, D., Wright, E., Skolnick, M., Kondoleon, S., Litt, M., Lalouel, J.-M. and White, R. (1988b). A mapped set of markers for human chromosome 17. *Genomics* **2**: 302-309.

Nakamura, Y., Lathrop, M., O'Connell, P., Leppert, M., Kamboh, M.I., Lalouel, J.-M. and White, R. (1989). Frequent recombination is observed in the distal end of the long arm of chromosome 14. *Genomics* **4**: 76-81.

Nei, M. and Roychoudhury, A.K. (1974). Genetic variation within and between the three major races of man, Caucasoids, Negroids and Mongoloids. *Am. J. Hum. Genet.* **26**: 421-443.

Neil, D.L. and Jeffreys, A.J. (1993). Digital DNA typing at a second hypervariable locus by minisatellite variant repeat mapping. *Hum. Mol. Genet.* **2**: 1129-1135.

Nelson, D.L. (1993). Six human genetic disorders involving mutant trinucleotide repeats. In. *Genome Analysis* **7**, *Genome rearrangement and stability.* Davies, K.E. and Warren, S.T. (eds). Coldspring Harbour Laboratory Press, New York: 1-25.

Newton, C.R., Graham, A., Hepstinall, L.E., Powell, S.J., Summers, C., Kalsheker, N., Smith, J.C. and Markham, A.F. (1989). Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Res.* **17**: 2503-2516.

Nickerson, D.A., Whitehurst, C., Boysen, C., Charmley, P., Kaiser, R. and Hood, L. (1992). Identification of clusters of biallelic polymorphic sequence-tagged sites (pSTSs) that generate highly informative and automatable markers for genetic linkage mapping. *Genomics* **12**: 377-387.

NIH/CEPH Collaborative Mapping Group. (1992). A comprehensive genetic linkage map of the human genome. *Science* **258**: 67-86.

Oakey, R. and Tyler-Smith, C. (1990). Y Chromosome DNA Haplotyping suggests that most European and Asian men are descended from one of two males. *Genomics* **7**: 325-330.

Olaisen, B., Bekkemoen, M., Hoff-Olsen, P. and Gill, P. (1993). Human VNTR mutation and sex. In. *DNA fingerprinting: State of the science.* Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. (eds). Birkhaüser Verlag, Basel: 63-69.

Oliver, S. (1994). Back to bases in biology. *Nature* **368**: 14-15.

O'Hoy, K.L., Tsilfidis, C., Mahadevan, M.S., Neville, C.E., Barcelo, J., Hunter, A.G.W. and Korneluk, R.G. (1993). Reduction in size of the myotonic dystrophy trinucleotide repeat mutation during transmission. *Science* **259**: 809-812.

Orita, M., Suzuki, Y., Sekiya, T. and Hayashi, K. (1989). Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics* **5**: 874-879.

Orr, H.T., Chung, M.-Y., Banfi, S., Kwiatkowski, T.J. Jr., Servadio, A., Beaudet, A.L., McCall, A.E., Duvick, L.A., Ranum, L.P.W. and Zoghbi, H.Y. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nature. Genet.* **4**: 221-226.

Oudet, C., von Koskull, H., Nordström, A.M., Peippo, M. and Mandel, J.L. (1993). Striking founder effect for the Fragile X syndrome in Finland. *Eur. J. Hum. Genet.* **1**: 181-189.

Owerbach, D. and Aagaard, L. (1984). Analysis of a 1963bp polymorphic region flanking the human insulin gene. *Gene* **32**: 475-479.

Paabo, S., Gifford, J.A. and Wilson, A.C. (1988). Mitochondrial sequences from 7000 year old brain. *Nucleic Acids Res.* **16**: 9755-9785.

Packer, C., Gilbert, D.A., Pusey, A.E. and O'Brien, S.J. (1991). A molecular genetic analysis of kinship and cooperation in African lions. *Nature* **351**: 562-565.

Page, D.C., Bieker, K., Brown, L.G., Hinton, S., Leppert, M., Lalouel, J-M., Lathrop, M., Nystrom-Lahti, M., De La Chappelle, A. and White, R. (1987). Linkage, physical mapping and DNA sequence analysis of pseudoautosomal loci on the human X and Y chromosomes. *Genomics* **1**: 243-256.

