POLYMORPHIC TANDEMLY REPEATED SEQUENCES IN HUMAN DNA

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

by

Ian Christopher Gray BSc (Leicester)
Department of Genetics
University of Leicester

November 1991

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

This work is dedicated to the memory of Alan Donoghue

Contents

Abstract

Tandemly repeated tracts of DNA are a ubiquitous feature of eukaryote genomes. One class of tandem repeats, 'minisatellites', have been shown to be highly variable both in overall length and in the internal arrangement of variant versions of the repeating unit along the array. Consequently, both length and internal variation at these loci can be exploited to generate individual-specific profiles of use in forensic science and the establishment of family relationships. Recently it has been demonstrated that short dinucleotide repeats, or 'microsatellites', and other simple tandem repeat arrays can also show length variation. This work describes the isolation and characterization of simple tandem repeat arrays, and their application in a forensic science case. The evolutionary persistence of variability at tandem repeat loci is also explored. Simple tandem repeats isolated were frequently associated with other tandem repeats and interspersed repetitive elements, a phenomenon previously described for minisatellites and perhaps indicative that certain genomic regions show relaxed fidelity in the maintenance of large-scale DNA structure, allowing tandem array expansion and retroposon insertion. Although variability is low relative to minisatellites, microsatellites, owing to limited length, can readily be amplified from highly degraded DNA using the polymerase chain reaction. Consequently it was possible to identify positively the skeletal remains of a murder victim by comparing microsatellite profiles of the skeleton with those of the presumptive parents. Comparative studies of microsatellite and minisatellite loci between man and other primates indicate that in evolutionary terms, the variable state is reasonably persistent at microsatellite loci, whereas highly variable minisatellites show extreme evolutionary transience. Such transience was also demonstrated for two large 'midisatellite' loci, suggesting that highly variable tandem repeat loci are extremely unstable and transient, whereas lower variability leads to evolutionary persistence of the variable state.

Aknowledgements

Initially I would like to thank members of G19 past and present for advice, assistance and useful discussions: thanks to Nicola Royle, John Armour, Andy Collick, Robert Kelly, Ila Patel, Vicky Wilson, Zilla Wong, Rita Neumann, Annette Macleod, Mark Gibbs, Esther Signer, Darren Monckton, Moira Crosier, Maureen Hill, David Neil and Keiji Tomaki.

I also owe thanks to my parents for their continual encouragement through six years of University education.

Thanks are also due to the Wellcome Trust for a grant which actually provided enough money to live on, with particular thanks to Mrs Sian Doughty for helpful and efficient administration.

Finally my greatest debt of thanks must be to Alec, for allowing me to work in his laboratory during an exciting period of discovery, and for providing expert supervision during that time.

Some of this work has been published:

Jeffreys, A.J., Royle, N.J., Patel, I., Armour, J.A.L., MacLeod, A., Collick, A., Gray, I.C., Neumann, R., Gibbs, M., Crosier, M., Hill, M., Signer, E. & Monckton, D. (1991). Principles and recent advances in DNA fingerprinting. In *DNA fingerprinting: approaches and applications*. pp 1-19. Ed: T. Burke, G. Dolf, A.J. Jeffreys & R. Wolff; Birkhauser Verlag, Bern.

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

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

Abbreviations

ATP	adenosine 5'-triphosphate
BCIG	5-bromo-4-chloro-3-indoyl-β-D-galactopyranoside
bp	base pairs
BSA	bovine serum albumin
CEPH	Centre d'Etude du Polymorphisme Humain
Ci	Curie
cM	centiMorgan
dATP	2'-deoxyadenosine 5'-triphospate
dCTP	2'-deoxycytosine 5'-triphospate
dGTP	2'-deoxyguanosine 5'-triphospate
dTTP	2'-deoxythymidine 5'-triphospate
dNTP	2'-deoxyribonucleoside 5'-triphosphate
ddATP	2'3'-dideoxyadenosine 5'-triphospate
ddCTP	2'3'-dideoxycytosine 5'-triphospate
ddGTP	2'3'-dideoxyguanosine 5'-triphospate
ddTTP	2'3'-dideoxythymidine 5'-triphospate
ddNTP	2'3'-dideoxyribonucleoside 5'-triphosphate
DMSO	dimethylsulphoxide
DNA	deoxyribonucleic acid
dpm	disintegrations per minute
DTT	dithiothreitol
EDTA	ethylenediaminetetra-acetic acid
g	grams
mg	milligrams
µg	micrograms
ng	nanograms
HEPES	N-(2-hydroxyethyl)piparizine-N'-2(ethanesulphonic acid)
IPTG	Isopropyl-β-D-galactopyranoside
kb	kilobase pairs
l	litres
ml	millilitres
µl	microlitres
LINE	long interspersed repeated element
LOD	log of the odds
LTR	long terminal repeat
LUA	Luria agar

LUB	Luria broth
M	molar
mM	millimolar
μM	micromolar
MOPS	3(N-morpholino)propanesulphonic acid
MVR	minisatellite variant repeat
MYA	million years ago
nt	nucleotides
OD	optical density
PCR	polymerase chain reaction
PEG	polyethylene glycol
PIC	polymorphism information content
PVP	polyvinyl pyrrolidone
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
rpm	revolutions per minute
SDS	sodium dodecyl sulphate
SINE	short interspersed repeated element
SSC	saline sodium citrate
TAE	Tris-acetate EDTA
TBE	Tris-borate EDTA
TEMED	N,N,N'N'-tetramethyl-ethylenediamine
Tris	Tris-(hydroxymethyl)-methylamine[2-amino-(2-hydroymethyl)-propan-1,3-diol]
θ	recombination fraction
UV	ultra violet
VNTR	variable number of tandem repeats
w/v	weight to volume
z	LOD score

CHAPTER 1

INTRODUCTION

1.1 The C-value paradox

The uncomfortable realization that amphibians and fish can have genomes twenty times as large as that of man was first made in 1951 by Mirsky & Ris. Furthermore, closely related species were found to have very different genome sizes or 'C-values'. Subsequent estimates of gene number, average gene size and genome size drew the conclusion that only a very small proportion of the eukaryote genome encodes protein products and that there is an enormous excess of apparently superfluous DNA in most eukaryote genomes (reviewed by Lewin, 1975). DNA reassociation kinetics demonstrated that much of this non-coding sequence is repetitive (Britten & Kohne, 1968), with more recent estimates suggesting that 20-30% of human DNA is of a repeated nature (see Schmid & Jelinek, 1982). The remaining excess DNA is apparently unique sequence, forming tracts between genes and non-coding intervening sequences within genes (Jeffreys & Flavell, 1979).

Although it was something of a relief to discover that despite being twenty times as large as that of man, the salamander genome was of similar sequence complexity, the role of the repetitive DNA element remained (and to a large extent still remains) enigmatic. In mammals, the bulk of this repetitive DNA can be divided into two broad categories: 'satellite' DNA comprising large tracts of simple tandemly repeated sequence, and dispersed repetitive DNA, consisting of elements which are not tandemly repeated but are interspersed with regions of unique sequence (see Schmid & Jelinek, 1982).

1.2 Dispersed repetitive DNA

The interspersed repetitive DNA component can be divided into two broad categories: short interspersed repetitive elements (SINES) and long interspersed repetitive elements (LINES; Singer, 1982). By far the most common SINE in human DNA is the Alu element, so called due to the presence of an *Alu*I restriction endonuclease cleavage site (Houck *et al.*, 1979). Alu is a ~300bp element scattered throughout the genome at a copy number of $>5x10^5$, representing ~5% of total human genomic DNA (Singer, 1982). Alu elements are ubiquitous in primates, and consist of two direct 130bp repeats, with a 31bp insertion in the second repeat

(Deininger *et al.*, 1981). A similar element in rodents, termed the B1 element, consists of a 130bp monomeric form of the Alu element (Krayev *et al.*, 1980).

Alu and B1 elements show all the characteristics of processed pseudogenes, inactive copies of functional genes which appear to have arisen following reverse transcription from mRNA and random insertion of the resulting DNA copy into a different region of the genome (see Rogers, 1985). Such elements, in common with Alu and B1 elements, are flanked by short direct repeats and have a polyadenosine tail. Most processed pseudogenes have been identified as single copies (Vanin, 1984); however, pseudogenes originating from the small nuclear RNA genes U1-U6 have copy numbers of 10^2-10^3 in human DNA (Denison & Weiner, 1982). B1 elements and Alu monomers show significant homology to the 5' and 3' ends of the 7SL RNA gene which encodes the RNA component of the signal recognition particle, a ribonucleoprotein involved in protein transport across the endoplasmic reticulum (Ullu & Tschudi, 1984). The source of all Alu elements would therefore appear to be a 7SL RNA pseudogene which has dimerized following deletion of the central region.

Comparison of Alu elements in man and other primates has revealed several distinct Alu subfamilies which are distinguishable by characteristic base differences (Britten *et al.*, 1988, 1989; Shen *et al.*, 1991). Each subfamily appears to have arisen at a particular stage of human evolution, strongly implying that all Alu elements are derived from a single active source gene (or limited number of source genes) and that the different subfamilies simply represent the accumulation of base substitutions in the source gene(s) during evolution (Shen *et al.*, 1991). This is in agreement with observations that few Alu elements are actively transcribed *in vivo* despite the the presence of an internal RNA polymerase III promoter (see Sharp, 1983). It has been estimated that the 'master' Alu element has a mutation rate similar to that of non-coding intervening sequence DNA and is consequently under no evolutionary constraints (Shen *et al.*, 1991). The internal promoter may therefore be defunct, suggesting that the master Alu might fortuitously reside in a genomic region which is highly transcriptionally active due to RNA polymerase II activity at the appropriate time for retroposon dispersal;

for example, in a region which is highly expressed in the early embryo (see below).

The second large class of dispersed repeats in the mammalian genome, the long interspersed elements, or LINES, are again dominated by a single element, the L1 element. The L1 family is ubiquitous in mammals (see Singer, 1982), and is sometimes referred to as the Kpn family in primates (Adams, 1980) and the MIF1 family in rodents (Brown & Dover, 1981). In man, this element has a length of 6.4kb, although elements are frequently truncated at the 5' end. There are estimated to be ~10^5 copies in human DNA (Singer, 1982). It has recently been demonstrated that L1 mobility involves an RNA intermediate, formally establishing these elements as retroposons (Evans & Palmiter, 1991). Unlike Alu and B1 elements, L1 elements have no internal RNA polymerase III priming site, and are generally thought to be fortuitously expressed by RNA polymerase II (Sun *et al.*, 1984). Two open reading frames have been identified in L1 elements, one of which bears some homology to the reverse transcriptase gene of certain retroviruses (Singer & Skowronski, 1985) leading to speculation that L1 elements provide their own reverse transcriptase activity following transcription.

The Alu and L1 families are the most prolific dispersed repeats in human DNA and together probably constitute ~10% of the genome. However, it has been estimated that there may be up to 1000 families of short (~250bp) dispersed repeats each with 200-2000 members per family (Sun *et al.*, 1984). Six such families, designated medium reiteration frequency or MER sequences have been identified following DNA database searches, and together comprise >17,000 new dispersed elements (Jurka, 1990). Other dispersed elements include solitary long terminal repeats (LTRs) from transposon and retroviral-like elements. Several endogenous retrovirus-like elements have been identified in human DNA, many as multiple (10-100) copy elements (Martin *et al.*, 1981; Bonner *et al.*, 1982; Steele *et al.*, 1984; Maeda, 1985). In addition, a solitary retroviral LTR-like element has been identified at moderate (~300) copy number (Noda *et al.*, 1982). (Solitary LTRs are thought to arise through deletion of a loop of DNA containing the provirus following homologous recombination between LTRs). A dispersed element with a copy number of ~10^4 and which appears to be a transposable element has also been

characterized from human DNA (Paulson *et al.*, 1985). The first such sequence identified in man, this element is 2.3kb long, has 350bp LTRs and has been designated THE (transposon-like human element). The LTRs are present at a copy number of ~30,000, suggesting that ~10,000 solitary LTRs are scattered throughout the genome.

Transcripts from dispersed repeats are abundant in teratocarcinoma cell lines, which are though to represent very early embryonic stages (Skowronski & Singer, 1985; La Mantia, 1989) and B2 elements have been shown to be transcribed in the early mouse embryo (Vasseur *et al.*, 1985) strongly suggesting that transposition events occur very early in development (a necessity for passage into the germ-line, unless transcription and transposition events occur in the germ-line itself). It has been proposed that endogenous retroviruses and transposable elements may also be transcribed in the early embryo, providing the necessary reverse transcriptase activity for reintroduction of retroposon transcripts into the genome (Rogers, 1986). Retroposons show a preference for insertion into A/T rich sites, and frequently insert into polyadenosine tails, giving a clustering effect (Rogers, 1986). Dispersed repeats are also frequently associated with tandem repeats, and on occasion tandem repeats actually arise from within dispersed repeats (Armour *et al.*, 1989; see chapters 3 and 7), suggesting that certain regions of the genome may be prone to the accumulation of such sequences. The genomic role of retroposons is debatable, but it is now largely accepted that such elements are likely to be 'selfish' or 'parasitic', ie self-propogating at the expense of the host genome (Doolittle & Sapienza, 1980; Orgel & Crick, 1980), or 'ignorant' - a passive by product of DNA turnover mechanisms (Dover, 1980).

Although documented incidences of novel retroposon insertion are rare, some have been identified as the direct cause of human genetic disease. A novel insertion of an L1 element into the clotting factor VIII gene has been shown to be the cause of a case of haemophilia A (Kazazian *et al.*, 1988). Similarly, a new insertion of an Alu element into the factor IX gene has been shown to be responsible for a case of haemophilia B (Vidaud *et al.*, 1988). In addition to influencing gene function by insertion, dispersed repeats in close approximation may also act as substrates for deletion and insertion events by unequal exchange. One

example is given by Nicholls *et al.* (1987), who describe a 62kb deletion involving the 5' end of the α-globin locus, resulting from unequal exchange between two Alu elements. Similarly, an unequal recombination event involving two Alu repeats has been shown to be responsible for a duplication in the LDL receptor gene, resulting in familial hypercholesterolaemia (Lehrman, 1987).

1.3 Highly repetitive 'satellite' DNA

The term 'satellite' DNA was first given to a genomic DNA fraction which migrated to a different position to the bulk of DNA on isopycnic centrifugation, due to a different cytosine/guanosine content, and hence gave a characteristic 'satellite' peak on fractionation (Sueoka, 1961; Kit, 1961). This fraction was subsequently found to be highly repetitive and analagous to highly repetitive sequence identified by reassociation kinetics. Not all large, highly repetitive DNA tracts migrate to a unique position on isopycnic centrifugation, but all are now known as satellite DNA.

At least five different classes of human satellite DNA have now been identified (as categorized by Tyler-Smith & Brown, 1987). By far the most abundant is the so-called α-satellite, consisting of a 171bp tandemly reiterated sequence which is ubiquitous in primates (Rosenberg, 1978). α-satellite sequences are associated with heterochromatin at the centromeres of human chromosomes (Manuelis, 1978), perhaps suggesting a role in chromosome pairing at meiosis, and extend for tracts of hundreds of kilobases (Waye & Willard, 1986; Tyler-Smith & Brown, 1987; Waye *et al.*, 1987). The most likely method of maintenance of repeat unit fidelity in satellite DNA is unequal sister chromatid exchange as described by Smith (1976). This mechanism is further implied by the identification of higher-order chromosome specific tandem repetition likely to be due to propagation of a novel (mutant) repeat unit types by crossover fixation (Willard & Waye, 1987).

1.4 Telomeres

A tandem array with a more obvious function is the telomeric repeat found at the ends of chromosomes. In man, this consists of a $(TTAGGG)_n$

array of approximately 10kb (Moyzis *et al.*, 1988). Telomere repeats can be added *de novo* to the ends of chromosomes by a ribonucleoprotein known as telomerase, a specialized reverse transcriptase which carries the template for telomere synthesis in its RNA component (Greider & Blackburn, 1989). This *de novo* addition of repeats maintains chromosome length in the germ-line; successive rounds of cell division would otherwise erode the telomeres due to the requirement of a primer for DNA synthesis of the lagging strand during replication (see Blackburn, 1991). Telomerase activity appears to be absent in somatic cells, resulting in such successive shortening of chromosomes; telomere length has been shown to be a function of age in both cultured cells and living individuals (Harley *et al.*, 1990; Hastie *et al.*, 1990). This has lead to speculation that telomere shortening, leading to chromosome instability, may be a direct cause of cell senescence; however, the telomeres of mouse chromosomes are 5-10 times longer than those of man and do not grow perceptibly shorter with successive cell divisions in somatic tissue (Kipling & Cooke, 1990). Several interstitial tracts of telomere-like sequence have been identified in human and mouse chromosomes (Hastie & Allshire, 1989). These tracts may result from internalization of telomeric sequences, or could simply reflect independent expansion of telomere-like tandem repeat arrays. In one case, an interstitial $(TTAGGG)_n$ repeat appears to have resulted from the ancient telomeric fusion of two acrocentric chromosomes which gave rise to human chromosome 2 (Allshire *et al.* 1989).

1.5 Ribosomal RNA genes

Other tandemly repeated arrays with obvious function are the ribosomal RNA genes; multiple copies of these genes are required to ensure that a sufficient number of ribosomes can be synthesized for translational activity. The 18S and 28S rRNA genes are combined in a single transcription unit, with each transcription unit separated by a non-transcribed spacer, to give an array of 43kb tandem repeats in man (Arnheim & Southern, 1977). These tracts are dispersed over the short arms of the acrocentric chromosomes (Henderson *et al.*, 1972) with a total copy number of 50-200 (Young *et al.*, 1976; Guabatz *et al.*, 1975). The 5S rRNA genes are also encoded as a tandem repeat, thought to exist

as a single locus on chromosome 1, comprising some 2000 repeats (Steffensen *et. al.*, 1974; Hatlen & Attardi, 1971).

1.6 Minisatellites - dispersed tandem repeats

(a) Historical aspects

Minisatellites, also known as VNTR (variable number of tandem repeat) loci are tandemly repeated DNA tracts, typically 1-20kb in total length, which can show extensive allelic length differences due to variation in repeat copy number (reviewed in Jeffreys, 1987). The first such hypervariable locus was identified in 1980 by Wyman & White, although the molecular basis for variability at this locus was not established until 1986 (Wyman *et al.*, 1986). Similarly, highly polymorphic loci, subsequently shown to be minisatellites, were identified in and around the α-globin gene complex (Higgs *et al.*, 1981; Goodburn *et al.*, 1983). The first demonstration that high levels of allelic variation were associated with tandem repetition resulted from sequence analysis of a highly polymorphic region near the human insulin gene (Bell *et al.*, 1982).

In 1985, a 33bp tandemly repeated tract, which had previously been isolated from an intron in the myoglobin gene, was used as a probe at low stringency to detect further minisatellites in human DNA, many of which were polymorphic (Jeffreys *et al.*, 1985a). Sequence analysis of these minisatellites revealed a common 15bp G-rich consensus or 'core' motif embedded in the repeat unit sequence of each minisatellite. Two of the minisatellites, designated 33.6 and 33.15, were found to consist entirely of tandem repeats of two variant forms of this core sequence. When used to probe Southern blots of restriction endonuclease cleaved human DNA at low stringency, minisatellites 33.6 and 33.15 detect two different subgroups of highly variable minisatellites, resulting in highly individual-specific 'DNA fingerprints' (Jeffreys *et al.*, 1985b).

Subsequent to the generation of DNA fingerprints with 33.6 and 33.15, many other tandem repeat probes were found to detect multiple variable loci at reduced stringency including probes derived from a minisatellite 3' to the α-globin gene complex (Fowler *et al.*, 1988), a tandem repeat from the M13 genome (Vassart *et al.*, 1987) and several chemically

synthesized simple repeat sequences and polymers of random oligonucleotides (Ali *et al.*, 1986; Vergnaud, 1989). However, many of the minisatellites detected by these probes have now been shown to be co-detected by either 33.6 or 33.15 (Armour *et al.*, 1990).