Palmer, M.S. and Collinge, J. (1993). Mutations and polymorphisms in the prion protein gene. *Hum. Mut.* **2**: 168-173.

Pardue, M.L. and Gall, J.G. (1970). Chromosomal location of mouse satellite DNA. *Science* **168**: 1356-1358.

Pauling, L., Itano, H.A., Singer, S.J. and Wells, I.C. (1949). Sickle cell anaemia: a molecular disease. *Science* **110**: 64-66.

Paulo di Nocera, P. and Sakaki, Y. (1990). LINEs: a super-family of retrotransposable ubiquitous DNA elements. *Trends Genet.* **6**: 29-30.

Paulson, K.E., Deka, N., Schmid, C.W., Misra, R., Schindler, C.W., Rush, M.G., Kadyk, L. and Leinwand, L. (1985). A transposon-like element in human DNA. *Nature* **316**: 559-361.

Peake, I.R.R.R., Bowen, D., Bignell, P., Lidell, M.B., Sadler, J.E., Standen, G. and Bloom, A.L. (1990). Family studies and prenatal diagnosis in severe von Willebrand disease by polymerase chain reaction amplification of a variable number tandem repeat region of the von Willebrand factor gene. *Blood* **76**: 555-561.

Pena, S.D.J., Souza, K.T., Andrade, M. and Chakraborty, R. (1994). Allelic associations of two polymorphic microsatellites in intron 40 of the human von Willebrand factor gene. *Proc. Natl. Acad. Sci. USA.* **91**: 723-727.

Polymeropoulos, M.H., Rath, D.S., Xiao, H. and Merril, C.R. (1992). Tetranucleotide repeat polymorphism at the human beta-actin related pseudogene H-beta-Ac-Psi-2 (ACTBP2). *Nucleic Acids Res.* **20**: 1432.

Prosser, J., Frommer, M., Paul, C. and Vincent, V.C. (1986). Sequence relationships of three human satellite DNAs. *J. Mol. Biol.* **187**: 145-155.

Prusiner, S.B. (1991). Molecular biology of prion diseases. *Science* **252**: 1515-1522.

Qu, L-H., Nicoloso, M. and Bachellerie, J-P. (1991). A sequence dimorphism in a conserved domain of human 28S rRNA. Uneven distribution of variant genes among individuals. Differential expression in HeLa cells. *Nucleic Acids Res.* **19**: 1015-1019.

Race, R.R. and Sanger, R. (1975). *Blood groups in man.* 6th. edition. Blackwell, Oxford.

Reed, P.W., Davies, J.L., Copeman, J.B., Bennett, S.T., Palmer, S.M., Pritchard, L.E., Gough, S.C.L., Kawaguchi, Y., Cordell, H.J., Balfour, K.M., Jenkins, S.C., Powell, E.E., Vignal, A. and Todd, J.A. (1994). Chromosome-specific microsatellite sets for fluorescence-based, semi-automated genome mapping. *Nature Genet.* **7**: 390-395.

Reeders, S.T., Breuning, M.H., Davies, K.E., Nicholls, R.D., Jarman, A.P., Higgs, D.R., Pearson, P.L. and Weatherall, D.J. (1985) A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. *Nature* **317**: 542-544.

Richards, R.I. and Sutherland, G.R. (1992). Dynamic mutations: A new class of mutations causing human disease. *Cell* **70**: 709-712.

Richards, R.I. and Sutherland, G.R. (1994). Simple repeat DNA is not replicated simply. *Nature Genet.* **6**: 114-116.

Richards, R.I., Holman, K., Friend, K., Kremer, E., Hillen, D., Staples, A., Brown, W.T., Goonewardena, P., Tarleton, J., Schwartz, C. and Sutherland, G.R. (1992). Evidence of founder chromosomes in fragile X syndrome. *Nature Genetics* **1**: 257-260.