(b) Applications of DNA fingerprints

By far the most common application of multilocus DNA fingerprints is the establishment of family relationships either in cases of disputed paternity (see Jeffreys *et al.*, 1991a) or immigration disputes (Jeffreys *et al.*, 1985c). Using DNA fingerprinting, family relationships can be established with a level of certainty far in excess of that given using conventional genetic markers (see Dodd, 1985). DNA fingerprints have also been used in forensic science, for the comparison of semen samples with suspected rapists (Gill *et al.*, 1985; Gill & Werrett, 1987), although the usefulness of multilocus probes is limited in such cases by lack of sensitivity, requiring at least 0.5μg of DNA per test (Jeffreys *et al.*, 1985b). This problem has been overcome by the use of single locus minisatellite probes (see below), which have 10-fold greater sensitivity (Wong *et al.*, 1987). Other applications of DNA fingerprinting include monitoring the success of bone marrow transplants (Thein *et al.*, 1986), determination of twin zygosity (Hill & Jeffreys, 1985) and the detection of chromosomal rearrangements in tumours (Thein *et al.*, 1987).

In addition to detecting multiple variable loci in human DNA, multilocus probes have been shown to generate DNA 'fingerprints' from a wide range of animal and plant species, including mice, dogs, cats, farm animals and birds (Jeffreys *et al.*, 1987a, 1987b; Jeffreys & Morton, 1987; Burke & Bruford, 1987, Dallas, 1988). Thus DNA fingerprints can be used to verify identity and parentage in economically important species, as well as in the establishment of family relationships in population biological studies (see for example Burke & Bruford, 1987; Parker *et al.*, 1991; reviewed in Burke *et al.*, 1991).

(c) The potential of minisatellites as linkage markers

A DNA fingerprint generated by a single multilocus probe is comprised of approximately 15 resolvable, unlinked bands (Jeffreys *et al.*, 1985a),

but provides no information as to the genetic location of each band; as such, a DNA fingerprint gives phenotypic, rather than genotypic information. In order to exploit the hypervariable nature of the minisatellites that make up a DNA fingerprint in genetic linkage analysis, it is necessary to clone individual loci.

An aside: the principle of linkage analysis

Due to the linear arrangement of DNA in eukaryotic chromosomes, two DNA markers on the same chromosome can be said to be physically linked to one another (syntenic). The proximity of syntenic markers is inversely proportional to the number of recombination events occurring between them (Morgan, 1910). This simple relationship forms the basis of genetic linkage analysis; the degree of linkage is usually determined by comparing the probability of obtaining a given set of pedigree data for two loci at a given recombination fraction with the probability of obtaining the same results in the absence of linkage (ie at a recombination fraction of 0.5). This expression is usually given as the logarithm of the odds in favour of linkage for the given recombination fraction over no linkage, and is termed the LOD (log of the odds) score (Fisher, 1935; Morton, 1955).

The 'reverse genetics' or 'positional cloning' approach to identifying gene defects associated with inherited disease involves identifying polymorphic markers linked to the gene of interest and assessing the degree of linkage between markers and the disease trait in affected families. Sequential 'chromosome walking' through informative markers plus physical mapping allows eventual identification of the gene itself. Consequently, the early stages of this approach depend heavily on an even and frequent scattering of highly informative markers across the chromosomes. The informativeness of a genetic marker is usually defined in terms of heterozygosity (ie the proportion of heterozygotes in the population) or the polymorphism information content (PIC), which describes the probability that an allele from a given parent can be identified on transmission to a given child (Botstein *et al.*, 1980). For highly variable loci, the PIC value is approximately the same as the heterozygosity.

The first polymorphic system to be studied was the ABO blood group system, which was identified serologically (Landsteiner, 1900). Subsequently, electrophoretic variants of proteins, including further cell surface antigens and enzymes were identified (see Harris & Hopkinson, 1972; Giblett, 1977). Unfortunately the majority of these systems are of low informativeness, with a limited number of alleles, the clear exception being the human leucocyte antigens (HLA) which are encoded as a cluster on chromosome 6 and generate an extremely informative haplotype (see Bodmer, 1981). In the late 1970s, it became possible to analyse polymorphism directly at the DNA level; nucleotide sequence variations were shown to occur once every few hundred base pairs on average (Jeffreys, 1979). These base substitutions are most easily exploited as genetic markers if they result in creation or destruction of a restriction endonuclease cleavage site, allowing identification of a restriction fragment length change on Southern blot analysis. This technique was first used to indirectly diagnose sickle-cell anaemia by virtue of a *Hpa*I site linked to the defective β-globin locus (Kan & Dozy, 1978). Restriction fragment length polymorphisms (RFLPs) have since been used to generate reasonably detailed linkage maps of the human chromosomes (Donis-Keller *et al.*, 1987) and have been instrumental in locating those genes responsible for some of the most common inherited diseases in man, including Duchenne muscular dystrophy (Monaco *et al.*, 1986) and cystic fibrosis (Rommens *et al.*, 1989).

New techniques, based on PCR, have broadened the range of single base polymorphisms which may be utilized beyond RFLPs, which rely on the presence or absence of a restriction site. A single base mismatch at the extreme 3' end of a PCR primer can completely prevent priming under some circumstances, a phenomenon which may be used to distinguish alleles differing due to a single base substitution (Newton *et al.*, 1989). PCR amplified alleles which differ by a single base may also be distinguished following denaturation and non-denaturing gel electrophoresis; the single stranded DNA produces different profiles for the two alleles as a consequence of conformational polymorphism resulting from the base difference (Orita *et al.*, 1989).

Although useful, single base polymorphisms are far from ideal markers for linkage analysis. Most are diallelic and consequently only three

phenotypic states exist. Heterozygosity can never exceed 50% in a population at Hardy-Weinberg equilibrium, and in practice is usually much lower. This severely limits the usefulness of single base polymorphisms in linkage analysis. In contrast, highly variable minisatellite arrays frequently show heterozygosities in excess of 90% (Wong *et al.*, 1986, 1987; Armour *et al.*, 1990) rendering them extremely promising linkage markers in this respect.

(d) Isolation of hypervariable minisatellites

Tandemly repeated DNA sequences are notoriously difficult to propagate in *E. coli* hosts, being prone to gross rearrangements usually resulting in deletion of much of the repeat array (Brutlag *et al.*, 1977). Minisatellites are no exception, and are extremely refractory to cloning, even in recombination deficient hosts (see Wyman *et al.*, 1985; Wong *et al.*, 1986). In spite of this, a number of highly informative minisatellite loci have been successfully cloned from genomic λ libraries following detection with the polycore probes 33.6 and 33.15 (Wong *et al.*, 1986, 1987; Nicola Royle, personal communication), and over 200 VNTR loci (albeit with somewhat lower mean variability) have been isolated from a genomic cosmid library (Nakamura *et al.*, 1987a). More recently, ordered array libraries in charomid vectors, amenable to multiple rounds of efficient screening with a battery of multilocus probes, have proved a rapid and efficient method for the isolation of minisatellites from human DNA (Armour *et al.*, 1990), and from the genomes of other species (Hanotte *et al.*, 1991; Burke *et al.*, 1991). Charomids are a series of cosmid derived vectors which include variable numbers of tandemly repeated 'spacers', allowing cloning of a range of DNA fragment sizes (Saito & Stark, 1986). This renders charomids relatively insensitive to reduction of insert size due to collapse, and consequently they are ideally suited to the cloning of minisatellites.

(e) Localization of minisatellites to the proterminal regions of human chromosomes

Unfortunately, despite the initial promise of random genetic distribution of minisatellites as shown by a lack of linkage between bands in a DNA fingerprint (Jeffreys *et al*, 1986), *in situ* hybridization and pedigree

analysis has shown that approximately 80% of minisatellites are found in the proterminal regions of human chromosomes (Royle *et al.*, 1987; Nakamura *et al.*, 1988; Armour *et al.*, 1990). Although this renders minisatellites of restricted usefulness in the construction of linkage maps, it adds weight to the argument favouring the involvement of minisatellites in homologous recombination (see below).

(f) Forensic application of single locus minisatellite probes

Despite apparently limited potential in linkage analysis, single locus minisatellite probes are of unparalleled value in forensic science, particularly in instances of rape, where the increased sensitivity over multilocus DNA fingerprinting probes mentioned above allows a profile to be generated from as little as 1µl of semen. Although not as discriminatory as a DNA fingerprint, sequential use of a panel of highly variable minisatellites gives a more than acceptable level of individual specificity (Wong *et al.*, 1987).

The sensitivity of single locus minisatellite typing can be extended further using the polymerase chain reaction (PCR), which in principle allows identification of a single cell (Jeffreys *et al.*, 1988). However, alleles >10kb long are refractory to amplification even from high molecular weight DNA, and minisatellites generally fail to amplify from degraded DNA samples (A.J. Jeffreys, personal communication). For this reason, much shorter PCR-amplifiable dinucleotide repeats may be preferable for analysis of degraded DNA, despite limited variability (see chapter 6). While PCR extends the range of forensic samples to which DNA analysis can be applied, it also presents unprecedented opportunities for erroneous conclusions through sample contamination from extraneous sources, and consequently requires strict laboratory procedures designed to prevent such occurrences.

(g) Minisatellites as potential recombination hot-spots

In addition to providing an invaluable new set of DNA markers for forensic science, the isolation of individual minisatellites also allowed some basic aspects of minisatellite biology to be explored. The ubiquity of G-rich minisatellites in higher eukaryotes initially implied conservation

and therefore function, a view supported to an extent by the identification of minisatellite binding proteins in a diverse range of species (Collick & Jeffreys, 1990; Wahls *et al.*, 1991). It had already been noted that the core sequence embedded within the repeat units of those minisatellites that constitute a 33.6 or 33.15 derived DNA fingerprint bore some similarity to the χ recombination signal of *E. coli*, leading to speculation that minisatellites may play a similar recombination-promoting role in eukaryotes, the variation in the number of tandem repeats arising as a consequence of unequal meiotic or mitotic exchange (Jeffreys *et al*, 1985a).

The localization of minisatellites to the proterminal regions of human chromosomes fuelled this speculation; chiasmata and recombination nodules associated with synaptonemal complexes, thought to be the visible cytological representation of meiotic recombination events, show a similar distribution (Hulten, 1974; Solari, 1980) and minisatellite core probes appear to hybridize preferentially to chiasmata in human bivalents (Chandley & Mitchell, 1988). Furthermore, a recombination hotspot identified near the murine Eβ MHC locus may involve a tandemly repeated region closely related in sequence to the minisatellite core (Steinmetz *et al.*, 1986). Additional evidence pointing to a role for minisatellites in recombination comes from Wahls *et al.* (1990); plasmids with inserts consisting of tandem repeats based on the core sequence show an elevated level of homologous recombination in adjacent sequences when introduced into EJ bladder carcinoma cells. Finally, direct estimates of minisatellite mutation rates to new length alleles in human pedigrees have shown that paternal and maternal mutations arise with equal frequency, surprising in view of the fact that oocytes arise through approximately 24 postzygotic cell divisions, whereas sperm are generated via ~400 divisions (Orgel & Rathenberg, 1975), and suggestive that the mutation process is restricted to one stage of gametogenesis, possibly meiosis (Jeffreys *et al.*, 1988).

However, there is also a considerable body of evidence against meiotic recombination as the major mutational mechanism operating at minisatellite loci. The generation of 13 new alleles at two minisatellite loci has failed to show any association with exchange of flanking markers, strongly suggesting that unequal exchange between homologous

chromosomes is not the major mechanism for the generation of new length alleles (Wolff *et al.*, 1988, 1989). Minisatellites can also show a significant degree of mutational mosaicism in both the germ line and somatic tissues; two highly unstable mouse minisatellites, *Ms6-Hm* and *Hm2* show clear length mosaicism in approximately 3% and 10% of mice respectively, with 10-60% of cells containing the new mutant allele. The events leading to the generation of these mutant alleles are clearly premeiotic, occurring early in development (Kelly *et al.*, 1989, 1991; M. Gibbs, personal communication). Although there has only been one instance of minisatellite mosaicism in man detectable by conventional Southern blotting techniques to date (A.J. Jeffreys and I. Patel, personal communication), single molecule PCR amplification of new mutant alleles of minisatellite MS32 from both sperm and blood DNA has revealed a significant level of mosaicism, with at least 40% of new mutants showing evidence of mosaicism (Jeffreys *et al.*, 1990).

(h) Internal variation at minisatellite loci

The tandemly repeated units that constitute a minisatellite are rarely identical, and frequently show sequence variation between repeat units (Jeffreys *et al.*, 1985a; Wong *et al.*, 1986, 1987; Jeffreys *et al.*, 1990; see chapter 3). This variation can be exploited to generate internal maps of minisatellite alleles, which reveal a far greater level of variability than that discernible by length differences; for example, the number of different alleles distinguishable by length at minisatellite MS32 has been estimated at approximately 40 (Wong *et al.*, 1987), whereas the true number of alleles in the human population as judged by internal variation may be greater than 10^8 (Jeffreys *et al.*, 1991b).

Study of the internal structure of minisatellite alleles, rather than measuring length differences alone, can give further insights into the biological processes operating at these loci. Internal maps of deletion mutants at MS32 isolated from sperm DNA and subsequently analysed following PCR amplification (see above) revealed all mutants isolated to be the result of intra-allelic events, with no evidence for inter-allelic recombination, and indicated that <6% of all such deletion mutants at MS32 arise through inter-allelic mechanisms, drawing the conclusion that MS32 alleles evolve largely along haploid lineages via mechanisms such

as sister chromatid exchange and DNA slippage during replication (Jeffreys *et al.*, 1990; discussed further in chapter 3). However, for technical reasons only large deletion events were studied; analysis of increases in allele length and of very small changes in allele length, known to be far more common, were not included. Very recent evidence resulting from extensive pedigree analysis of MS32 internal maps has revealed 7 mutant alleles, all resulting from very small increases in repeat unit copy number and perhaps reflecting a bias in favour of gain of small numbers of repeats. Furthermore, at least one and probably two of these mutants have arisen through inter-allelic exchange, revitalizing the concept of a role for minisatellites in meiotic recombination (Jeffreys *et al.*, 1991b). The mutational processes operating at minisatellite loci may therefore be more complex than previously imagined, possibly involving several different mechanisms including both intra and inter-allelic exchange.

In addition to providing an insight into the fundamental biological processes operating at minisatellite loci, minisatellite variant repeat (MVR) mapping is also potentially a very powerful tool in forensic science, overcoming many of the criticisms recently levelled at individual identification based on minisatellite allele length (Lander, 1989). The greatest limitation of minisatellites as currently used in forensic analysis is the inability to unequivocally identify an allele of given length due to subtle length differences between similarly sized alleles. Consequently, in database construction it is necessary to group or 'bin' alleles of similar length, with a subsequent large reduction in statistical power (Budowle *et al.*, 1991). Even when comparing two profiles side by side on the same gel, electrophoretic 'band shifts' can lead to identically sized alleles migrating different distances, giving a false exclusion (see Norman, 1989). Furthermore, the statistical evaluation of a match between two profiles requires the estimation of allele frequencies (as judged by length) and the assumption that alleles associate at random in human populations, an assumption that has recently been brought into question (Cohen, 1990). In contrast, MVR mapping generates unambiguous, non-subjective digital codes, allowing data from different laboratories to be pooled for the construction of large databases for forensic interrogation. In addition, MVR maps of MS32 generated from genomic DNA samples consist of a ternary code resulting from superimposition of the two allelic profiles,

effectively generating phenotypic rather than genotypic data and consequently obviating the need for allele frequency estimates and the assumption of random allele association (Jeffreys *et al.*, 1991b).

1.7 Simple tandemly repeated sequences

Following the finding that minisatellites can show a high degree of length variability, the possibility of variation at simple tandemly repeated loci has been explored. Simple tandem repeats may be loosely defined as tandemly repeated arrays where the length of the repeat unit is $\leq$ 4bp, and the overall array length is typically much shorter than a minisatellite, usually less than 1kb. However, the dividing line between simple tandem repeats and minisatellites is arbitrary; the highly variable mouse minisatellite *Hm2* has a repeat unit just 4bp long but alleles several kb in length (Kelly *et al.*, 1991; M. Gibbs, personal communication). Furthermore, many tracts of satellite DNA, particularly in arthropods, consist of simple tandem repeats (see John & Miklos, 1988).

The first simple tandem repeat sequences to be assayed for length variability were $(dA\text{-}dC)_n\cdot(dT\text{-}dG)_n$ dinucleotide repeats, where n is 15-30 (Litt & Luty, 1989; Weber & May, 1989). Such elements are widely dispersed in eukaryote genomes (for more details see section 5.1). As these elements are frequently embedded in unique sequence DNA, PCR was used to assay variability. It was consequently found that $(dA\text{-}dC)_n\cdot(dT\text{-}dG)_n$ repeat loci or 'microsatellites' are indeed variable, with a range of alleles typically varying by 2bp or one repeat unit, consistent with variability arising as a consequence of DNA slippage during replication (Litt & Luty, 1989; Weber & May, 1989). Unfortunately the degree of variability at such microsatellite loci is limited; heterozygosity is relatively low compared to minisatellites, and the number of alleles is limited, with allele frequency distributions typically dominated by one or two common alleles (Weber, 1990). As a consequence, microsatellites are of limited usefulness in forensic science, although they have the advantage of being applicable to the typing of highly degraded DNA samples (see chapter 6). However, the genomic distribution of microsatellites appears to be random, with no evidence for clustering (Luty *et al.*, 1990), rendering them extremely promising markers for linkage analysis.

Further to the discovery of variability at $(dA\text{-}dC)_n\cdot(dT\text{-}dG)_n$ repeat loci, other short di- tri- and tetranucleotide repeat arrays have also been shown to be variable (Tautz, 1989; see chapter 7; see for example Mercier *et al.*, 1991). Even mononucleotide repeats such as Alu element polyadenosine tails can show significant degrees of variability (Economou *et al.*, 1989; see chapter 7; Mant *et al.*, 1991). Tetranucleotide repeat arrays can show high levels of heterozygosity and smoother allele frequency distributions than than dinucleotide repeats; furthermore, longer repeat unit length and greater overall array length allows tetranucleotide repeats to be typed on high percentage agarose gels, largely removing the need for typing on cumbersome polyacrylamide sequencing type gels, as is required for resolving closely spaced alleles at dinucleotide repeat loci (although polyacrylamide gels are required for unequivocal allele identification in, for example, a forensic context).

1.8 This work

Although polymorphic tandemly repeated loci appear to be a ubiquitous feature of the genomes of higher eukaryotes, very little information exists concerning the evolutionary dynamics of such loci. By comparing individual minisatellite loci in man and other primates, the degree of evolutionary persistence of the variable state was ascertained and related to the level of variability and mutation rate, allowing a simple model of the evolutionary progress of a minisatellite, starting from a two repeat unit ground state, to be proposed. This is described in chapter 3. Following the success in mapping internal variability at the minisatellite MS32 (Jeffreys *et al.*, 1991b), chapter 4 describes an attempt to exploit internal variation as a source of polymorphism at a second minisatellite locus, MS1.

Chapter 5 concerns the isolation and characterization of $(dC\text{-}dA)_n.(dT\text{-}dG)_n$ microsatellite loci as potential genetic markers, and also touches on the persistence of variability at these loci as determined by species comparisons. Chapter 6 involves the application of such loci to forensic science, and describes the positive identification of the skeletal remains of a murder victim following the extraction of highly degraded DNA from bone, marking the first instance of PCR data being accepted as evidence in a British court.