Riggins, G.J., Lokey, L.K., Chastain, J.L., Leiner, H.A., Sherman, S.L., Wilkinson, K.D. and Warren, S.T. (1992). Human genes containing polymorphic trinucleotide repeats. *Nature Genetics* **2**: 186-191.

Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., Anand, S., Smith, J.C. and Markham, A.F. (1990). A novel method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Res.* **18**: 2887-2890.

Rogstad, S.H., Patton, J.C.II. and Schall, B.A. (1988). A human minisatellite probe reveals RFLPs among individuals of two angiosperms. *Nucleic Acids Res.* **18**: 1081.

Rommens, J.M., Ianuzzi, M.C., Kerem, B-S., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, L.J., Kennedy, D., Hidaka, N., Zsiga, M., Buchwald, M., Rordan, J.R., Tsui, L-C. and Collins, F.S. (1989). Identification of the cystic-fibrosis gene: chromosome walking and jumping. *Science* **245**: 1059-1065.

Rouyer, F., Simmler, M.-C., Johnsson, C., Vergnaud, G., Cooke, H.J. and Weissenbach, J. (1986). A gradient of sex linkage in the pseudoautosomal region of the human sex chromosomes. *Nature* **319**: 291-295.

Royle, N.J., Clarkson, R.E., Wong, Z. and A.J. Jeffreys. (1988). Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. *Genomics* **3**: 352-360.

Royle, N.J., Armour, J.A.L., Webb, M., Thomas, A. and Jeffreys, A.J. (1992). A hypervariable locus D16S309 located at the distal end of 16p. *Nucleic Acids Res.* **20**: 1164.

Saiki, R.K., Bugawan, T.L., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1986). Analysis of enzymatically amplified β-globin and HLA-DQα DNA with allele-specific oligonucleotide probes. *Nature* **324**: 163-166.

Saiki, R.K., Bugawan, T.L., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988). Primer directed amplification of DNA with a thermostable DNA polymerase. *Science* **239**: 487-491.

Saito, I. and Stark, G.R. (1986). Charomids: cosmid vectors for efficient cloning and mapping of large or small restriction fragments. *Proc. Natl. Acad. Sci. USA.* **83**: 8664-8668.

Sakar, G., Kapelner, S. and Sommer, S. (1990). Formamide can dramatically improve the specificity of PCR. *Nucleic Acids Res.* **18**: 7465.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). *Molecular cloning, a laboratory manual.* 2nd Edition. Coldspring Harbour Laboratory Press. New York.

Schlotterer, C. and Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **20**: 211-215.

Schmid, C.W. and Deininger, P.L. (1975). Sequence organisation of the human genome. *Cell* **6**: 345-358.

Schmid, C.W. and Jelinek, W.R. (1982) The Alu family of dispersed repetitive sequences. *Science* **216**: 1065-1070.

Schmitt, K. and Goodfellow, P.N. (1994). Predicting the future. *Nature Genetics* **7**: 219.

Schultes, N.P. and Szostak, J.W. (1991). A poly (dA.dT) tract is a component of the recombination initiation site at the *ARG4* locus in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **11**: 322-328.

Schumm,J.,Knowlton,R.,Braman,J.,Barker D., Vovis G., Akots, G., Brown, V., Gravius, T., Helms, C., Hsiao, K., Rediker, K., Thurston, J., Botstein, D. and Donis-Keller, H. (1985). Detection of more than 500 single-copy RFLPs by random screening. *Cytogenet. Cell Genet.* **40**: 739.

Shelbourne, P., Winqvist, R., Kunert, E., Davies, J., Leisti, J., Thiele, H., Bachmann, H., Buxton, J., Williamson, B. and Johnson, K. (1992). Unstable DNA may be responsible for the incomplete penetrance of the myotonic dystrophy phenotype. *Hum. Mol. Genet.* **1**: 467-473.

Sherrington, R., Melmer, G., Dixon, M., Curtis, D., Mankoo, B., Kalsi, G. and Gurling, H. (1991). Linkage disequilibrium between two highly polymorphic microsatellites. *Am. J. Hum. Genet.* **49**: 966-971.