In chapter 7, the isolation and characterization of tetranucleotide repeat arrays is described, and their potential as genetic markers assessed. Finally, chapter 8 describes the characterization and apparent extreme evolutionary transience of two 'midisatellite' arrays fortuitously isolated from human DNA.

CHAPTER 2

MATERIALS & METHODS

2.1 Materials

2.1.1 Chemical Reagents

All chemical reagents, unless otherwise stated below, were supplied by Fisons, Loughborough. Antibiotics, bovine serum albumin (BSA), N-2-hydroxyethylpiperazine-N'-2-ethanesulphonic acid (HEPES), isopropyl-β-D-galactopyranoside (IPTG), polyethylene glycol (PEG), ficoll 400, N,N,N',N'-tetramethylethylenediamine (TEMED), dithiothreitol (DTT), thymine, spermidine trichloride, salmon sperm DNA and agarose were supplied by the Sigma Biochemical Company, Poole. Seachem HGT agarose and nusieve agarose were supplied by ICN Biochemicals Ltd, High Wycombe. Ammonium persulphate was from Bio-Rad, Watford. 5-bromo-4-chloro-3-indoyl-β-D-galactopyranoside (BCIG) was obtained from Anglian Biotechnology, Colchester. Acrylamide and urea were supplied by Serva, Heidelberg. N,N'-methylene-bisacrylamide was supplied by Uniscience, Cambridge. Deoxyribonucleotides and dideoxynucleotides were from Pharmacia, Milton Keynes. Radiochemicals were supplied by Amersham International plc, Little Chalfont. All reagents used were of analytical grade.

2.1.2 Oligonucleotides

Hexadeoxyribonucleotides for random oligonucleotide priming (section 2.2.10a) were supplied by Pharmacia. 20 and 24-mers for PCR amplification and 17-mers for sequencing M13 and phagemid inserts were synthesized in the Department of Biochemistry, University of Leicester on an Applied Biosystems 380B DNA synthesizer.

2.1.3 Enzymes

Restriction endonucleases were supplied by Gibco-BRL, Paisley, New England Biolabs (via CP Laboratories, Bishop's Stortford) and the Boehringer Corporation, Lewes. Digests were performed in ReAct buffers obtained from Gibco-BRL. Pharmacia supplied T4 polynucleotide kinase, T7 DNA polymerase and the klenow fragment of DNA polymerase I. T4 ligase was supplied by Gibco-BRL. Calf intestinal alkaline phosphatase was obtained from the Boehringer Corporation.

DNA polymerase from *Thermophilus aquaticus* was from Amersham International.

2.1.4 Molecular weight markers

λ DNA digested with *Hind*III and ϕx174 DNA digested with *Hae*III were obtained from Gibco-BRL.

2.1.5 Media

All media were obtained from Oxoid, Basingstoke, with the exception of yeast extract and tryptone, which were supplied by Difco, East Molesley.

2.1.6 Bacterial strains

Two bacterial strains were used, NM522: Δ(*lac-pro*AB), *thi*, *sup*E, *hsd*R17 (Rk^-,Mk^+), [F' *pro*AB, lacI^{q}ZΔM15] (Gough & Murray, 1983) and XL1Blue: *end*A, *hsd*R, *sup*E, *thi*, *rec*A, *gyr*A, *rel*A [F' *pro*AB, *lac*I^{q}ZΔM15 Tn10(Tet^r)] (Bullock *et al.*, 1987), supplied by Stratagene Cloning Systems Inc.

2.1.7 Cloning vectors

M13 sequencing vectors M13mp18 and M13mp19 (Yanisch-Perron *et al.*, 1985) were used, replicative forms supplied by Gibco-BRL. The phagemid vectors pBluescript II KS^+ and SK^+ (Short *et al.*, 1988) from Stratagene were also employed.

2.1.8 DNA samples

Human DNA samples from lymphoblastoid cell lines derived from large families were supplied by Professors Howard Cann and Jean Dausset of the Centre d'Etude du Polymorphisme Humain (CEPH), Paris. Human placental DNA samples were provided by Dr. Raymond Dalgleish, Leicester. Primate DNA samples had previously been prepared using procedures described in Jeffreys & Flavell (1977) and Jeffreys (1979). Blood samples from the great apes had been supplied by Dr A.F. Scott (John Hopkins University, Baltimore, USA), and post-mortem tissue

samples from other primates had been obtained from the Zoological Society of London and the Jersey Wildlife Preservation Trust. DNA samples from human/rodent hybrid somatic cell lines for chromosomal assignments were provided by Dr. Sue Povey, University College, London, and are described in table 2.1.

2.2 Methods

2.2.1 Estimation of DNA concentration

DNA concentration was assayed by measurement of UV absorbance at a wavelength of 260nm in a Cecil Instruments CE235 spectrophotometer, given that 0.02 OD units represents 1μg/ml of DNA (or 0.05 units for 1μg/ml of oligonucleotides). Alternatively, DNA concentration was estimated by visual comparison following agarose gel electrophoresis against a standard range of samples of known concentration.

2.2.2 Ethanol precipitation

In order to concentrate, de-salt or recover DNA following manipulation, ethanol precipitation was used. 1/10 volume of 2M sodium acetate (pH 5.6) and 2.5 volumes of ethanol were added to the DNA solution. The solution was then mixed and left on ice for 20 minutes or at -80° for 5 minutes, followed by centrifugation at 15,000rpm for 10 minutes. The ethanol was removed and the DNA pellet rinsed in 80% ethanol, dried under vacuum and redissolved in the required amount of distilled water. For recovering DNA molecules <200bp, $MgCl_2$ was added to a final concentration of 0.01M prior to addition of ethanol, and centrifugation time was increased to 20 minutes.

2.2.3 Restriction endonuclease digests

Restriction digests were performed in the appropriate ReAct buffer (Gibco-BRL) with ≥1unit of enzyme per μg of DNA. Incubations were carried out at the temperature recommended by the enzyme suppliers, usually (but not invariably) 37°, for 1-15 hours. Spermidine trichloride was added to a final concentration of 4mM to enhance enzyme activity

Table 2.1. Human-rodent somatic cell hybrids used for chromosomal assignments. References: (1) Kielty *et al.*, 1982; (2) Edwards *et al.*, 1985; (3) Solomon *et al.*, 1979; (4) Swallow *et al.*, 1986: (5) Goodfellow *et al.*, 1982; (6) Jones *et al.*, 1980; (7) Varesco *et al.*, 1989; (8) Sykes & Solomon, 1978; (9) Sue Povey, University College, London, personal communication; (10) Van Heyningen *et al.*, 1975; (11) Croce & Koprowski, 1974; (12) Edwards *et al.*, 1986; (13) Solomon *et al.*, 1983.

HYBRID	CHROMOSOME																					
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X
FG10E8B[1]		+			+			+							+			+			+		
FST9/10[1]			+	+		+		+	+	+		+	+	+	+			+		+		+	+
SIF15P5[2]		+				+	+			+				+	+					+			+
SIF4A31[2]				+	+	+								+									+
MOG2C2[3]	+		+	+	+		+	+	+	+	+			+	+	+	+	+	+	+	+	+	+
MOG2E5[3]	+		+	+	+	+	+	+	+	+		+	+	+	+	+	+	+			+	+	+
TWIN19C5[4]			+	+				+				+		+			+	+		+	+	+	
TWIN19F9[4]	+	+	+	+		+		+	+			+		+			+	+		+	+	+	
TWIN19F6[4]		+	+	+		+		+	+			+		+			+	+		+	+	+	
TWIN19D12[4]	+		+	+		+		+	+			+		+		+	+	+		+	+	+	
MCP6[5]						+																	+
SIF4A24E1[2]				+													+				+		+
640-63a12[6]									q														
PNTS1[7]					+																		
F4sc12sc13[8]	p													+									+
1αA9602+[9]						+						+									+		+
HORL411B6P[10]	+		+								+		+		+								
Clone 21[11]							+																
C4A/G[12]								+															
FIR5R3A4[13]														+									+

and specificity (Pingourd, 1985). For high molecular weight DNA, spermidine was added after the digestion had progressed for 5 minutes, to prevent DNA precipitation by the spermidine.

2.2.4 Alkaline phosphatase treatment

To remove the terminal 5' phosphate groups from cloning vector DNA following endonuclease digestion (to prevent recircularization on ligation to inserts), approximately 0.01 units of calf intestinal alkaline phosphatase (CIP) per picomole of 5' ends of DNA were added directly to the reaction mix and incubation continued at 37° for 30 minutes. The CIP was then deactivated by heating to 68° for 10 minutes. (Modified from Maniatis *et al.*, 1982).

2.2.5 Ligations

Ligation of DNA fragments into cloning vectors (M13 or phagemid) was performed as follows: a 2:1 molar ratio of vector to insert was mixed at low concentration (≤30µg/ml) in the presence of 1mM ATP, 1x ligase buffer (50mM Tris-HCl pH 7.5, 10mM $MgCl_2$, 10mM DTT), 4mM spermidine and 0.1 Weiss units of T4 ligase per µl of reaction mix. Ligation reactions containing DNA molecules with 'sticky ends' were incubated at room temperature overnight, wheras blunt-ended ligations were left at 4° overnight.

2.2.6 Polynucleotide kinase treatment

Addition of 5' phosphate groups to $(dA\text{-}dC)_{12}$ and $(dG\text{-}dT)_{12}$ synthetic oligonucleotides prior to ligation (see section 2.2.18d) was done as follows: 2µg of oligonucleotides were incubated with 2 units of T4 kinase at 37° for 1 hour, in a 20µl reaction mixture containing 50mM Tris-HCl, 10mM $MgCl_2$ and 1mM ATP. Kinase was inactivated by heating the reaction to 68° for 10 minutes.

In order to visualize PCR-amplified microsatellite alleles on denaturing polyacrylamide gels (chapters 5 and 6), 10 pmoles of one of the amplimers were radioactively labelled using 3 pmoles of 3000Ci/mmol [γ^{32}P]ATP, following the protocol described above.

2.2.7 Agarose gel electrophoresis

Agarose gels in the concentration range 0.8 - 4% were used, depending on the size of DNA fragments to be resolved. 0.8 - 2% gels were cast from Sigma agarose or Seachem HGT (ICN Biomedicals). 4% gels were a 3:1 mixture of Nusieve agarose (ICN Biomedicals) and HGT agarose. Gels <2% in concentration were run in TAE (40mM Tris-acetate, 20mM sodium acetate, 0.2mM EDTA, pH8.3) containing 0.5µg/ml ethidium bromide. To obtain greater resolution, gels with a concentration ≥2% were run in TBE (0.089M Tris-borate, 0.089M boric acid, 0.002M EDTA) in the presence of 0.5µg/ml ethidium bromide.

Gel length varied from 10-40cm, depending on the separation required. Loading mix was added to DNA samples prior to loading; 5x loading mix comprised 12.5% ficoll 400, 0.1% bromophenol blue in 5x TAE. DNA samples were run alongside markers of known molecular weight, either λ DNA cut with *Hin*dIII or ϕx174 DNA cut with *Hae*III. Gels were run at 4-8 volts/cm until the desired degree of separation had been achieved. DNA was then visualized by UV flourescence on a transilluminator (Chomato-vue C-63, UV Products Inc., California), and photographed using Kodak negative film (T-max Professional 4052). Films were processed with Kodak LX24 developer, FX40 fixer and HX40 hardener.

2.2.8 Preparative gel electrophoresis

In order to purify DNA fragments prior to ligation, random oligonucleotide labelling etc., preparative gel electrophoresis (Yang *et al.*, 1979) was used. The DNA samples were run in an agarose gel until adequate separation had been achieved. Dialysis membrane (Scientific Industries International Inc., Loughborough) was cut into sheets just wider than the gel slots and longer than the thickness of the gel and boiled in 10mM Tris-HCl pH 7.5, 1mM EDTA for 10 minutes. The gel was viewed using a hand-held UV source (Chromato-vue 57, UV Products Inc.) and an incision made with a scalpel immediately ahead of the band or smear to be purified. The dialysis membrane was inserted into this cut, and the DNA run onto the membrane at 10-15 volts/cm. When all of the DNA was loaded onto the membrane (as judged by eye), the incision was

extended across the gel to loosen the membrane. With the current still on, the membrane was rapidly removed to a 1.5ml microfuge tube using forceps. The tube was then spun for 3 minutes at 15,000 rpm with the corner of the membrane trapped in the lid to gravitate the buffer containing the DNA to the bottom of the tube. The DNA was subsequently precipitated with ethanol and redissolved in the required volume of distilled water.

2.2.9 Southern blotting (Southern, 1975).

The positions of the DNA markers were visualized by UV fluorescence and marked by notching the sides of the gel to be blotted. The gel was shaken gently in 0.25M HCl for 2 x 7 minutes (depurination), 0.5M NaOH, 1M NaCl for 2 x 15 minutes (denaturation) and 0.5M Tris-HCl pH 7.5, 3M NaCl for 2 x 15 minutes (neutralization). The gel was placed on a wick of 3MM blotting paper (Whatman International Ltd., Basingstoke) which had been soaked in 20x SSC (1x SSC is 0.15M NaCl, 15mM trisodium citrate pH 7) and draped over a reservoir of 20x SSC. A nylon filter (Hybond-N, Amersham) which had been pre-soaked in 3x SSC was placed on top of the gel, followed by a sheet of 3MM blotting paper, also soaked in 3x SSC. A stack of 10 sheets of Quick-Draw blotting towels (Sigma) was placed on top of the 3MM paper, followed by a glass plate and a 0.5-1kg weight. Blots were left for 2-15 hours, with regular changing of towels during the first hour to ensure even transfer.

2.2.10 DNA hybridization

2.2.10a Random oligonucleotide labelling of DNA fragments (Feinberg & Vogelstein, 1984)

5-10ng of gel-purified DNA were labelled as follows. The DNA was boiled for 3 minutes and allowed to cool to room temperature, and the volume adjused with distilled water such that the final volume of the reaction mixture would be 30µl. 1.2µl of 10mg/ml BSA (enzyme grade, Pharmacia) and 6µl of oligo-labelling buffer (OLB) were added. OLB consisted of a mixture of 3 solutions: solution A: 1.25M Tris-HCl (pH 8.0), 125mM $MgCl_2$, 0.18% v/v 2-mercaptoethanol, 0.5mM dATP, 0.5mM dGTP, 0.5 mM dTTP; Solution B: 2M HEPES (pH 6.6) and

solution C: hexadeoxyribonucleotides (Pharmacia) suspended in 3mM Tris-HCl, 0.2mM EDTA (pH 7.0) at 90 OD units/µl. Solutions A B and C were mixed in the ratio 2:5:3 respectively to give OLB. Following addition of OLB, 2µl α-^{32}P-dCTP (1000Ci/mmol, Amersham) and 2.5 units of the klenow fragment of DNA polymerase I were added. The reaction mixture was incubated for 1 hour at 37° or overnight at room temperature. After labelling, 70µl of 'stop solution' (20mM NaCl, 20mM Tris-HCl pH 7.5, 2mM EDTA, 0.25% sodium dodecyl sulphate - SDS) was added to the reaction, followed by 100µg of high molecular weight carrier herring sperm DNA. The probe was then ethanol precipitated and washed in 80% ethanol to remove unincorporated α-^{32}P-dCTP, and redissolved in 500µl distilled water. A specific activity of $>10^9$ dpm/µg DNA was routinely obtained. Probes were boiled for 3 minutes immediately prior to use.

2.2.10b Hybridization

Hybridization reactions were in either Denhardt's solution (0.2% ficoll 400, 0.2% polyvinyl pyrrolidone - PVP, 0.2% BSA in 3x SSC; Denhardt, 1966) for colony screening following plate lifts onto nitrocellulose filters, or in phosphate/SDS solution (0.5M sodium phosphate, pH 7.2, 1mM EDTA, 7% SDS; Church & Gilbert, 1984) for nylon membranes following Southern transfer. All hybridizations were performed in sealed perspex chambers in a shaking water bath.

Hybridizations in Denhardt's solution (modified by Jeffreys *et al.*, 1980).

Filters were pre-hybridized for 30 minutes in complete filter hybridization mix (CFHM: 1x Denhardt's solution, 0.1% SDS). Filters were then transferred to a minimal volume of CFHM + 6% polyethylene glycol (PEG) and the boiled probe and hybridized overnight at 65° with gentle shaking. Following hybridization, filters were washed at low stringency as described in section 2.2.10c.

Phosphate/SDS hybridization

Filters were pre-hybridized in 0.5M sodium phosphate, pH 7.2, 1mM EDTA, 7% SDS for 10 minutes, and then hybridized in a fresh aliquot of

the same solution plus the boiled probe. Hybridizations were at 65° with gentle shaking overnight, except for MS1 internal mapping experiments (chapter 4), where hybridization time was reduced to 3 hours.

2.2.10c Post-hybridization washing

Following hybridization, filters were washed at low (1x SSC, 0.1% SDS), intermediate (0.5x SSC, 0.05% SDS) or high (0.1x SSC, 0.01% SDS) stringency, according to the requirements of the experiment. Several washes were performed at 65° until the level of radioactivity stabilized, as monitored by a hand-held Geiger counter (Series 9000 mini-monitor, Mini Instruments Ltd., Essex).

2.2.10d Autoradiography

Filters were placed in autoradiographic cassettes separated from a sheet of Fuji RX100 X-ray film by a sheet of aluminium foil. Exposures were at room temperature or -80° for 1 hour to 14 days depending on the strength of the signal. Exposures at -80° were in the presence of an intensifying screen. Densitometric scanning, when required, was done with an Ultrascan Laser Densitometer (LKB).

2.2.10e Stripping filters for rehybridization

Filters required for rehybridization with a different probe were first stripped in 0.4M NaOH for 5 minutes at 65°, and then neutralized in 0.1x SSC, 0.1% SDS, 0.2M Tris-HCl pH 7.5 for 5 minutes.

2.2.11 M13 and phagemid cloning

Following restriction endonuclease digestion, gel purified DNA fragments were ligated into the multiple cloning site of the chosen vector (M13mp18 or19, Yanisch-Peron *et al.*, 1985 or pBluescript II KS$^+$ or SK$^+$, Short *et al.*, 1988) which had been linearized with the appropriate restriction enzyme, CIP-treated and gel purified (see sections 2.2.3, 2.2.4 and 2.2.8). The M13 library described in chapter 5 was constructed

following ligation of 20ng of insert DNA into 200ng of M13mp18 DNA in a 100µl reaction.

2.2.12 Transformation of *E. coli* with recombinant vector DNA (Hanahan, 1983)

2.2.12a Generation of competent cells

A 5ml culture of *E. coli* NM522 (Gough *et al.*, 1983) was grown at 37° overnight in Luria broth (LUB: 1% w/v Difco bacto tryptone, 0.5% w/v Difco bacto yeast extract, 0.5% w/v NaCl). 0.5ml was diluted 1 in 100 with fresh LUB and grown to mid-log phase (A_{600} ~0.45). Cells were pelleted in a microfuge tube at 12,000 rpm for 30 seconds, the supernatant removed, and the cells resuspended in 0.5ml of MR (10mM MOPS pH 7.0, 10mM RbCl). The cells were pelleted as before, resuspended in 0.5ml MRC (100mM MOPS pH 6.5, 10mM RbCl, 50mM $CaCl_2$), and placed on ice for 30 minutes. Following this, the cells were again pelleted and finally resuspended in 0.15ml MRC. Cells were held on ice prior to use.

2.2.12b Transformation

5µl of ligated DNA and 3µl of dimethyl sulphoxide (DMSO) were added to the competent cells, and the cells left on ice for 1 hour. Cells were heat-shocked at 55° for 35 seconds, cooled on ice and then held at room temperature. 200µl of log-phase cells, 10µl of 25mg/ml BCIG (in formamide) and 10µl 25mg/ml IPTG were added. This mixture was transferred to a glass test-tube, 3ml of molten (50°) BTL (1% w/v BBL trypticase, 0.5% w/v NaCl, 0.25% w/v $MgSO_4.7H_2O$, 0.6% w/v Difco agar) added, and the mixture poured onto a base of Luria agar (LUA: 1.5% w/v Difco agar, 1% w/v Difco bacto tryptone, 0.5% w/v Difco bacto yeast extract, 0.5% w/v NaCl). Plates were incubated at 37° overnight.

2.2.13 M13 library screening

Plate lifts onto nitrocellulose filters (Amersham) were by the method of Benton & Davis (1977). A filter was placed on the surface of each plate

using sterile forceps. Holes were punched around the edge of the filter with a sterile needle and the pattern of holes reproduced on the underside of the plate with a red-hot needle for purposes of orientation. After 5 minutes the filter was carefully lifted off the plate and floated phage-side down in 1.5M NaCl, 0.1M NaOH for 1 minute. The filter was then submerged in 2x SSC, 0.2M Tris-HCl (pH 7.5) for 1 minute, and blotted dry on 3MM paper. Filters were then baked at 80° for 3-4 hours prior to hybridization. Library plates were frequently stored for several weeks at 4° following lifts.

2.2.14 Second-round screening

Positive plaques were picked from the library plates using a sterile pasteur pipette and deposited in 1ml of phage buffer (6mM tris-HCl pH 7.2, 10mM $MgSO_4$, 0.005% gelatin). 100µl aliquots of 10^4, 10^5 and 10^6-fold dilutions of this stock were mixed with 100µl of log-phase NM522 in LUB and incubated at room temperature for 15 minutes. 3ml of molten BTL was added, and the cells plated onto a base of LUA. Plates were incubated at 37° overnight. Plates of suitable plaque density were lifted onto nitrocellulose as above and hybridized. Single well separated positive plaques were picked into 1ml of phage buffer and stored at 4°.

2.2.15 Recovering single-stranded DNA from recombinant M13 phage

Plaques resulting from recombinant phage were picked into 1ml of phage buffer. 0.5ml of a culture of NM522 which had been grown at 37° overnight in LUB was diluted 1 in 100 into fresh LUB. 1.5ml of the diluted NM522 was mixed with 0.1ml of phage suspension in a large capped test-tube and shaken vigourously for 5-6 hours at 37°. The culture was transferred to a microfuge tube and spun for 10 minutes at 15,000 rpm. 1ml of supernatant was removed to a fresh tube, and the phage precipitated by the addition of 200µl of PEG 6000, 2.5M NaCl followed by incubation at 4° for 15 minutes. Phage particles were pelleted by centrifugation at 15,000 rpm for 10 minutes. Following removal of all traces of supernatant, the phage particles were resuspended in 100µl of 1.1M sodium acetate (pH 7.0). The protein component of the phage particles was removed by adding an equal volume of phenol/chloroform (phenol, chloroform, isoamyl alcohol and 8-hydroxyquinoline in the ratio

100: 100: 4: 1, saturated with Tris-HCl pH 7.5) and emulsifying by mixing vigo rously. The two layers were separated by centrifugation at 15,000rpm for 5 minutes and the upper aqueous layer removed. To maximise recovery the phenol layer was re-emulsified with a second aliquot of 1.1M sodium acetate which was pooled with the first aqueous phase after centrifugation. Following ethanol precipitation and a rinse with 1ml 80% ethanol, the DNA was redissolved in 25µl of distilled water, and stored at -20°.

2.2.16 Introduction of phagemids into *E. coli* by electroporation

Recombinant pBluescript phagemids (Short *et al.*, 1988) were introduced into XL1Blue (Bullock *et al.*, 1987) by the electroporation method of Dower *et al.* (1988). A 1 litre culture of XL1Blue in LUB plus 25µg/ml ampicillin and 12.5µg/ml tetracycline was grown to log phase. The cells were sequentially pelleted and resuspended in decreasing volumes of 10% glycerol as follows:

1 litre culture: 4000rpm for 15 minutes at 4°, resuspended in 1 litre of 10% glycerol.

Cells in 1 litre of 10% glycerol: 4000rpm for 15 minutes at 4°, resuspended in 500ml of 10% glycerol.

Cells in 500ml of 10% glycerol: 4000rpm for 15 minutes at 4°, resuspended in 20ml of 10% glycerol.

Cells in 20ml of 10% glycerol: 4000rpm for 15 minutes at 4°, resuspended in 2ml of 10% glycerol.

Ligated DNA was precipitated with ethanol to remove salts and redissolved in distilled water, and 40µl of cells plus 5µl ligated DNA were subjected to an exponential pulse of 1.5kV (with a capacitance of 25µF and a parallel resistance of 940Ω) in a cuvette with a 2mm electrode gap, in a Bio-Rad Gene Pulser. (The time constant was approximately 20 milliseconds). After electroporation, the cells were shaken in 1ml SOC (2% w/v Difco bacto tryptone, 0.5% w/v yeast extract, 10mM NaCl, 10mM $MgCl_2$, 10mM $MgSO_4$, 20mM glucose;

Hanahan *et al.*, 1983) at 37° for 30 minutes, pelleted by centrifugation at 15000rpm for 30 seconds, resuspended in 50µl SOC and spread onto LUA plates containing 25µg/ml ampicillin, 12.5µg/ml tetracycline, 25µg/ml BCIG and 25µg/ml IPTG. Plates were incubated at 37° overnight.

2.2.17 Recovering single-stranded DNA from recombinant phagemids

The XL1Blue colony containing the recombinant phagemid (pBluescript KS^+ or SK^+) was picked from the plate surface and grown in 5ml LUB plus 25µg/ml ampicillin and 12.5µg/ml tetracycline for approximately 6 hours. A further 3ml of LUB/tetracycline/ampicillin was inoculated with enough of these cells to give an OD of approximately 0.05 (typically a 10-fold dilution). Following dilution, the cells were grown for a further hour (to an OD of about 0.1) and VCSM13 helper phage (Stratagene) added at low multiplicity of infection (~1:10: 10µl of a $1x10^9$ pfu/ml solution). The culture was grown overnight at 37° with vigo rous shaking. The following morning, the culture was heated to 65° for 15 minutes and the cell debris pelleted by centrifugation at 15,000rpm in a microfuge for 10 minutes. The supernatant was transferred to a fresh tube and the phage precipitated and DNA prepared as for M13 (see above).

2.2.18 Polymerase Chain Reaction (Saiki *et al.*, 1988).

2.2.18a General PCR protocol

PCR amplifications from genomic DNA were generally performed in a 10µl reaction mix comprising 45mM Tris-HCl (pH 8.8), 11mM $(NH_4)_2SO_4$, 4.5mM $MgCl_2$, 6.7mM 2-mercaptoethanol, 4.5µM EDTA, 1mM dATP, 1mM dCTP, 1mM dGTP, 1mM dTTP, 113µg/ml BSA, 1µM of each primer and 1 unit of Taq polymerase. Reaction mixtures were overlaid with paraffin oil in a 0.5ml microfuge tube and amplifications carried out in either a Perkin-Elmer Cetus DNA thermal cycler (Perkin-Elmer Cetus, Connecticut, USA) or a Techne programmable Dri-block PHC1 (Techne, Cambridge). Amplimer sequences and parameters for the various loci amplified are given in table 2.2. Following amplification, PCR products were analysed by gel electrophoresis and ethidium staining,

Table 2.2. PCR parameters. *followed by two chases of 67° for 1 minute, 70° for 10 minutes. (Continued overleaf).

Locus	Amplimers	Input DNA(ng)	Denaturing		Annealing		Extension		No. Cycles	Thermal Cycler	Chapter
			°C	mins	°C	mins	°C	mins			
MS32	A: TCACCGGTGAATTCCACAGACACT B: AAGCTCTCCATTTCCAGTTTCTGG C: CTTCCTCGTTCTCCTCAGCCCTAG D: CGACTCGCAGATGGAGCAATGGCC D1: TCACCGGTGAATTCCGACTCGCAG-ATGGAGCAATGGCC	100	95	1.5	60	1.5	70	10	15	Techne	3
MS1	A: GCTTTTCTGTGATGAGCCTTGATG B: AAGAAGCATATGCAACCCATGAGG	100	95	1.5	60	1.5	70	10	15	Techne	3
MS1	A: GCTTTTCTGTGATGAGCCTTGATG B: AAGAAGCATATGCAACCCATGAGG A-R: TCATGCGTCCATGGTCCGGAT$^{A}/_{G}$-TCCACCCT$^{A}/_{G}$TCCACCCTA C-R: TCATGCGTCCATGGTCCGGAT$^{A}/_{G}$-TCCACCCT$^{A}/_{G}$TCCACCCTG TAG: TCATGCGTCCATGGTCCGGA	100	96	1.3	67	1.0	70	5	18*	Cetus	4
TG9	A: CTGCACAAATCCCCACTCCC B: GCAAGACCCTGCCTCTTTGT	10	95	1.5	60	1.5	70	2	33	Techne	5
TG10	A: CCACAGGTAGAGCCCCTTTG B: TAAAGGCACCAGATCCTTCG	10	95	1.5	60	1.5	70	2	33	Techne	5,6
TG29	A: TTTTGCCCTAGCCACACGCT B: GAAAGACTGAAAACCAGCCC	10	95	1.5	60	1.5	70	2	33	Techne	5

Locus	Amplimers	Input DNA(ng)	Denaturing		Annealing		Extension		No. Cycles	Thermal Cycler	Chapter
			°C	mins	°C	mins	°C	mins			
Actin[1]	A: TTGACCTGAATGCACTGTGA B: TTCCATACCTGGGAACGAGT	10	96	1.3	58	1.0	70	2	33	Cetus	6
Mfd3[2]	A: GGTCTGGAAGTACTGAGAAA B: GATTCACTGCTGTGGACCCA	10	96	1.3	58	1.0	70	2.0	33	Cetus	6
Mfd5[2]	A: CATAGCGAGACTCCATCTCC B: GGGAGAGGGCAAAGATCGAT	10	96	1.3	58	1.0	70	2.0	33	Cetus	6
Mfd49[3]	A: GATAAATGCCAAACATGTTG B: TGCTCTCAGGATTTCCTCCA	10	96	1.3	50	1.0	70	2.0	33	Cetus	6
Mfd64[4]	A: ACGAACATTCTACAAGTTAC B: TTTCAGAGAAACTGAGGTGT	10	96	1.3	50	1.0	70	2.0	33	Cetus	6
GGAA1	A: AGACTATGGTGTGGGAAGCC B: TCTCTGCATGTTTGTTGCAG	10	95	1.5	60	1.5	70	2.0	33	Techne	7
GGAA6	A: AGCCTGGGCAACTCTGACTC B: CCAACAACTTTGGTCCTCTC	10	95	1.5	60	1.5	70	2.0	33	Techne	7
GGAA7	A: GATTTTGGAGGGCTACATGG B: TCTTAACTGGCTGGATAGGC	10	95	1.5	60	1.5	70	2.0	33	Techne	7

or by Southern hybridization. All PCR reactions were performed in conjunction with zero DNA controls, which consistently gave no product.

2.2.18b Generation of DNA for sequencing by PCR

To generate further DNA for sequencing, MS32 and MS1 (chapter 3), PCR products were re-amplified by seeding a 50µl reaction mixture (as above) with 1µl from the initial reaction, and cycling at 95° for 1.5 minutes, 67° for 1.5 minutes and 70° for 3 minutes (Techne Dri-block). Cycling was continued until 1-2µg of product was generated (typically 15-20 cycles). Following gel purification, this DNA was sequenced directly using the protocol of Winship (section 2.2.19b) or used as a substrate for the generation of single-stranded DNA (see below). Alleles of simple repeat sequence GGAA.6 were simply amplified from genomic DNA as described in table 2.2, and sequenced directly following gel purification.

2.2.18c Generation of single-stranded DNA by PCR

200-500ng of gel-purified PCR product was used as a template for generating single-stranded DNA using a single primer in a further 50µl reaction performed as above. After ether extraction to remove paraffin oil, the single-stranded DNA was precipitated by adding 5M ammonium acetate to a final concentration of 2M, followed by an equal volume of isopropanol and chilled at -80° for 5 minutes. This allows precipitation of DNA in the absence of dNTPs (Maniatis *et al.*, 1982). Following centrifugation in a microfuge (15,000rpm for 10 minutes) and vacuum drying, the DNA pellet was redissolved in 7µl of distilled water.

2.2.18d Generation of a $d(C\text{-}dA)_n \cdot (dG\text{-}dT)_n$ probe by PCR

2µg aliquots of $(AC)_{12}$ and $(TG)_{12}$ oligonucleotides were independently kinased and ligated in 20µl reactions as described in sections 2.2.5 and 2.2.6. 1µl aliquots were then used as a substrate in a 10µl PCR reaction. Out of register annealing during PCR generated products several kilobases in length, which were gel-purified (section 2.2.8) and used to screen an M13 library (chapter 5) following random oligonucleotide priming (section 2.2.10a).

2.2.19 DNA sequencing

2.2.19a Sequencing reactions

Single-stranded M13 and phagemid DNAs, and single-stranded DNA generated by PCR were sequenced by the chain termination method of Sanger *et al.* (1977) using the sequenase protocol (USB) with T7 DNA polymerase (Tabor & Richardson, 1987). 1-2μg of DNA was mixed with 2ng the appropriate primer (M13 universal 17-mer, pBluescript SK or KS primer, or the opposite PCR primer to that used for generation of the template) in sequencing buffer (40mM Tris-HCl pH 7.5, 20mM $MgCl_2$, 50mM NaCl) in a total volume of 10μl. Template and primer were annealed by heating to 65° for 2 minutes in a water bath and allowing to cool to room temperature over a period of 0.5-1 hour. To the annealed template primer, the following were added: 1μl 0.1M DTT, 2μl labelling mix (1.5μM dCTP, 1.5μM dGTP, 1.5μM dTTP), 0.5μl [α–^{35}S]dATP (10μCi/μl) and 2 units of T7 DNA polymerase (Pharmacia). This mixture was incubated for 5-10 minutes at room temperature, and then divided equally into four 2.5μl aliquots of termination mixes:

T: 80μM dATP, 80μM dCTP, 80μM dGTP, 80μM dTTP, 8μM ddTTP, 50mM NaCl.
C: 80μM dATP, 80μM dCTP, 80μM dGTP, 80μM dTTP, 8μM ddCTP, 50mM NaCl.
G: 80μM dATP, 80μM dCTP, 80μM dGTP, 80μM dTTP, 8μM ddGTP, 50mM NaCl.
A: 80μM dATP, 80μM dCTP, 80μM dGTP, 80μM dTTP, 8μM ddATP, 50mM NaCl.

The reaction temperature was elevated to 37°, and incubation continued for a further 10 minutes. 4μl of stop solution (95% formamide, 20mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol) was added to each termination reaction, and the reactions stored at -20° prior to electrophoresis.

2.2.19b Sequencing double-stranded PCR products

Double-stranded gel purified PCR products were sequenced using the modified sequenase protocol described by Winship (1988). 0.2pmol of template DNA (~100ng for a 500bp template) was mixed with 10pmol of primer in 40mM Tris-HCl pH 7.5, 20mM $MgCl_2$, 50mM NaCl, 10% DMSO in a total volume of 6µl. Template and primer were annealed by boiling for 3 minutes and immediately cooling in a dry ice/IMS bath. After thawing at room temperature, 4µl of a solution comprising 0.5µl [α-^{35}S]dATP, 0.025M DTT and 2 units of T7 DNA polymerase was added to the template-primer, and the resulting mixture divided equally into 2µl each of 'T' 'C' 'G' and 'A' mix:

T mix: 50mM NaCl, 10% DMSO, 80µM dCTP, 80µM dGTP, 80µM dTTP, 8µM ddTTP.
C mix: 50mM NaCl, 10% DMSO, 80µM dCTP, 80µM dGTP, 80µM dTTP, 8µM ddCTP.
G mix: 50mM NaCl, 10% DMSO, 80µM dCTP, 80µM dGTP, 80µM dTTP, 8µM ddGTP.
A mix: 50mM NaCl, 10% DMSO, 80µM dCTP, 80µM dGTP, 80µM dTTP, 0.08µM ddATP.

These reactions were incubated at 37° for 5 minutes, after which 2µl of chase mix (50mM NaCl, 10% DMSO, 0.25mM dATP, 0.25mM dCTP, 0.25mM dGTP, 0.25mM dTTP) was added to each reaction. Incubation was continued for a further 5 minutes, and 4µl of stop mix added. Reactions were stored at -20° prior to electrophoresis.

2.2.19c Sequencing polyacrylamide gel electrophoresis

6% polyacrylamide gels containing 8M urea with a buffer gradient of 0.5-2.5x TBE were used. Gels were 0.4mm thick, poured between two glass plates ~400 x 330mm. Gels were warmed by running at 2000 volts for 15 minutes prior to loading. Samples were denatured by boiling for 3 minutes or alternatively by heating to 80° in an oven for 10 minutes prior to loading; 2.5µl of each sample was loaded per well. Gels were run at 1500-2000 volts initially, until the plates had warmed to 50-60°, after which the voltage was reduced to ~1200 volts. Gels were run until

the bromophenol blue dye had run off the bottom (~2.5 hours). To resolve sequence >200bp from the primer extension gels with a buffer gradient of 0.5-1.5x TBE were used. These were typically run for 6-8 hours. When the run was complete, the gels were fixed in 10% methanol, 10% acetic acid for 10 minutes and transferred to a sheet of 3MM blotting paper. Gels were dried under vacuum at 80° for 1-2 hours in a Drygel Sr SE1160 (Hoefer Scientfic Instruments, San Francisco).

2.2.20 Computing

DNA sequences were analysed using a VAX 8650 mainframe computer operating on VMS 5.3-1, using programs developed at the University of Wisconsin (Devereux *et al.*, 1984). The EMBL DNA sequence database was scanned using the FASTN program (Lipman & Pearson 1985). Linkage analysis was performed using the LINKAGE programs (Lathrop *et al.*, 1985) and the CEPH database version 2. Software for computer simulations was written by Professor Alec Jeffreys in VAX BASIC V3.4.

CHAPTER 3

EVOLUTIONARY TRANSIENCE OF HYPERVARIABLE MINISATELLITES IN MAN AND THE PRIMATES

3.1 Introduction

Multilocus 'DNA fingerprinting' probes detect multiple variable loci in a wide range of animal species (Jeffreys *et al.*, 1987a, 1987b; Jeffreys & Morton, 1987; Wetton *et al.*, 1987; Burke & Bruford, 1987). However, little information exists concerning the dynamics of evolution of individual hypervariable loci. This chapter describes the use of PCR to amplify the primate homologues of two extremely variable human minisatellites, MS32 and MS1, in order to explore their ancestral state.

3.2 Human minisatellite MS32

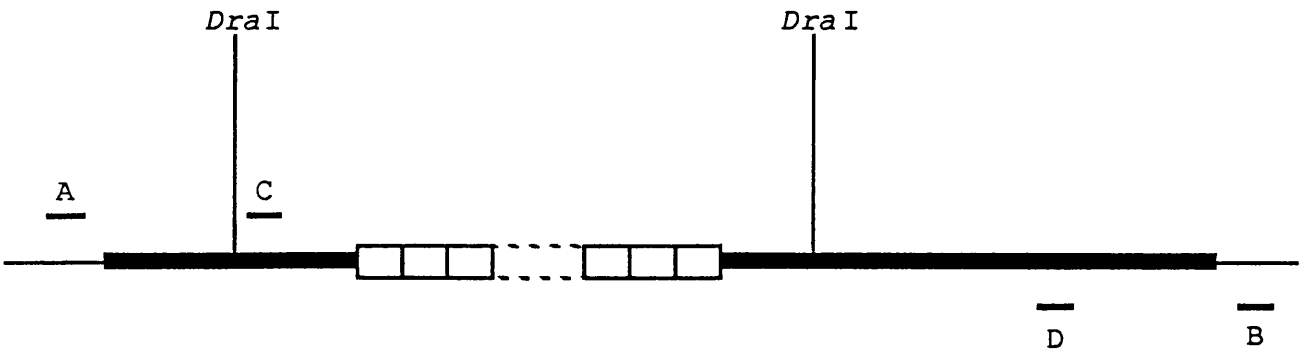
The highly variable human minisatellite MS32 is located interstitially on chromosome 1 at 1q42-43 (Royle *et al.*, 1988), and consists of a 29bp repeat sequence which has apparently expanded from within a retroviral long terminal repeat (LTR)-like element (Wong *et al.*, 1987, Armour *et al.*, 1989b). The allele length heterozygosity at this locus has been estimated at 97.5% (Armour *et al.*, 1989a), with alleles containing from 12 to more than 600 repeats.