Signer, E.N. and Jeffreys, A.J. (1993). Application of human minisatellite probes to the development of informative DNA fingerprints and the isolation of locus-specific markers in animals. In. *DNA fingerprinting: State of the science*. Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. (eds). Birkhaüser Verlag Basel: 421-428.

Signer, E.N., Gu, F., Gustavsson, I., Andersonn, L. and Jeffreys, A.J. (1994). A pseudoautosomal minisatellite in the pig. *Mammalian. Genome* **5**: 48-51.

Simon, M., Phillips, M. and Green, H. (1991). Polymorphism due to variable number of repeats in the human involucrin gene. *Genomics* **9**: 576-580.

Singer, M.F. (1982). SINES and LINES: highly repeated short and long interspersed sequences in mammalian genomes. *Cell* **28**: 433-434.

Smit, A. (1993). Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res.* **21**: 1863-1872

Smith, G.P. (1976) Evolution of repeated DNA sequences by unequal crossover. *Science* **191**: 528-535.

Smith, J.C., Anwar, R., Riley, J., Jenner, D., Markham, A.F. and Jeffreys, A.J. (1990). Highly polymorphic minisatellite sequences: allele frequencies and mutation rates for five locus specific probes in a Caucasian population. *J. For. Sci. Soc.* **30**: 19-32.

Solari, A.J. (1980). Synaptonemal complexes and associated structures in microspread human spermatocytes. *Chromosoma* **81**: 307-314.

Solomon, E., Voss, R., Hall, V., Bodmer, W.F., Jass, J.R., Jeffreys, A.J., Lucibello, F.C., Patel, I. and Rider, S.H. (1987). Chromosome 5 allele loss in human colorectal carcinomas. *Nature* **328**: 616-619.

Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E. and Moyzis, R.K. (1991). Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes. *Genomics* **10**: 807-815.

Stephan, W. (1989). Tandem-repetitive noncoding DNA: forms and forces. *Mol. Biol. Evol.* **6**: 198-212.

Strand, M., Prolla, T.A., Liskay, R.M. and Petes, D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365**: 274-276.

Sueoka, N. (1961). Variation and heterogeneity of base composition of deoxyribonucleic acids: a compilation of old and new data. *J. Mol. Biol.* **3**: 31-40.

Sullivan, K.M., Hopgood, R., Lang., B. and Gill, P. (1991). Automated amplification and sequencing of human mitochondrial DNA. *Electrophoresis* **12**: 17-21.

Sullivan, K.M., Hopgood, R. and Gill, P. (1992) Identification of human remains by amplification and automated sequencing of mitochondrial DNA. *Int. J. Leg. Med.* **105**: 83-86.

Sullivan, K.M., Walton, A., Kimpton, C., Tully, G. and Gill, P. (1993). Fluorescence-based DNA segment analysis in forensic science. *Biochem. Soc. Trans.* **21**: 116-120.

Sun,L., Paulson, K.E., Schmid, C.W., Kadyk, L. and Leinwand, L. (1984). Non-Alu family interspersed repeats in human DNA and their transcriptional activity. *Nucleic Acids Res.* **12**: 2669-2690.

Sun, H., Treco, D., Schultes, N.P. and Szostak, J.W. (1989). Double-strand breaks at an initiation site for meiotic gene conversion. *Nature* **338**: 87-90.

Sun, H., Dawson, D. and Szostak, J.W. (1991a). Extensive 3′-overhanging, single-stranded DNA associated with the meiosis-specific double-strand breaks at the *ARG4* recombination initiation site. *Cell* **64**: 1155-1161.

Sun, H., Dawson, D. and Szostak, J.W. (1991b). Genetic and physical analyses of sister-chromatid exchange in yeast meiosis. *Mol. Cell. Biol.* **11**: 6328-6336.

Sutton, W.S. (1903). The chromosome in heredity. *Biol. Bull.* **4**: 231-251.

Swallow, D.M., Gendler, S., Griffiths, B., Corney, G., Taylor-Papadimitriou, J. and Bramwell, M.E. (1987). The human tumour-associated epithelial muncins are coded by an expressed hypervariable gene locus PUM. *Nature* **328**: 82-84.

Szostak, J.W., Orr-Weaver, T.L. and Rothstein, R.J. (1983). The double-strand-break repair model for recombination. *Cell* **33**: 25-35.

Tamaki, K., Yamamoto, T., Uchihi, R., Katsumata, Y., Kondo, K., Mizuno, S., Kimura, A. and Sasazuki, T. (1991). Frequency of HLA-DQA1 alleles in the Japanese population. *Hum. Hered.* **41**: 209-214.

Tamaki, K., Monckton, D.G., MacLeod, A., Neil, D.L., Allen, M. and Jeffreys, A.J. (1992). Minisatellite variant repeat (MVR) mapping: analysis of "null" repeat units at D1S8. *Hum. Mol. Genet.* **1**: 401-406.

Tamaki, K., Monckton, D.G., MacLeod, A., Allen, M. and Jeffreys, A.J. (1993). Four-state MVR-PCR: increased discrimination of digital DNA typing by simultaneous analysis of two polymorphic sites within minisatellite variant repeats at D1S8. *Hum. Mol. Genet.* **2**: 1629-1632.

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic markers. *Nucleic Acids Res.* **17**: 6463.

Tautz, D and Renz, M. (1984). Simple sequences are ubiquitous repetitive elements of eukaryotic genomes. *Nucleic Acids. Res.* **12**: 4127-4138.

Tautz, D., Trick, M. and Dover, G.A. (1986) Cryptic simplicity in DNA is a major source of genetic variation. *Nature* **322**: 652-656.

Timofeeva, M.Y., Kost, M.V., Tikhomirova, T.P., Sarafanov, A.G., Elbert, B.L., Altimov, A.A., Kvitsiniya, S.L., Beritashili, D.R. and Zelenin, A.V. (1993). Organization of the 5S ribosomal-RNA cluster in the human genome. *Mol. Biol.* **27**: 531-535.

Thacker, J., Webb, M.B.T. and Debenham, P.G. (1988). Fingerprinting cell lines: Uses of human hypervariable DNA probes to characterise mammalian cell cultures. *Somat. Cell Mol. Genet.* **14**: 519-525.

Thein, S.L., Jeffreys, A.J. and Blacklock. (1986). Identification of a post-transplant cell population by DNA fingerprint analysis. *Lancet* **2**: 37.

Thein, S.L., Jeffreys, A.J., Gooi, H.C., Cotter, F., Flint, J., O'Connor, N. and Wainscoat, J.S. (1987). Detection of somatic changes in human cancer DNA by DNA fingerprint analysis. *Br. J. Cancer* **55**: 353-356.

Thompson, W.C. and Ford, S. (1991). The meaning of a match: sources of ambiguity in the interpretation of DNA prints. In. *Forensic DNA technology*. Farley, M.A. and Harrington, J.J. (eds). Lewis Publishers, Inc, Michigan: 93-152.

Tyler-Smith, C. and Brown, W.R.A. (1987). Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J.Mol. Biol.* **195**: 457-470.

Vassart, G., Georges, M., Monsieur, R., Brocas, H., Lequarre, A.S. and Christophe, D. (1987). A sequence in M13 phage detects hypervariable minisatellites in human and animal DNA. *Science* **235**: 683-684.

Vergnaud, G. (1989). Polymers of random short oligonucleotides detect polymorphic loci in the human genome. *Nucleic Acids Res.* **17**: 7623-7630.

Vergnaud, G., Mariat, D., Apiou, F., Aurias, A., Lathrop, M. and Lauthier, V. (1991). The use of synthetic tandem repeats to isolate new VNTR loci: cloning of a human hypermutable sequence. *Genomics* **11**: 135-144.

Vogel, F. and Rathenberg, R. (1975). Spontaneouss mutation in man. *Adv. Hum. Genet.* **5**: 223-318.

Vogelstien, B., Fearon, E.R., Kern, S.E., Hamilton, S.R., Preisinger, A.C., Nakamura, Y. and White, R. (1989). Allelotype of colorectal carcinomas. *Science* **244**: 207-211.

Vuorio, E. and de Crombrugghe, B. (1990). The family of collagen genes. *Ann. Rev. Biochem.* **59**: 837-872.