3.3 Cross-hybridization of MS32 to primate DNA

To determine whether the MS32 locus is also detectable and variable in primates, a cloned *Dra*I fragment of human DNA consisting of the MS32 minisatellite plus 160 bp of flanking sequence (figure 3.1A) was used to probe *Alu*I-digested genomic DNA from 4 chimpanzees, 5 gorillas and 2 orang-utans. Two human samples were also probed for comparison. Under conditions of intermediate washing stringency (0.5x SSC, 1% SDS, 65°), several apparently monomorphic DNA fragments ranging from approximately 0.1 to 2kb were detected, some of which were species specific, whereas others were common to more than one species (figure 3.2Ai). Such monomorphic fragments were also detected in humans, in addition to the variable MS32 locus itself, and presumably reflect cross-hybridization of MS32 to other loci in primate DNA. (Because 160bp of DNA flanking the minisatellite were included in the MS32 probe, it is possible that some or all of the monomorphic loci detected may be due to cross-hybridization to other, related, LTRs). What appears to be a polymorphic locus is also detected in the gorilla,

Figure 3.1. Structure of the human MS32 and MS1 minisatellite loci and positions of PCR primers.
A. MS32. Minisatellite tandem repeats are shown as open boxes. Dashed lines indicate multiple tandem repeat units. The LTR-like element in which MS32 is embedded is shown as a thick line. The primer sequences are: A, 5'-TCACCGGTGAATTCCACAGACACT-3', B, 5'-AAGCTCT-CCATTTCCAGTTTCTGG-3', C, 5'-CTTCCTCGTTCTCCTCAGCCCT-AG-3' D, 5'-CGACTCGCAGATGGAGCAATGGCC-3'. Primer D1 is a derivative of primer D with the 5' extension TCACCGGTGAATTC, which contains an *Eco*RI site (underlined).
B. MS1. The primer sequences are: A, 5'-GCTTTTCTGTGATGA-GCCTTGAT-G-3', B, 5'A-AGAAGCATATGCAACCCATGAGG-3'.

A. MS32



B. MS1

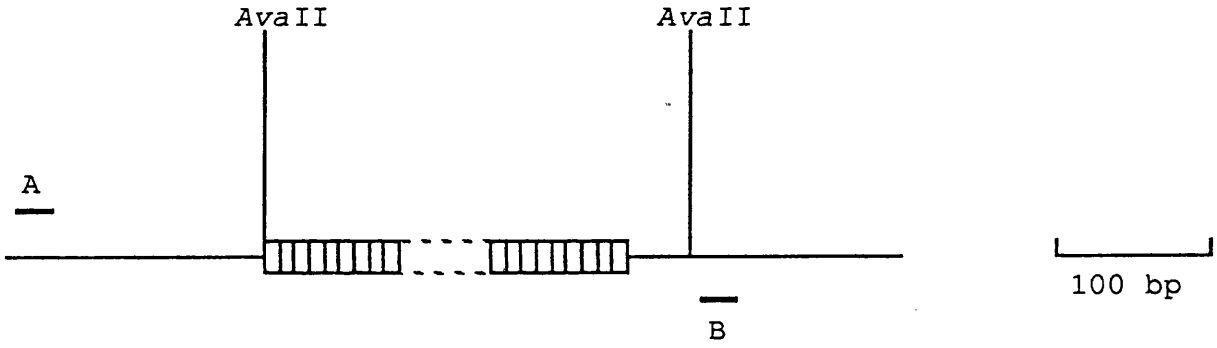
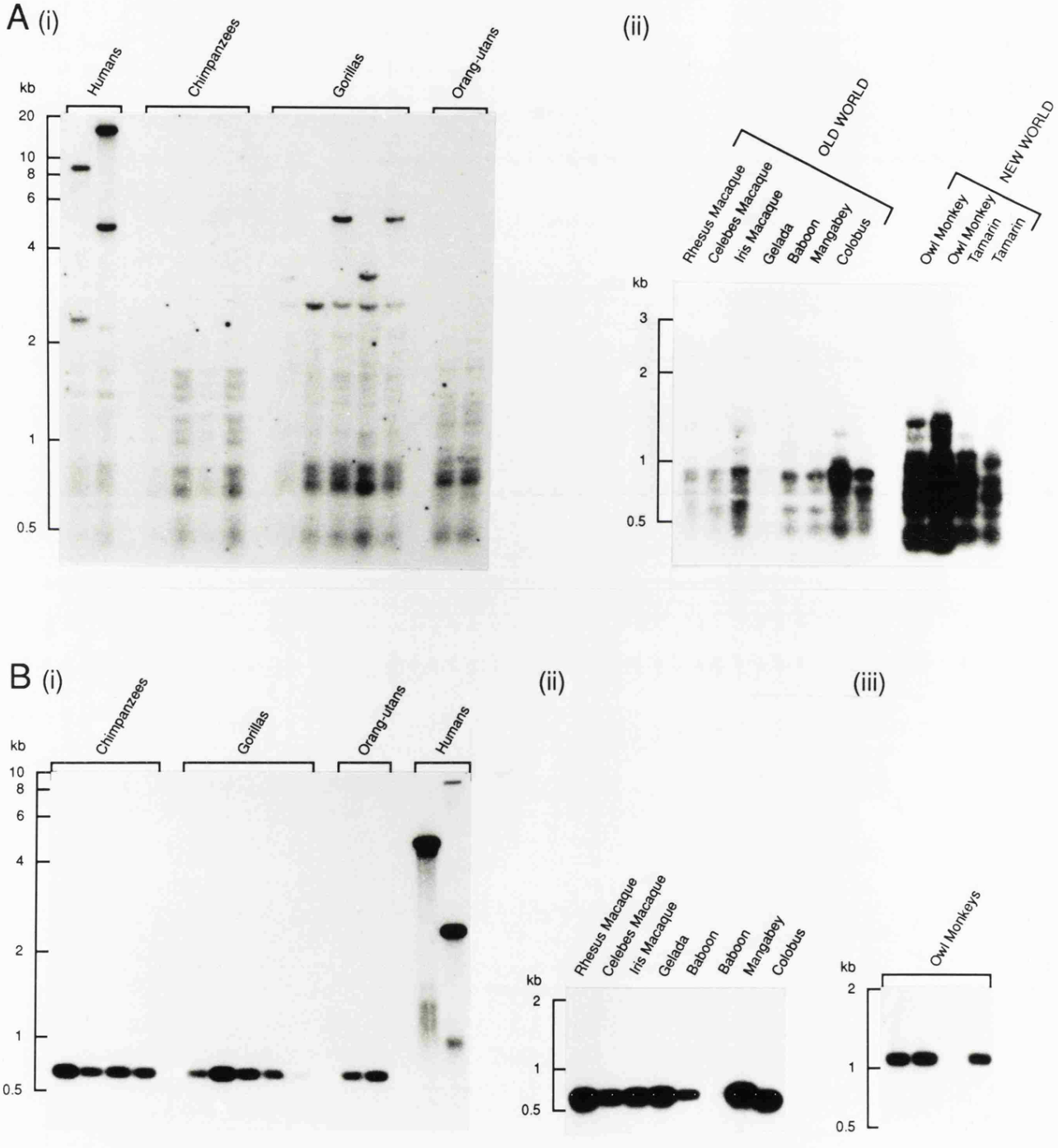


Figure 3.2. Minisatellite MS32 in primates.
A. Cross-hybridization of human MS32 to ape and monkey DNA. Genomic DNA was digested with *Alu*I, separated by agarose gel electrophoresis and Southern blot hybridized with the human MS32 probe (figure 1a).
(i). At intermediate stringency, human MS32 detects several monomorphic DNA fragments in AluI-digested DNA from the great apes. In addition, a variable locus is detected in gorilla DNA. No variable loci are apparent in either chimpanzee or orang-utan DNA. Two human DNAs are shown for comparison.
(ii). A similar profile is obtained with Old World and New World monkey DNAs.
B. Detection of the primate MS32 locus by PCR amplification followed by hybridization with the human MS32 probe.
(i). PCR amplification shows that MS32 is a monomorphic locus of approximately 600bp in the great apes. 100ng of genomic DNA was amplified for 15 cycles using primers A and B (figure 1A). Two human DNAs were amplified for comparison (note that the first human has an allele which is too large for PCR amplification (>10kb), hence only the smaller allele appears).
(ii). Similarly, PCR amplified MS32 in a range of Old World monkeys is a monomorphic locus of about 600bp. Amplifications were performed as above, but with primers C and B, rather than A and B. There is a slight size difference between the Cercopithecines (rhesus -> mangabey) monkeys and the colobus monkey. Amplification of MS32 in the baboon is poor, probably due to primer mismatches.
(iii). The PCR product generated from the South American owl monkey represents an apparently monomorphic locus of 1.1kb. Amplifications were performed as above, but with primers A and D, rather than A and B. The locus has amplified poorly in the third individual, for unknown reasons.


A (i)
Humans
Chimpanzees
Gorillas
Orang-utans
kb
20
10
8
6
4
2
1
0.5
(ii)
OLD WORLD
NEW WORLD
Rhesus Macaque
Celebes Macaque
Iris Macaque
Gelada
Baboon
Mangabey
Colobus
Owl Monkey
Owl Monkey
Tamarin
Tamarin
kb
3
2
1
0.5
B (i)
Chimpanzees
Gorillas
Orang-utans
Humans
kb
10
8
6
4
2
1
0.5
(ii)
Rhesus Macaque
Celebes Macaque
Iris Macaque
Gelada
Baboon
Baboon
Mangabey
Colobus
kb
2
1
0.5
(iii)
Owl Monkeys
kb
2
1
0.5

with allele sizes of about 2.5, 3 and 5kb. No variable bands were detected in either the chimpanzee or the orang-utan, suggesting that MS32 is either monomorphic in these species, or has polymorphic alleles of less than 2kb which are obscured by the ladder of monomorphic DNA fragments.

Under the same conditions, MS32 generates similar hybridization patterns in a selection of both Old World and New World monkeys, with several monomorphic DNA fragments ranging from approximately 0.1 to 2kb and no obvious detection of variable loci (figure 3.2Aii). No cross-hybridization to prosimian (lemur or dwarf lemur) DNA was observed at this hybridization stringency (not shown).

3.4 Amplification of the primate MS32 locus by PCR

In order to establish which if any of the DNA fragments detected in great ape DNA by the human MS32 clone represented the MS32 locus in these species, the locus was amplified by PCR. Several primers have been synthesized for PCR amplification of human MS32 (figure 3.1A). To amplify this locus in the great apes, the most distal human primers A and B were used. Both lie outside the putative LTR. The amplified alleles were detected by Southern blot hybridization using the human *Dra* I MS32 probe which lies entirely within the amplification primers. PCR revealed that MS32 in the chimpanzee, gorilla and orang-utan is an apparently monomorphic locus with a length of about 600bp (between primers A and B) in all three species (figure 3.2Bi). The variable DNA fragments detected by MS32 in the Southern blot hybridization of gorilla genomic DNA were not amplified and presumably represent a second, cross-hybridizing minisatellite similar in sequence to MS32.

MS32 failed to amplify in both the Old World and New World monkeys using primers A and B (not shown). However, by replacing primer A with primer C (see figure 3.1A) successful amplification of MS32 in the Old World monkeys was achieved (figure 3.2Bii). Again the locus was found to be monomorphic, though slightly larger than the great ape MS32 allele in the Cercopithecine monkeys (gelada, mangabey, baboon and macaque). The black and white colobus monkey, a representative of the Colobinae, was found to have the same size allele as the great apes.

Although successful in amplifying MS32 in the Old World monkeys, the primer combination of C plus B failed to amplify the locus in New World monkeys (not shown). However, primer A coupled with primer D resulted in successful amplification of a 1.1kb apparently monomorphic locus in the South American owl monkey (figure 3.2Biii). Incidentally, this primer combination would not allow the amplification of MS32 in the Old World monkeys. Subsequent DNA sequence analysis revealed that primer D has critical mismatches with the Old World monkey sequences at its 3' end, whereas the primer C sequence is heavily diverged between human and owl monkey. Attempts to amplify MS32 from prosimian DNA (lemur and dwarf lemur) were completely unsuccessful with all combinations of primers (not shown).

3.5 DNA sequence organization of MS32 in primates

MS32 alleles were further amplified from primates to the stage where PCR products could easily be visualized on an ethidium-stained gel (figure 3.3). The sequence of MS32 in the chimpanzee, gorilla, orang-utan, rhesus macaque and colobus monkey was determined by direct sequencing of these PCR products (figure 3.4). The larger PCR product from the owl monkey was sequenced by re-amplifying with primers A and D1, cleaving with *Eco*RI and cloning into the *Eco*RI site of M13mp19 (Yanis-Perron *et al.*, 1985); primer A has a natural *Eco*RI site while primer D1 is a longer version of primer D with an *Eco*RI site engineered into an additional 5' extension (figure 3.1A legend).

In man, MS32 consists of a 29bp tandem repeat, with almost all repeat units virtually identical, other than a T->G transversion which occurs in approximately two thirds of the repeat units (Wong *et al.*, 1987, Jeffreys *et al.* 1990). The penultimate repeat unit has fairly extensive base substitutions at its 3' end, and is 4bp longer than a 'standard' repeat, due to an additional CGGG motif near its 5' end. The final repeat unit is heavily diverged, but is homologous to the first 20bp of a standard repeat, showing a 13/20 match, with 6 of the 7 mismatches being transitional base substitutions (figure 3.4).

The great apes have just two complete repeat units, the first of which is identical to the repeat unit which makes up the majority of the human

Figure 3.3. PCR-amplified primate MS32 alleles. PCR products shown in figure 3.2B were re-amplified for 15 cycles and electrophoresed through a 1% agarose gel in TAE with ethidium bromide. The size discrepancy between the rhesus monkey and the other Old World primates is clearly visible.

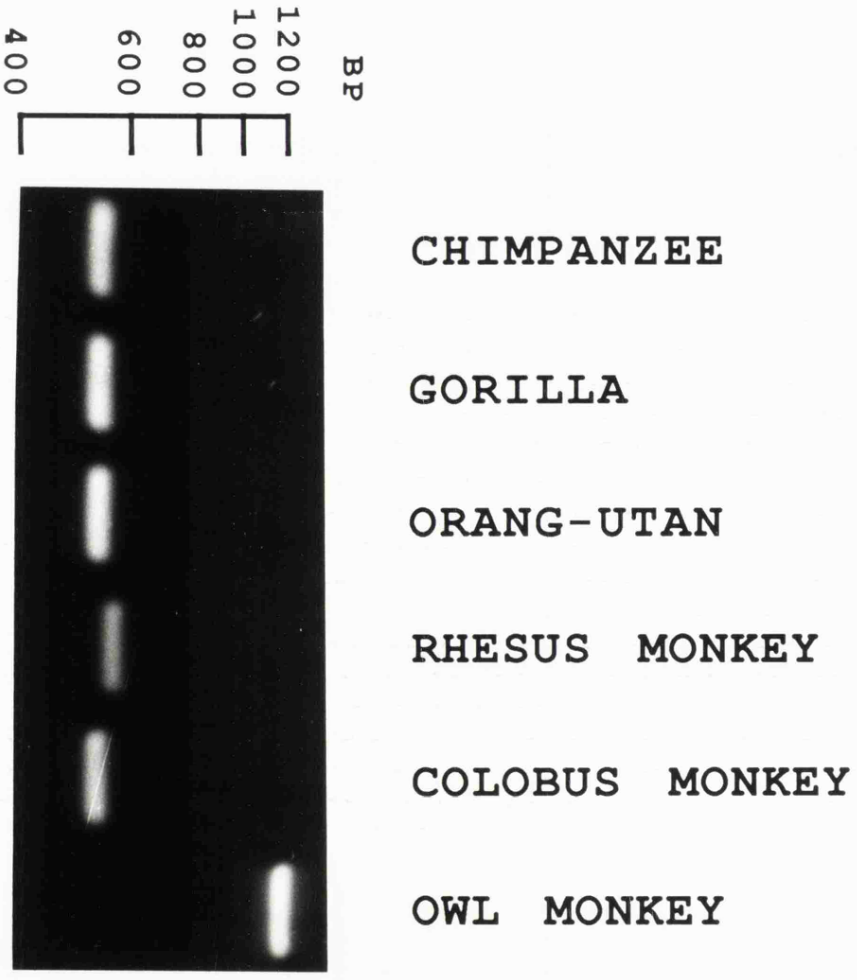

400
600
800
1000
1200
BP
CHIMPANZEE
GORILLA
ORANG-UTAN
RHESUS MONKEY
COLOBUS MONKEY
OWL MONKEY

Figure 3.4. DNA sequence organization of the MS32 locus in primates. Sequences are aligned, with gaps (.) introduced to improve alignments. Base changes from the consensus are shown in bold type. The minisatellite is shown in uppercase with an arrow over each tandem repeat. A duplicated region downstream of the minisatellite in the rhesus monkey (see figure 3.5) is underlined. Uncertainties in the colobus sequence are denoted by 'n'. The extra sequence upstream of the minisatellite in the owl monkey is shown, and PCR primer A is underlined in the human sequence.


owl monkey ctgcctgaaaatgcctaaaactagtc
owl monkey aactcatctctcctactcctcctaacgagtgctcacctttccaatcggtaaccttgctgcagttatcaacctcacgcattcatttctgatgagaaagtcattatataagttccaggagaa
owl monkey aatatttccacagtgttaaaatgttttattctcttatagttaacagcagtgaagagatagaaagagggtgaaagagcttatttagaggcataaaccaaaaatcaatgacagcccaagtat
owl monkey attggttaggaaatgggtaaaaaaaaatctgctttatatttatagagcaggcatcattgttttccaggagaaaagtcaatgttgtcttaagtttgaagttttgttccaggactaagaga
owl monkey gccattcactccatttccaacaatgactttccactcagatattcaagagtgatgggtgcttcatacaccttctatcttctttacttcagactaggccatatcacatttgatctgtgagaa
human gatcaccggtgaattccacagacacttaaaagcaaaaataataat
owl monkey tgaagtggtgaagggaagagatgagaaagaggaaaaatggtcaataaaggacgagtggttttatatataaaatcaagtacgttcctaaggattataataacagttttttcttgttacaa
|----------> Start of LTR
human ttgttgaatacagtgagttctaaatttctcttcaaagaatcagtatgtcagtatgttca.....gttctttgttctccattttaaagttgaacttcctcgttctcctcagccctagt
chimp ttgttgaatacagtgagttctaaatttctcttcaaagaatcagtatgtcagtgtgttca.....gttctttgttctccattttaaagttgaacttcctcgttctcctcagccctagc
gorilla ttgttgaatacagtgagttctaaatttctcttcaaagaatcagtatgtcagtatgttca.....gttctttgttctccattttaaagttgaacttcctcgttctcctcagccctagt
orang-utan ttgttgaatacagtgagttctaaatttctcttcaaagaatcagtatgtcagtatgttca.....gttctttgttctccattttaaagttgaacttcctcgttctcctcagccctagt
owl monkey gtgttaaatacagtgag...tcagtttctcttcaaagaatcagtatgtcagtatgttcaatttcctgttctttgttctccattttaaagtttaacttccttatctccctgccccctggt
|----------->|----------> |----
human ttcactaaacaacctttccaacagttctaatcagtagttaacgtctatt.....CCCCTGGTCACCTGCTCCATTCTGAGTCACCCCTGGTCACCTGCTCCATTCTGCGTCA(CCCCT
chimp ttcactaaacaacctttccaactgttctaatcagtagttaacatctatt.....CCCCTGGTCACCTGCTCCATTCTGAGTCA.....
gorilla ttcactaaacaacctttccaacagttctaatcagtagttaacatctatt.....CCCCTGGTCACCTGCTCCATTCTGAGTCA.....
orang-utan ttcactaaacaacctttccaacagttctaatcagtagttaacatctattcctattCCCCCTGGTCACCTGCTCCATTCTGAGTCA.....
rhesus ctccaacagctccaatgagtagttcacatctatt.....CCCCTGGTCACCTGCTCCATCCTGAGTCA.....
colobus ctccaacagctccaatgagtagttcacatctgtt.....CCCCTGGTCACCTGTTCCATCCTGAGTCA.....
owl monkey ttcagtaaacaaccttc.ccaccagctctaatcagtagttcacatctatg.....TCCCTGGTCACCTGCTCCATCCTGAGTCACCCCTGGTCACCTGCTCTGTCCTGAGACA.....


human
chimp
gorilla
orang-utan
rhesus
colobus
owl monkey
End of LTR

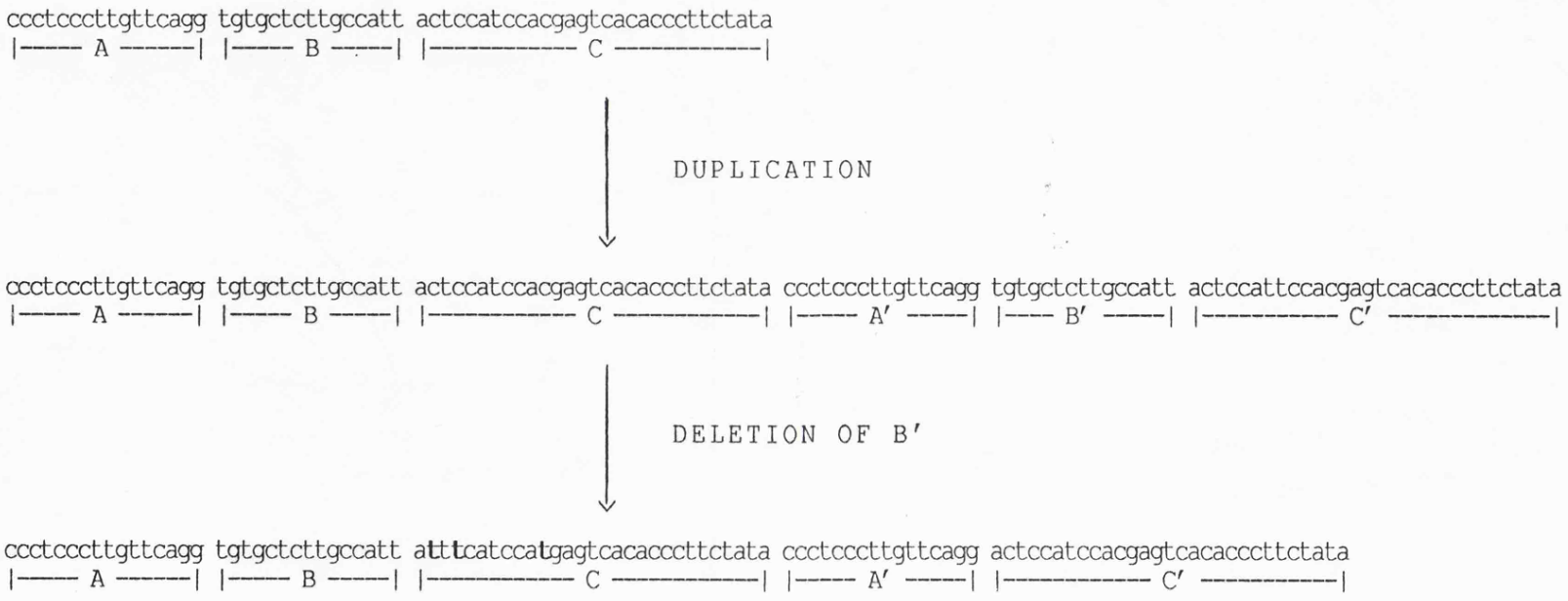
minisatellite. The second repeat is similar to the variant penultimate repeat in the human array, but lacks the additional CGGG sequence and thus maintains the 29bp periodicity which is characteristic of the rest of the human minisatellite. The heavily diverged, 20bp final repeat unit which terminates the human minisatellite also completes MS32 in the apes. The orang-utan has a 6bp duplication immediately preceding the minisatellite. This duplication is not present in man, the chimpanzee or the gorilla. Other than this small duplication in the orang-utan, the overall arrangement of the DNA flanking the minisatellite in the great apes is identical to that in man.

The MS32 minisatellite has the same arrangement in the rhesus macaque and the black and white colobus as in the great apes. However, as noted above, the rhesus monkey, and all other Old World monkeys typed bar the colobus, give a slightly larger PCR product than the great apes. This is due to 45bp of extra sequence downstream of the minisatellite which appears to have arisen through a duplication in the Cercopithecines of a 60bp stretch of DNA, followed by a deletion of 15bp from the centre of the duplicated segment, resulting in a net gain of 45bp (figure 3.5).

The sequence of MS32 in the South American owl monkey reveals that the minisatellite has three complete repeat units, rather than the two seen in the great apes and Old World monkeys. It has the same gross structural characteristics as a human MS32 allele: two similar repeat units, a variant penultimate repeat unit, and a heavily diverged 20bp final repeat unit. Unlike the human minisatellite, the standard repeat units are not identical but have accumulated a considerable number of base substitutions (figure 3.4).

The main reason for the greater length of the PCR product from the owl monkey when compared to the Old World monkeys and apes is the presence of approximately 600bp of additional sequence between the priming site of primer A and the beginning of the LTR. Homology between the owl monkey and human sequences ends abruptly at the 5' boundary of the LTR, and is not regained as primer A is approached, suggesting that primer A was fortuitously annealing to a site 600bp upstream of the LTR, rather than at the intended priming site (20bp upstream of the LTR in humans).

Figure 3.5. A duplication/deletion event downstream of MS32 in the rhesus monkey. The extra 45bp of sequence downstream of MS32 appears to have arisen as shown; duplication of the sequence ABC gives the tandem repeat ABCA'B'C', and subsequent deletion of B', followed by 3 base substitutions in C, gives the observed ABCA'C'.


ccctcccttgttcagg tgtgctcttgccatt actccatccacgagtcacacccttctata
|----- A ------| |----- B -----| |---------- C ----------|
DUPLICATION
ccctcccttgttcagg tgtgctcttgccatt actccatccacgagtcacacccttctata ccctcccttgttcagg tgtgctcttgccatt actccattccacgagtcacacccttctata
|----- A ------| |----- B -----| |---------- C ----------| |----- A′ -----| |---- B′ -----| |---------- C′ ----------|
DELETION OF B′
ccctcccttgttcagg tgtgctcttgccatt atttcatccatgagtcacacccttctata ccctcccttgttcagg actccatccacgagtcacacccttctata
|----- A ------| |----- B -----| |---------- C ----------| |----- A′ -----| |---------- C′ ----------|

3.6 Elevated sequence divergence in the DNA regions flanking MS32

Inter-species DNA sequence divergence estimates for the DNA region flanking the MS32 tandem repeat (figure 3.4) are given in table 3.1. Comparisons of these divergences with corresponding estimates from thermal stability measurements (Sibley & Ahlquist, 1987) and from non-coding sequences spanning the η-globin pseudogene locus (Goodman *et al.*, 1990) show an elevated level of MS32 divergence for all inter-species comparisons.This elevation is highly significant for comparisons between Hominoidea, Cercopithecoidea and Ceboidea, and suggests an elevated level of base substitutions around the MS32 minisatellite locus compared with that seen in the unique sequence component of the human genome (from thermal stability measurements) or the η-globin locus (from DNA sequence analysis).

3.7 The human minisatellite MS1

The human MS1 minisatellite is located on chromosome 1 at 1p33-35 (Royle *et al.*, 1988) and consists of a 9bp tandem repeat with the number of repeats ranging from approximately 140 to 2500, and an allele length heterozygosity estimated to be >99% (Wong *et al.*, 1987; Jeffreys *et al.*, 1988). MS1 is the most variable and unstable human minisatellite isolated to date, with a spontaneous germline mutation rate to new length alleles of approximately 0.05 per gamete (Jeffreys *et al.*, 1988).

3.8 Cross-hybridization of MS1 to primate DNA

A cloned *Ava*II fragment of human DNA consisting of the MS1 minisatellite plus 45bp of 3' flanking sequence (figure 3.1B) was used to probe *Alu*I digested DNA from a range of primates (figure 3.6A). Under conditions of intermediate washing stringency (0.5x SSC, 1% SDS, 65°) a complex pattern of hybridizing DNA fragments was detected, rendering the identification of alleles derived from the MS1 locus difficult in all species other than man.

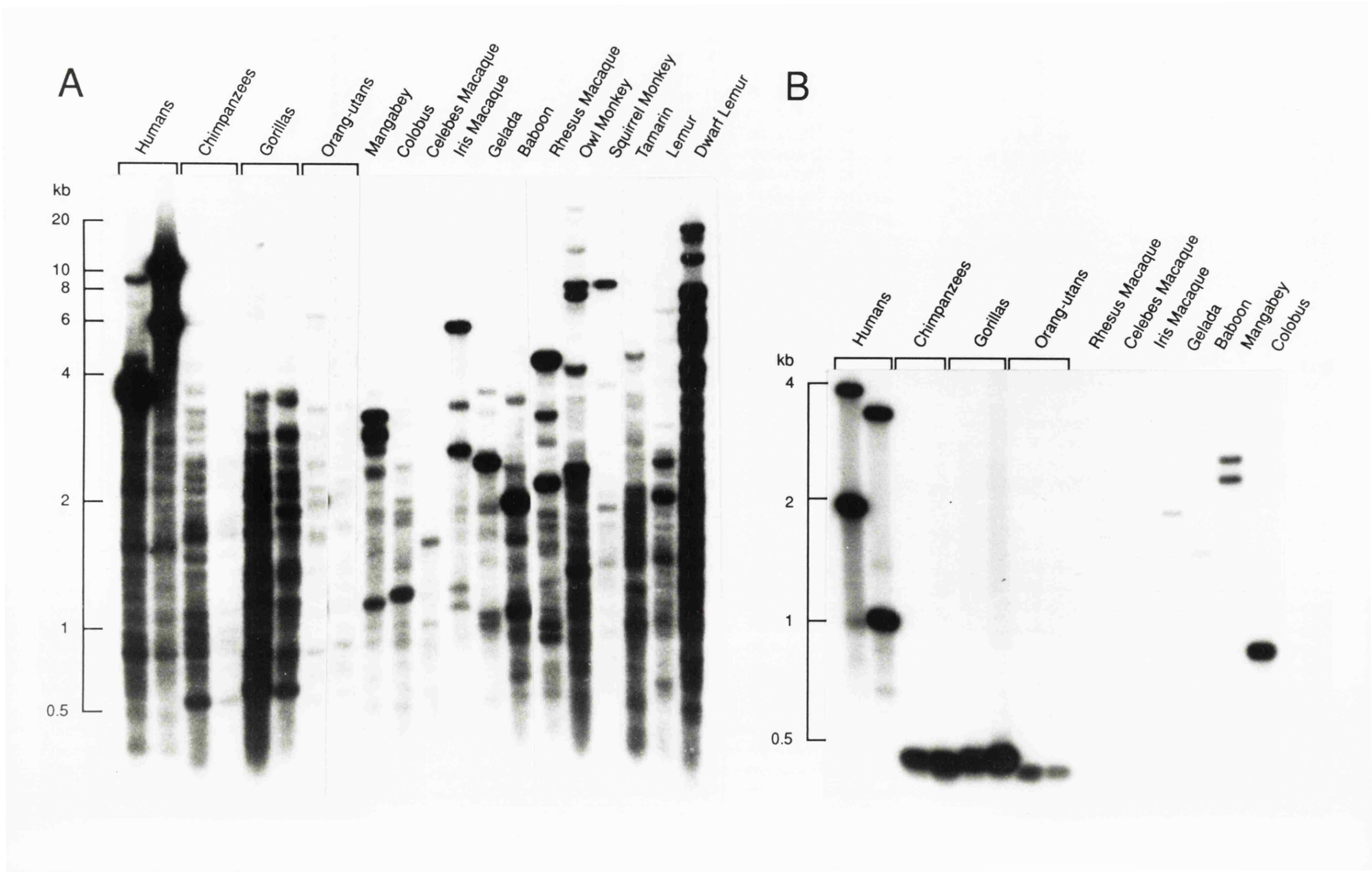
Table 3.1. Elevated sequence divergence in the regions flanking minisatellite MS32. For inter-species comparisons, the following abbreviations were used: H, human; C, chimpanzee; G, gorilla; O, orang-utan; Homi, Hominoidea; Cerc, Cercopithecoidea (rhesus, colobus); Cebo, Ceboidea (owl monkey). Inter-species MS32 distances were estimated by pairwise comparisons of all available DNA sequences flanking minisatellite MS32 (figure 3.4), and are calculated as the total number of substitutions plus gaps divided by the total number of shared nucleotide positions plus the total number of gaps, without correction for superimposed substitutions. ΔT50H indicates median sequence divergence estimated from thermal stability measurements (Sibley & Ahlquist, 1987). Uncorrected divergences for a 10.8kb non-coding segment spanning the ψη-globin pseudogene locus are taken from Goodman *et al.* (1990). The ratio of the MS32 to ψη-globin divergence is given together with the significance of the deviation between MS32 and ψη-globin divergences, estimated from the χ^2 test.

Comparison	MS32	ΔT50H	ψη-globin	Ratio	*p*
H-C,G	0.026	0.020	0.017	1.5	NS
H,C,G-O	0.037	0.036	0.035	1.1	NS
Homi-Cerc	0.119	0.073	0.073	1.6	<0.01
Homi-Cebo	0.162	0.112	0.108	1.5	<0.01
Cerc-Cebo	0.201	-	0.119	1.7	<0.001

Figure 3.6. Minisatellite MS1 in primates.

A. Cross-hybridization of human minisatellite MS1 to primate DNA. Genomic DNAs were digested with AluI, separated by gel electrophoresis and Southern blot hybridized with the human MS1 probe (figure 1A). At intermediate stringency, the human MS1 probe detects several fragments in *Alu*I digested primate DNAs.

B. Detection of the primate MS1 locus by PCR amplification followed by hybridization with the human MS1 probe. On amplifying MS1 in primates by PCR, it becomes clear that the locus is short and shows minimal variability in the great apes, but is apparently highly variable between some Old World monkey species, and polymorphic within at least one species (the mangabey). Experimental procedures were as for MS32.


A
Humans
Chimpanzees
Gorillas
Orang-utans
Mangabey
Colobus
Celebes Macaque
Iris Macaque
Gelada
Baboon
Rhesus Macaque
Owl Monkey
Squirrel Monkey
Tamarin
Lemur
Dwarf Lemur
kb
20
10
8
6
4
2
1
0.5
B
Humans
Chimpanzees
Gorillas
Orang-utans
Rhesus Macaque
Celebes Macaque
Iris Macaque
Gelada
Baboon
Mangabey
Colobus
kb
4
2
1
0.5

3.9 Amplification of the primate MS1 locus by PCR

To identify MS1 alleles in genomic DNA, this locus was amplified by PCR, again using primers based on the human flanking DNA sequence (figure 1B), and detected by Southern hybridization using the human MS1 probe described above (figure 3.6B). Like MS32, MS1 appeared to be a short, monomorphic locus in the chimpanzee, gorilla and orang-utan, although there was a small degree of size variation between species. Subsequent analysis of a larger panel of apes revealed that very small size differences can occur between different gorilla alleles, suggesting that MS1 is not completely monomorphic in this species, but can show a limited degree of length variability (figure 3.7).

In contrast to the great apes, the Old World monkeys had MS1 alleles which were variable between species, and also appeared to show polymorphism within some species. This is particularly evident in the mangabey; two presumptive alleles of different length can be seen. No amplification products could be obtained from New World monkey or prosimian DNAs (not shown).

3.10 DNA sequence organization of MS1 in primates

As for MS32, the structure of MS1 in the chimpanzee, gorilla, orang-utan and colobus monkey was determined by sequencing of PCR products. In addition, eight of the smallest human MS1 alleles detected in the CEPH panel of large families were amplified (figure 3.8), cloned and the 5' and 3' regions of the minisatellite sequenced. The overall arrangement of the minisatellite and immediate flanking sequence was identical in all species analysed, with each species having multiple copies of the consensus CCCTATCCA minisatellite repeat sequence with variant repeat monomers arising due to base substitutions. By assigning a letter code to each variant repeat, the repeat unit array of each minisatellite allele could be encoded as an alphabetical string (figure 3.9).

Comparison of the encoded allele structures of the eight human alleles shows a fairly stable 5' cluster of seven repeat units (ABCDEAC), with minor variation between alleles, and a highly stable 3' end (AACKMKKJBBL). Internally, the alleles are comprised primarily of

Figure 3.7. PCR-amplified ape MS1 alleles. PCR products shown in figure 3.6B were re-amplified for 15 cycles and electrophoresed through a 4% agarose gel (3 parts nusieve, 1 part HGT) in TBE with ethidium bromide. Two alleles are distinguishable in the gorilla, the first two individuals clearly being heterozygotes.


300
400
500
600
BP
CHIMPANZEES
GORILLAS
ORANG-UTANS

Figure 3.8. Small human MS1 alleles, PCR amplified. 8 of the smallest MS1 alleles, originally identified in the CEPH panel of large families by Southern hybridization, were PCR amplified for 30 cycles (from 100ng of genomic DNA) and the products electrophoresed through a 1% agarose gel in TAE with ethidium bromide.

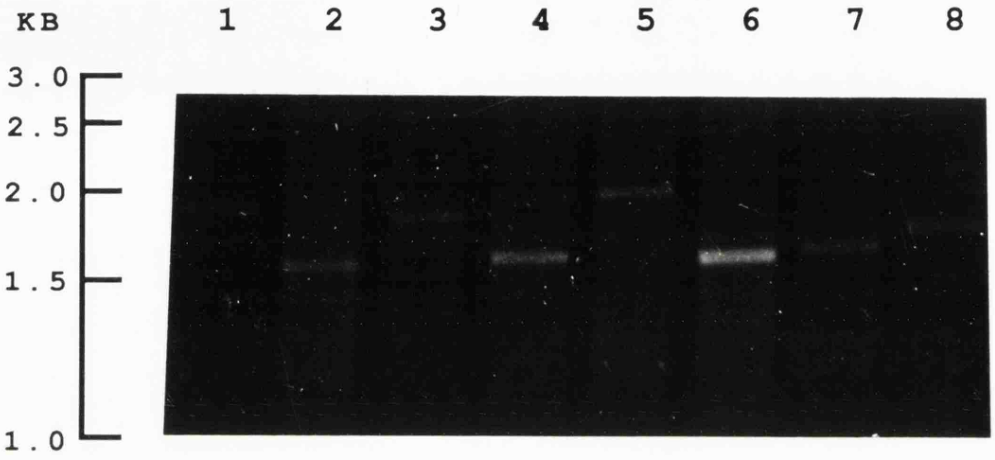

KB
1
2
3
4
5
6
7
8
3.0
2.5
2.0
1.5
1.0

Figure 3.9. MS1 variant repeats and allele structures.
A. 19 variations on the MS1 9bp repeat unit sequence are listed, with base changes from the consensus (con) shown in lower case. Each variant is assigned a letter code.
B. MS1 alleles from man, chimpanzee, gorilla, orang-utan and colobus monkey, written according to the letter code assigned to variant repeats. The origin of each of the short human alleles analysed is given: F = French, M = Mormon. Stable 5' and 3' haplotypes are shown in bold uppercase, with these and other related sequences joined by vertical lines. Gaps (.) have been introduced in the chimpanzee, gorilla and orang-utan alleles to improve alignments. A higher-order tandem repeat in the colobus allele is marked by arrows. $(A/C)_n$ denotes multiple A and C type repeats presumed to constitute the bulk of the human minisatellite; $(X)_n$ denotes multiple repeat units of unknown sequence at the centre of the colobus minisatellite. The number of repeats in each allele is given; for the human and colobus alleles, the number of repeats was estimated from the size of the PCR product.