Wallis, G.A., Starman, B.J. and Byers, P.H. (1989). Clinical heterogeneity of OI explained by molecular heterogeneity and somatic mosaicism, *Am. J. Hum. Genet.* **45**: A895.

Wahls, W.P., Swenson, G. and Moore, P.D. (1991). Two hypervariable minisatellite DNA binding proteins. *Nucleic Acids Res.* **19**: 3269-3274.

Wang, Z.Y., Weber, J.L., Ludwigsen, S. and Buetow, K. (1993). Human chromosome-8 linkage map based on short tandem repeat polymorphisms. *Cytogenet. Cell Genet.* **64**: 145.

Warburton, P.E., Grieg, G.M., Haaf, T. and Willard, H.F. (1991). PCR amplification of chromosome specific alpha-satellite DNA: definition of centromeric STS markers and polymorphic analysis. *Genomics* **11**: 324-333.

Watson, J.D. (1990) The human genome project-past, present, and future. *Science* **248**: 44-49.

Waye, J.S. and Willard, H.F. (1986) Nucleotide sequence heterogeneity of alpha-satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res.* **15**: 7549-7568.

Waye, J.S., Richard, M., Carmody, G. and Newall, P.J. (1994). Allele frequency data for VNTR locus D17S79: identification of an internal *Hae*III polymorphism in the Black population. *Hum. Mut.* **3**: 248-253.

Weatherall, D.J. and Clegg, J.B. (1976). Molecular genetics of human haemoglobin. *Ann. Rev. Genet.* **10**, 157-178.

Weatherall, D.J. and Clegg, J.B. (1982). Thalassemia revisited. *Cell* **29**: 7-9.

Weber, J.L. (1990). Informativeness of human $(dA\text{-}dC)_n.(dG\text{-}dT)_n$ polymorphisms. *Genomics* **7**: 524-530.

Weber, J.L. and May, P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**: 388-396.

Weber, J.L. and Wong, C. (1993). Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**: 1123-1128.

Weber, J.L., Wang, Z.Y., Hansen, K., Stephenson, M., Kappel, C., Salzman, S., Wilkie, P.J., Keats, B., Dracopoli, N.C., Brandriff, B.F. and Olsen, A.S. (1993). Evidence for human meiotic recombination interferance obtained through construction of a short tandem repeat-polymorphism linkage map of chromosome-19. *Am. J. Hum. Genet.* **53**: 1079-1095.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morisette, J., Millasseau, P., Vaysseix G., and Lathrop, M. (1992). A second-generation linkage map of the human genome. *Nature* **359**: 794-801.

Wells, R.D. and Sinden, R.R. (1993). Defined ordered sequence DNA, DNA structure and DNA-directed mutation. In. *Genome Analysis* **7**, *Genome rearrangement and stability*. Davies, K.E. and Warren, S.T. (eds). Coldspring Harbour Laboratory Press, New York: 107-138.

Wetton, J.H., Carter, R.E., Parkin, D.T. and Walters, D. (1987). Demographic study of a wild house sparrow population by DNA fingerprinting. *Nature* **327**: 147-149.

Wilkie, A.O.M. and Higgs, D.R. (1992). An unusually large $(CA)_n$ repeat in the region of divergence between subtelomeric alleles of human chromosome 16p. *Genomics* **13**: 81-88.

Wilkie, P.J., Krizman, D.B. and Weber, J.L. (1992). Linkage map of chromosome 9 microsatellite polymorphisms. *Genomics* **12**: 607-609.

Willard, H.F. (1990) Centromeres of mammalian chromosomes. *Trends Genet.* **6**: 410-416.

Willard, H.F. 1991. Evolution of alpha-satellite. *Current Opin. Genet. Dev.* **1**: 509-514.

Willard, H.F. and Waye, J.S. (1987). Hierarchical order in chromosome-specific human alpha-satellite DNA. *Trends in Genet.* **3**: 192-195.

Willard, H.F., Waye, J.S., Skolnick, M.H., Schwartz, C.E., Powers, V.E. and England, S.B. (1986). Detection of restriction fragment length polymorphisms at the centromeres of human chromosomes by using chromosome-specific alpha-satellite DNA probes: implications for development of centromere-based genetic-linkage maps. *Proc. Natl. Acad. Sci. USA.* **83**: 5611-5615.