A.
B.
No. rpts
con: CCCTATCCA
A
B c
C g
D t c
E t a
F c
G gc
H tg a
I t
J t
K a
L t
M t
N t cc
O c t
P t ca
Q a aatg
R c
S ca
HUMAN 1 (F) ABCDEACaacacaacaa<---(a/c)n-->ccccccccccaaaccccccccacaaaaaaccaaAACKMKKJBBL 173
2 (F) ABCDEACabbbaaca<--------(a/c)n-------->cccaaaaaaccccccccaaccccccaAACKMKKJBBL 131
3 (M) AcCDEACaccaaacaac<--------(a/c)n-------->aaaaaaaaaccccccccacccaaaAACKMKKJBBL 163
4 (M) ABCDEACabbbaac<----------(a/c)n---------->aaaaaaaaaccccaccccccccAACKMKKJBBL 139
5 (M) AcCDEACbbbbbbccccccccaccccaaaaac<-(a/c)n->ccccccccccccccaccccaaaAACKMKKJBBL 186
6 (M) ABCDEbCcbcaaacaca<---------(a/c)n---------->aaaaaaakaaaaaaaakakmAACKMKfJBBL 139
7 (M) AcCbEACbbbbcccc<--------(a/c)n-------->ccccccccccaaccccccacccaaaAACKMKKJBBL 145
8 (M) ABCDEACaacaa<---------(a/c)n--------->ccccccaccccccaaccaaaaaccccAACKMKKJBBL 159
CHIMPANZEE ABbDEfba.....AACK.aKJBBL 18
GORILLA ABCDEAbbbabb.....ACK.aKJBBL 21
ORANG-UTAN ABbDbgbbc.....hiJkaL 15
COLOBUS nfaoaaaofkafboafbofe<----------(X)n---------->aopafqlraskfaao 84
--->--->">

intermingled clusters of A- and C-type repeat units, which differ by a single base transition, with extreme variation between alleles in the order of A- and C-type units along the allele. No significant inter-allelic alignment of these internal variant repeat maps is evident, despite the fact that these alleles are all of similar length. This is in marked contrast to alleles at MS32 at which internal maps can show clear inter-allelic homologies (Jeffreys *et al.*, 1990). Several alleles show B-type repeat units extending from the stable 5' end into the central region of the repeat array, before the familiar A/C-type combination of repeats commences. One unusual allele (no. 6) has A, K and M repeat units, rather than A/C, at the 3' end of the central region; this allele also has unique haplotypes in the stable 5' and 3' regions.

The chimpanzee, gorilla and orang-utan have 18, 21 and 15 repeat units respectively at MS1 in the individuals studied (figure 3.9). There is considerable repeat unit sequence variation in all species, although shortened versions of the stable 5' and 3' human minisatellite regions are preserved in chimpanzee and gorilla. Other than the first two and the last repeat unit, these stable 5' and 3' regions are not evident in the orang-utan. No length polymorphism at MS1 was detected in the chimpanzee and orang-utan individuals tested, consistent with the substantial inter-repeat divergence observed. In contrast, low level polymorphism is detected in the gorilla (figure 3.7), which also shows a relatively homogeneous central run of B-type repeat units possibly analogous to the variable internal region of the human alleles.

Finally, the colobus monkey MS1 allele analyzed contains approximately 84 repeat units. While this allele is composed of repeats of the same minisatellite consensus sequence seen in the Hominoidea, there are none of the groupings of variant repeat units characteristic of the ends of alleles in man and the African apes. Instead, the colobus allele has many unique variants extending a considerable distance into the minisatellite, with no evidence for a relatively homogeneous internal region.

3.11 Discussion

PCR amplification and DNA sequence analysis of primate minisatellite loci enable the ancestral states and evolutionary dynamics of human hypervariable loci to be explored. Inter-species comparisons of minisatellite organization at MS1 and MS32 determined in the present study are summarised in figure 3.10.

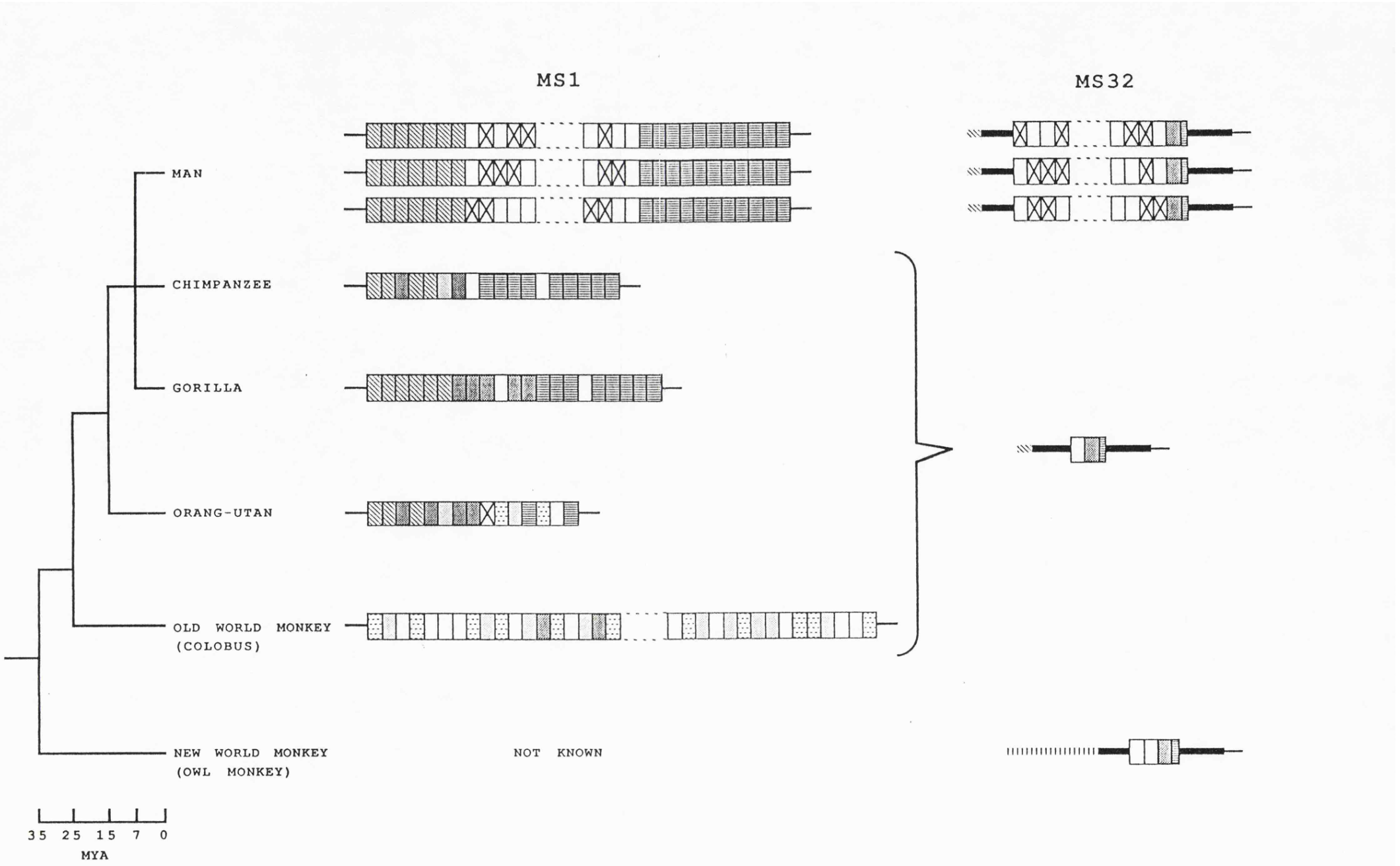
(a) The evolutionary history of MS32

MS32 has evolved by duplication within a retroviral LTR homologous to a haptoglobin associated retrovirus-like element (Maeda 1985; Armour *et al.*, 1989). The age of the retroviral insertion is unknown, but must have predated the divergence of the Cercopithecoidea and Ceboidea approximately 35 million years ago. The homology between the MS32 loci in man and the owl monkey ends abruptly at the 5' boundary of the LTR, suggesting that excision of an associated provirus has occurred in one species but not the other, perhaps by recombination between the two retroviral LTRs. Curiously though, neither the human or the owl monkey sequence 5' of the LTR resembles the haptoglobin retrovirus, even though 70% sequence similarity exists between the minisatellite and haptoglobin LTRs.

The initial duplications of minisatellite repeat units within the LTR also predated the divergence of Old World and New World monkeys, and presumably generated a short and very stable tandem array consisting of two complete and similar repeat units X and Y and a truncated diverged repeat unit Z. This stable monomorphic ancestral state of the human locus, XYZ, has remained unchanged in the last 25 million years and is found in all great apes and Old World monkey species examined. In the South American owl monkey, the MS32 locus is longer (XXYZ); it is not known whether this change in repeat copy number has occurred recently in this species, or whether it is the relic of other ancient changes in repeat copy number early in primate evolution.

In contrast to the stable organization of MS32 in great apes and Old World monkeys, this minisatellite has attained extreme variability in man over a remarkably short period, subsequent to the divergence of man and

Figure 3.10. Evolutionary comparisons of minisatellite structure in primates. Divergence times are taken from Koop *et al.* (1986). For MS1, the stable 5' and 3' human haplotypes of variant repeat units are shown by diagonally and horizontally shaded boxes respectively. Stippled boxes represent a variety of other variant repeats. Dashed lines represent multiple internal repeat units. For MS32, the flanking LTR is shown as a thick line, preceded by a different sequence in the owl monkey compared with other species. T- and A-type repeats are shown as open and crossed boxes respectively, and the 3' variant repeats are shaded. The three MS32 repeat units in great apes and Old World monkeys are referred to in the text as XYZ.


MS1
MS32
MAN
CHIMPANZEE
GORILLA
ORANG-UTAN
OLD WORLD MONKEY
(COLOBUS)
NEW WORLD MONKEY
(OWL MONKEY)
NOT KNOWN
35 25 15 7 0
MYA

the great apes approximately 7 million years ago (Koop *et al.*, 1986). Contemporary human alleles have the structure (X)nYZ, where n = 10 - 600+, indicating that amplification of repeats in the ancestral allele was restricted to the first ancestral repeat unit.

(b) Evolutionary transience of hypervariable loci

The question now arises whether the mutation rate to new length alleles at MS32 seen in contemporary human populations is sufficiently high to have enabled MS32 to evolve high repeat copy number in the last 7 million years (~700,000 generations), assuming a model of unequal sister chromatid exchange in which mutational increases and decreases in copy number occur with equal frequency (Jeffreys *et al.*, 1988, 1990). Hypothetical pathways of the evolution of a minisatellite are shown in figure 3.11. All pathways eventually terminate with an allele containing a single repeat unit, which, in the absence of *de novo* reduplication, should be immune to further changes in repeat unit copy number. A minority of pathways will show, by chance, a transient phase of high numbers of repeat units. Assuming that mutation rate increases with allele length (no data yet exist to confirm or refute this suggestion) then the phase of high repeat copy number will be associated with high mutation rate and thus population hypervariability. To estimate the number of generations required to generate, maintain and lose a hypervariable locus, single haploid chromosome lineages of MS32 were simulated by computer, commencing with two repeat units (program written by Professor. Alec Jeffreys). Mutations in repeat copy number were generated with a frequency and magnitude governed by current best estimates of MS32 mutational processes (table 3.2 legend; see Jeffreys *et al.*, 1990), and assuming proportionality between allele length and mutation rate. The simulation data are summarized in table 3.2.

Contemporary human alleles of MS32 contain on average 200 repeat units. If mutation rate is proportional to allele length, then most haploid lineages will show alleles slowly drifting around 2 repeat units prior to eventual terminal collapse to a single repeat unit, consistent with MS32 stability in great apes and Old World monkeys. Only one lineage in 250 will ever attain a transient phase of high (>200) repeat unit copy number. MS32 in man, which was experimentally chosen for its hypervariability,

Figure 3.11. Hypothetical evolutionary pathways of a single haploid chromosome lineage of a minisatellite, commencing with 2 repeat units. Solid line, a pathway which by chance exceeds a target number of repeat units prior to reduction to a single repeat unit, at which point (X) variation in repeat unit copy number is assumed to cease. Dotted line, a pathway which fails to exceed the target. The following sequential phases, in numbers of generations, can be defined along the first pathway, where n = number of repeat units: lag, from the start to the last generation at which n = 2, prior to n exceeding the target; gain, from the last generation where n = 2 until n exceeds the target; dwell, from the first generation where n = target to the last when n = target; decay, from the last generation with n = target to the first generation at which n is reduced to ≤ 2; extinction, from last n = target to n = 1.

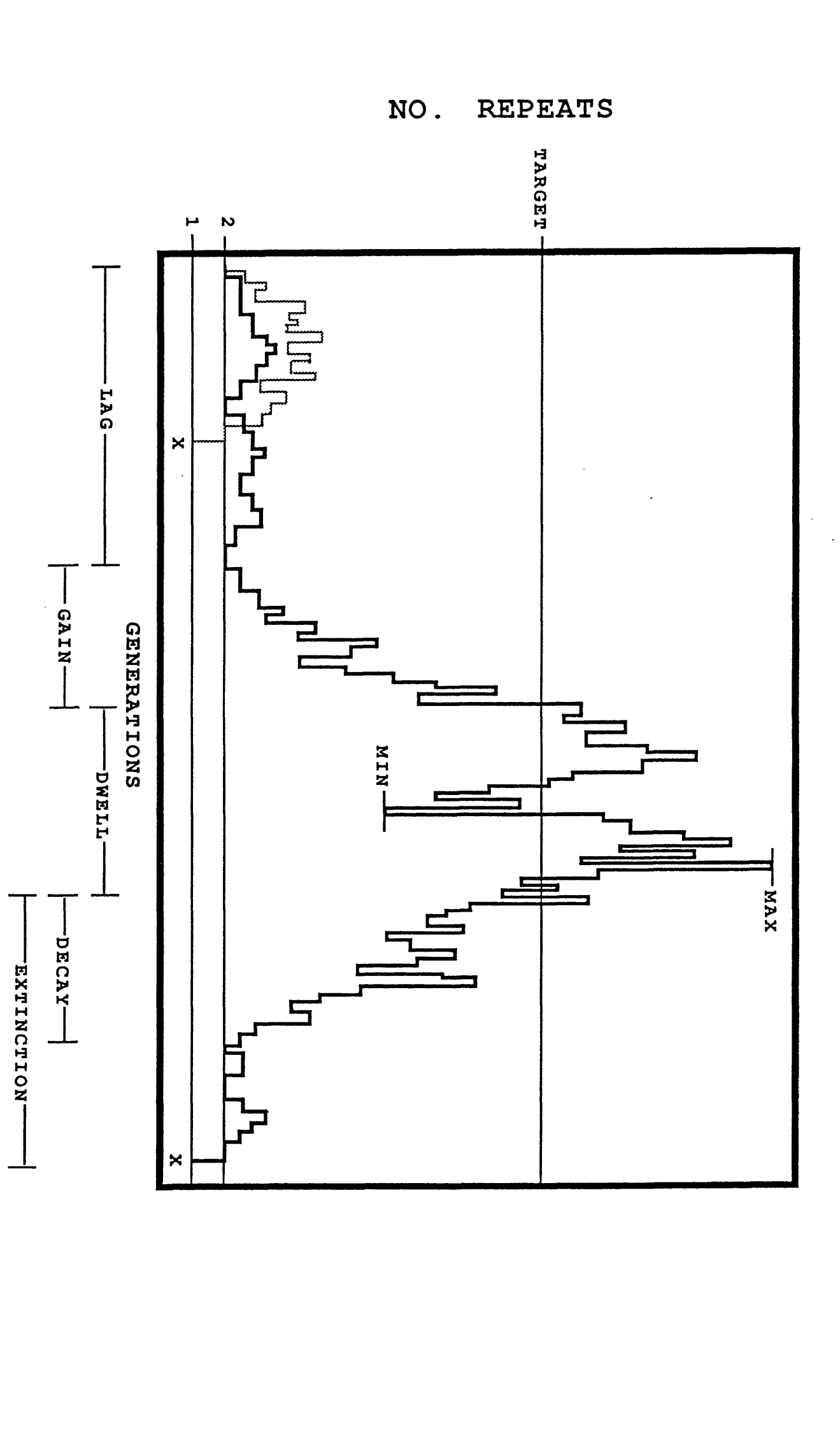

NO. REPEATS
TARGET
1
2
X
X
MIN
MAX
LAG
GAIN
DWELL
DECAY
EXTINCTION
GENERATIONS

Table 3.2. Computer simulations of the evolutionary dynamics of a single haploid lineage of a minisatellite, commencing with two repeat units. Phases of the evolutionary pathway are shown in figure 3.11. Simulations were performed as follows. The repeat-unit copy number n of an allele with initially two repeat units was sequentially altered at each generation with a probability defined by the mutation frequency μ per generation, until n eventually reached 1, at which point no further change in n is assumed to occur and the simulation was terminated. Mutation rate was assumed to be proportional to array length. Data from minisatellite MS32 suggest a mutation rate of 0.0014 per gamete for alleles 150 repeat units long (Jeffreys *et al.*, 1990). The mutation rate μ in the simulation was therefore defined as $9\text{x}10^{-6}(n\text{-}1)$, whence $\mu=0$ when $n=1$. Mutational increases and decreases in n were assumed to occur with equal frequency. At each mutation event, n was assumed to increase or decrease by between 1 and $(n\text{-}1)$ repeat units as expected for an unequal exchange process. From MS32 mutation data given in Jeffreys *et al.* (1990), it is assumed for all alleles that the probability at each mutation event of gaining or losing precisely i repeat units is given by

$$(n\text{-}i)^{3.4} \Big/ \sum_{i=1}^{i=n-1} (n\text{-}i)^{3.4}$$

This empirical relation reflects the fact that most mutation events involve the gain or loss of a relatively small number of repeat repeat units, and was used to select the precise number of repeat units gained or lost at each simulated mutation event. The table summarizes data from 100 simulated lineages where n transiently exceeded the the target number of repeats. Standard deviations are given in brackets. For highly skewed distributions, median and 95% limit values are also given.