Wohlrab, F., McLean, M. and Wells, R.D. (1987). The segment inversion site of Herpes Simplex Virus Type 1 adopts a novel DNA structure. *J. Biol. Chem.* **262**: 6407-6416.

Wohlrab, F., Chatturjee, S. and Wells, R.D. (1991). The Herpes Simplex Virus Type 1 segment inversion site is specifically cleaved by a virus-induced nuclear endonuclease. *Proc. Natl. Acad. Sci. USA.* **88**: 6432-6436.

Wolff, R.K., Nakamura, Y. and White, R. (1988). Molecular characterization of a spontaneously generated new allele at a VNTR locus: no exchange of flanking DNA sequence. *Genomics* **3**: 347-351.

Wolff, R.K., Plaetke, R., Jeffreys, A.J. and White, R. (1989). Unequal crossingover between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci. *Genomics* **5**: 382-394.

Wolff, R., Nakamura, Y., Odelberg, S., Shiang, R. and White, R. (1991). Generation of variability at VNTR loci in human DNA. In. *DNA fingerprinting: approaches and applications*. Burke, T., Dolf. G., Jeffreys, A.J. and Wolff, R. (eds). Birkhaüser Verlag, Basel: 20-38.

Wong, Z., Wilson, V., Jeffreys, A.J. and Thein, S.L. (1986). Cloning a selected fragment from a human DNA 'fingerprint': isolation of an extremely polymorphic minisatellite. *Nucleic Acids Res.* **14**: 4605-4616.

Wong, Z., Wilson, V., Patel, I., Povey., S. and Jeffreys, A.J. (1987). Characterization of a panel of highly variable minisatellites cloned from human DNA. *Ann. Hum. Genet.* **51**: 269-288.

Wong, Z., Royle, N.J. and Jeffreys, A.J. (1990). A novel human DNA polymorphism resulting from transfer of DNA from chromosome 6 to chromosome 16. *Genomics* **7**: 222-234.

Wyman, A.R. and White, R. (1980) A highly polymorphic locus in human DNA. *Proc. Nat. Acad. Sci. USA.* **77**: 6754-6758.

Wyman, A.R., Wolfe, L.B. and Botstein, D. (1985). Propagation of some human DNA sequences in bacteriophage λ vectors requires mutant Escherichia coli hosts. *Proc. Nat. Acad. Sci. USA.* **82**: 2880-2884.

Wyman, A.R., Mulholland, J. and Botstein, D. (1986). Oligonucleotide repeats involved in the highly polymorphic locus D14S1. *Am. J. Hum. Genet.* **39**: A226

Yamazaki, H., Nomoto, S., Mishima, Y. and Kominami, R. (1992). A 35-kDa protein binding to a cytosine-rich strand of minisatellite DNA. *J. Biol. Chem.* **267**: 12311-12316.

Yanagawa, T., Mangklabruks, A., Chang, Y.B., Okamoto, Y., Fisfalen, M.E., Curran, P.G. and Degroot, L.J. (1993). Human histocompatability leucocyte antigen-DQα-*-0501 allele associated with genetic susceptibility to Graves-disease in a Caucasian population. *J. Clin. Endocrinol. Metabol.* **76**: 1569-1574.

Yandell, D.W. and Dryja, T.P. (1989). Detection of DNA sequence polymorphisms by enzymatic amplification and direct genomic sequencing. *Am. J. Hum. Genet.* **45**: 547-555.

Yu, S., Mulley, J., Loesch, D., Turner, G., Donnelly, A., Gedeon, A., Hillen, D., Kremer, E., Lynch, M., Pritchard, M., Sutherland, G.R. and Richards, R.I. (1992). Fragile-X syndrome: unique genetics of the heritable unstable element. *Am. J. Hum. Genet.* **50**: 968-980.

Zhong, N., Dobkin, C. and Brown, W.T. (1993). A complex mutable polymorphism located within the fragile X gene. *Nature Genet.* **5**: 248-253.