	Target no. of repeats			
	20	50	200	500
Proportion of lineages reaching target	0.053	0.018	0.0040	0.0018
Duration in thousands of generations:				
LAG	340 (400)	360 (470)	290 (370)	340 (460)
GAIN	470 (300)	580 (360)	730 (480)	650 (320)
DWELL	250 (360)	190 (370)	60 (150)	15 (35)
median (95% limits)	66 (3-870)	43 (1-890)	9 (0.1-480)	5 (0.05-55)
DECAY	360 (310)	460 (380)	600 (450)	620 (410)
EXTINCTION	660 (410)	830 (570)	1000 (690)	980 (580)
MAXIMUM NO. REPEATS	71 (130)	390 (970)	1400 (4700)	4100 (9300)
median (95% limits)	32 (20-240)	95 (50-1500)	300 (200-3700)	1000 (500-26000)
MINIMUM NO. REPEATS	13 (6)	32 (16)	140 (69)	330 (150)
median (95% limits)	15 (4-20)	35 (4-50)	164 (10-200)	350 (66-500)

would therefore represent the product of such a successful lineage. Simulated lineages which achieve in excess of 200 repeat units will have done so on average within the last 700,000 generations (the 'gain' phase, figure 3.11; see table 3.2), a period similar to that elapsed since man diverged from the great apes. It can therefore be concluded that mutation rates at MS32 are sufficiently high to rapidly generate a hypervariable locus such as MS32, even under a model of unbiased sister chromatid exchange. It is therefore not necessary as yet to invoke additional processes, such as saltatory amplification or a mutational bias in favour of gains in repeat copy number (Charlesworth *et al.*, 1986; Stephan, 1986, 1989; Walsh 1987), to account for the rapid evolution of MS32. However, uncertainties in mutation rates for the shortest alleles, which occupy most of the 'gain' phase (figure 3.11) could seriously distort these simulation estimates. In addition, very recent data resulting from variant repeat analysis of MS32 show that evolution does not occur exclusively along haploid lineages, but can occasionally involve inter-allelic exchange. Furthermore, there would appear to be a bias in favour of gain of repeat units (Jeffreys *et al.*, 1991b).

Single chromosome lineage simulations also make it possible to estimate the length of survival of a high copy number state (the 'dwell' phase, figure 3.11) prior to eventual reduction to a single repeat unit (table 3.2). While the lengths of the 'lag', 'gain', 'decay' and 'extinction' phases (figure 3.11) are relatively insensitive to the target number of repeat units, the 'dwell' phase shortens dramatically with repeat copy number, indicating that an unstable locus such as MS32 will tend to be very transient in the genome, typically surviving as a high copy number locus for only 10,000 generations. However, further simulations on populations, rather than single chromosome lineages, are required to obtain more detailed estimates of the length of the 'dwell' phase of population hypervariability at such loci.

(c) MS32 allele structure

Contemporary MS32 alleles in humans consist of arrays of two intermingled variant repeats termed A and T, which differ by a single base substitution (Jeffreys *et al.*, 1990). The almost complete homogeneity of human minisatellite repeat units appears to be due not to concerted

evolution of an ancient repeat sequence family, but rather to the very recent amplification of MS32 in man. The ancestral minisatellite in great apes contains only T-type repeat units. Presumably, the earliest stage of evolution of the human variable minisatellite involved the expansion from TT alleles in apes to TTT, TTTT etc. At some stage, a base substitution occurred which introduced an A-type repeat unit into a (T)n array, followed by spread of the variant repeats through descendant arrays to give the enormous number of different A + T type alleles currently present in human populations. Neither the ancestral TT allele, nor the hypothetical descendant TTT, TTTT etc alleles have yet been detected in man (Jeffreys *et al.*, 1990), implying that these ancestral allelic states have become extinct in contemporary human populations.

(d) Evolution of MS1

A similar evolutionary picture emerges for the human minisatellite MS1 (figure 3.10). Again, the tandem repetitive state is ancient, and precedes the divergence of great apes and Old World monkeys approximately 25 million years ago. Unlike MS32, minisatellite MS1 has attained high repeat unit copy number and variability on at least two separate occasions, first within recent human evolution and second within some or all Old World monkeys. The immediate ancestral state of the human minisatellite appears to be the short and relatively invariant locus seen in contemporary great apes, although the possibility of a high copy number ancestor cannot be excluded. Amplification of MS1 in man appears again to have occurred very recently, but has involved only the central region of the minisatellite, the 5' and 3' regions of the array being largely invariant between human alleles and shared with the short African ape alleles. Similarly, contemporary human polymorphism in the number of tandem repeats and the distribution of variant repeats appears to be confined to this relatively homogeneous central region, in contrast to MS32, where variation in the distribution of A- and T-type repeats can extend over the entire minisatellite array (Jeffreys *et al.*, 1990). It is not known why there should be such a sharp transition between the stable 5' and 3' ends of MS1 and the extremely variable central region (figure 3.9). One possible explanation is that base substitutions in these terminal regions interfere with repeat unit turnover via interactions (through

slippage, unequal crossing over and branch migration at Holliday junctions) with the central variable region.

(e) Base substitutions in minisatellite flanking DNA

Human minisatellites are not randomly dispersed in the genome but are preferentially localized in sub-telomeric regions, frequently showing clustering with other minisatellites and associated with dispersed repetitive elements such as Alu, L1 and retroviral LTRs; in some cases, such as MS32, the minisatellite has amplified from within a retroposon (Royle *et al.*, 1988; Armour *et al.*, 1989; Kelly *et al.*, 1990). One possibility is that minisatellites mark unstable genomic regions which are prone to processes such as minisatellite amplification and retroposon insertion. To test whether instability extends to base mutation, we have analysed the level of sequence divergence in the region flanking minisatellite MS32 (table 3.1). Interestingly, there is a highly significant (approximately 50%) elevation of divergence compared with that predicted from thermal stability and non-coding DNA sequence analysis, implying an elevated substitution rate in the region of the minisatellite. It is not known whether this increase is specifically associated with minisatellite containing regions, or instead reflects shifts in mutation rate correlated for example with compositional isochores in the mammalian genome (Wolfe *et al.*, 1989).

(f) Other evolutionary studies of minisatellite loci

Few other studies have focussed on the evolutionary dynamics of minisatellite loci. Two polymorphic tandem arrays associated with the human ψζ-globin gene, a 36bp tandem repeat in the 5' flanking region, and a 14bp tandem repeat in the first intron (Goodburn *et al.*, 1983) have been examined in apes by Southern blot analysis (Chapman *et al.*, 1981; Chapman & Wilson, 1982; Chapman *et al.*, 1985). This revealed a series of allele lengths identical in man and the great apes at each locus, suggesting that the same sets of alleles have persisted for millions of years. However, sequence comparisons of human and chimpanzee alleles of identical length from the minisatellite in the first intron of the ψζ-globin gene demonstrated disparity in the arrangement of variant repeat units (Willard *et al.*, 1985), possibly due to multiple length changes on

one or both lineages, or due to gene conversion-type events between repeat units. Both ψζ-globin associated loci show limited variability; the array in the intervening sequence having only 4 alleles with a heterozygosity of approximately 65%, and that in the immediate 5' region having just 3 alleles and an estimated heterozygosity of 50% (Chapman *et al.*, 1985). Consequently these loci do not provide a good comparison with the hypervariable loci MS1 and MS32, each of which has in excess of 40 alleles distinguishable by length (Wong *et al.*, 1987), and probably $>10^8$ alleles distinguishable by internal structure (Jeffreys *et al.*, 1991b). Indeed, this finding is in agreement with the hypothesis that minisatellites with low variability are more stable and enduring in evolutionary terms than highly variable loci, which are by contrast highly transient.

Minisatellite MS29, with what may be described an intermediate level of variability in man (9 alleles with a heterozygosity of 61%) has, like MS1, been shown by Southern blotting to be much shorter and relatively invariant in the chimpanzee and gorilla (Wong *et al.*, 1990). A more thorough evolutionary study has recently been performed on two human minisatellites, cMS608 and cMS630 (Armour *et al.*, 1991). cMS608 has 10 alleles and a heterozygosity of approximately 70% in man, and consists of a 67-68bp repeat unit which has expanded from the junction point of two Alu elements fused in tandem. The great ape homologue of this array, as analysed by PCR, comprises of just the two Alu elements, with no evidence of a minisatellite other than an extended polyadenosine tail of the structure $(AAY)_n$ from the 5' Alu repeat, a feature also characteristic of the human locus. Interestingly, a very short allele of just 3 repeat units is present in man at a frequency of 30% at this locus, possibly a persisting relic of the initial amplification events leading to the minisatellite, and perhaps suggestive that this hypervariable array is very young indeed.

cMS630 consists of an 8bp tandem repeat with a heterozygosity of 84% in man, and interestingly shows a bimodal allele distribution; approximately 5 alleles of around 500 repeat units and a further 5 around 100 repeat units (Armour *et al.*, 1991). A limited number of chimpanzees and gorillas typed at this locus by PCR showed allele distributions similar to, but distinct from, that of the smaller human alleles, implying that the variable tandem repeated state is ancient, and that allele frequency

distributions have altered with species divergence. The larger human alleles are likely to be recent, possibly resulting from the generation of a single large allele with further smaller mutation events within this allele giving rise to the observed secondary distribution.

Finally, Wolff *et al.* (1991) have cross-hybridized 12 human VNTR probes to ape DNA by Southern hybridization and found patterns of hypervariability versus non-variability between species which appear to be random and do not reflect established phylogenetic relationships, in agreement with the notion of extreme evolutionary transience of these loci. PCR analysis of one such locus (variable in man but monomorphic in apes) revealed a single repeat unit in chimpanzee, gorilla and orang-utan, presumably representing the ancestral precursor to the human minisatellite.

(g) Conclusion

These studies show that large fluctuations in the degree of polymorphism at variable loci can occur over remarkably short evolutionary time periods, rendering the ability of single locus minisatellite probes to detect informative loci by cross-species hybridization largely unpredictable. Furthermore, those loci with the greatest level of variability are likely to be short-lived in evolutionary terms, whereas a lower mutation rate confers relative stability and a greater chance of cross-species variation. Although multilocus 'DNA fingerprinting' probes have been shown to detect variable loci in a wide range of species, these results suggest that the patterns generated by such probes do not necessarily represent homologous groups of hypervariable loci in different species, even if those species are closely related.

CHAPTER 4

MAPPING VARIANT REPEAT UNITS WITHIN THE MINISATELLITE MS1

4.1 Introduction

Sequence analysis of cloned minisatellites has shown that the repeat units which constitute the tandem array are rarely identical, but frequently display sequence variation between repeat units (Owerbach & Aagaard, 1984; Jeffreys *et al.*, 1985; Wong *et al.*, 1986, 1987; see previous chapter). The evidence would suggest that those minisatellites that show the greatest level of length variability have the lowest degree of sequence variation between repeat units, probably because such arrays are very young and have had insufficient time to accumulate variants (see chapter 3), or alternatively due to higher rates of sequence homogenization during mutation to new length alleles through processes such as unequal exchange (Jeffreys *et al.*, 1985). However, repeat unit sequence variability has been shown to be present even in the most variable loci (Wong *et al.*, 1986,1987; see previous chapter).

With the advent of PCR, it has been possible to explore such inter-repeat variation within the minisatellite MS32, dramatically extending the number of distinguishable alleles within the population and giving an insight into the processes that generate variability (Jeffreys *et al.*, 1990). The initial approach for minisatellite variant repeat (MVR) mapping at MS32 involved PCR amplification of minisatellite alleles from genomic DNA followed by gel purification. Gel-purified alleles were then end-labelled, divided into two aliquots and partially digested with *Hin*fI, which cleaves every repeat unit, and *Hae*III; ~2/3 of repeat units contain a *Hae*III site whereas the remainder do not, due to a G ->A transition. The two variants are designated a- and t-type. Agarose gel electrophoresis and subsequent autoradiography of the resulting fragments produced a ladder of bands from which the number of minisatellite repeats could be determined, and comparison of *Hin*fI and *Hae*III digests allowed each repeat to be scored a- or t-type (Jeffreys *et al.*, 1990). Although a powerful technique for studying internal variation at MS32, this method was cumbersome, and more significantly was limited to those alleles which could be efficiently amplified by PCR (≤ 6kb).

4.2 Internal mapping using repeat-specific PCR primers (Jeffreys *et al.*, 1991b)

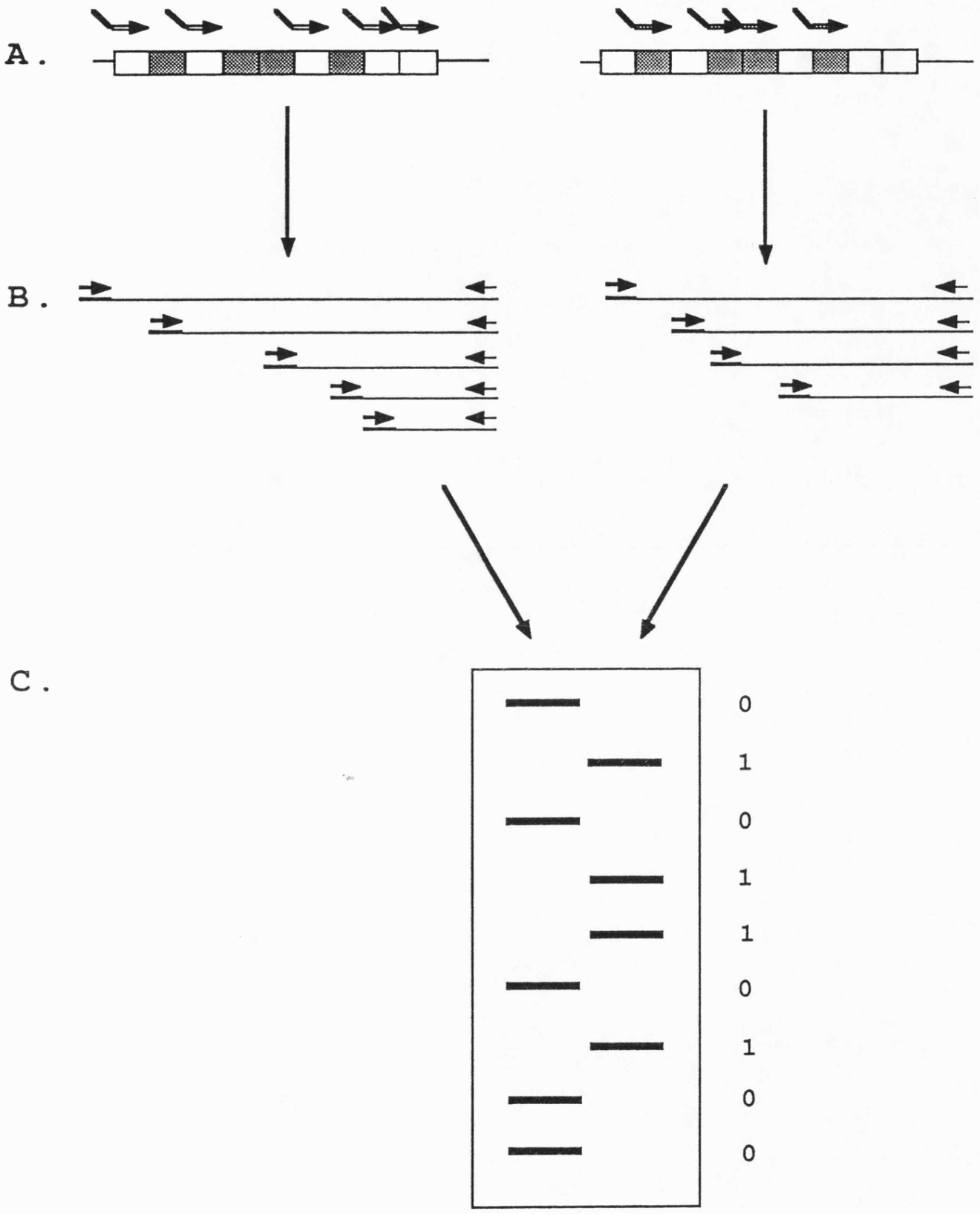
An improvement over the original MVR mapping technique came with the realization that a single base mismatch at the extreme 3' end of a PCR primer could completely prevent priming under some circumstances (Newton *et al.*, 1989). This was exploited to design two PCR primers internal to the minisatellite MS32, with the 3' end of each primer designed such that priming would be possible from one variant but not the other and vice versa. Using an MS32 allele which has been PCR-amplified and subsequently gel purified as a template, amplification using one or other of the MVR specific primers plus a primer in the DNA immediately flanking the minisatellite should generate two sets of complementary products, from which the MVR map may be deduced. However, progressive shortening of products due to internal priming of nascent PCR products by the MVR specific primer would lead to a successive 'fading out' of the MVR map above a limited number of repeat units. Consequently, each MVR specific primer has an additional non-complementary 5' extension or 'tag'; PCR reactions may then be performed with a limited concentration of the MVR specific primer and a high concentration of the flanking primer and a primer consisting of the tag sequence (figure 4.1A, B). Priming by the MVR specific primer plus the flanking primer generates an initial set of products which are subsequently amplified by the flanking and tag primers. Shortening of products by internal priming is minimized due to the low concentration of the MVR specific primer. Agarose gel electrophoresis of the PCR products followed by Southern blotting, hybridization with an MS32 probe and autoradiography gives two complementary ladders representing an internal map of the minisatellite allele (figure 4.1C). This technique is not restricted to gel-purified PCR amplified alleles, but can be applied directly to genomic DNA, to give the two allelic profiles superimposed. >60 'rungs' on such a ladder can easily be resolved for MS32, to provide a digital (and hence non-subjective) individual-specific profile amenable to database storage (A.J. Jeffreys *et al.*, 1991b). The following chapter describes the application of this technique to a second minisatellite, MS1.

Figure 4.1. Strategy for mapping internal variation in a single minisatellite allele.

A. A PCR amplimer capable of priming from just one of the two variant repeat unit types towards the 3' end of the minisatellite array is added at limiting concentration to a PCR reaction containing the minisatellite allele. A second reaction with an amplimer designed to prime from the other variant is also performed. One variant is depicted as an unfilled box and the corresponding primer as an unfilled arrow, whereas the other variant is drawn as a filled box and the primer as a filled arrow. Each of the variant repeat-specific primers has an additional non-complementary 5' extension or 'tag', shown as a thick line.

B. Following generation of single-stranded products of lengths corresponding to the internal arrangement of variant repeats, two further PCR amplimers, one which is complementary to the immediate 3' flanking DNA (thin arrow) and a second consisting of the tag sequence (thick arrow), allow generation of a series of heterogenously lengthed double-stranded products. This uncoupling of MVR pattern detection from subsequent amplification prevents progressive shortening of PCR products due to internal priming by the MVR specific primers. In practice, stages A and B can be performed as a single reaction which includes the flanking sequence primer and the tag primer in conjunction with the internal repeat primer at limiting concentration.

C. The products generated from each variant-specific primer are electrophoresed alongside one another through an agarose gel and detected by Southern blotting and hybridization with a minisatellite probe to give a binary code. When applied to total genomic DNA, the profiles of the two alleles are superimposed, giving a ternary code.


A.
B.
C.
0
1
0
1
1
0
1
0
0

4.3 Primer design for MVR mapping MS1

As described in chapter 3, MS1 is the most variable minisatellite cloned to date, with a heterozygosity of >99% based on allele length (Wong *et al.*, 1987; Armour *et al.*, 1989). It consists of a 9bp tandem repeat, with two common variants apparently making up the variable section of the array in most cases. These two variants differ by an A -> G substitution at the centre of the repeat unit and are designated A and C-type (see chapter 3). Consequently, two 20-mer primers were designed, one for amplification from each repeat type toward the 3' end of the minisatellite (figure 4.2). As described above, a 20-mer tag sequence was added to the 5' end of each variant-specific primer. Due to the limited length of MS1 repeat units (9bp) it was necessary to introduce redundancy at two positions in each primer, to allow for either variant upstream of the repeat unit of interest.

4.4 MVR mapping MS1

Varying primer concentration, annealing time and temperature and cycle number determined the optimum conditions for MVR mapping as follows: a 10µl reaction containing 100ng DNA in the PCR buffer described previously (chapter 2, section 2.2.18a) with 10nM repeat primer, 1µM tag primer and 1µM flanking primer B (see chapter 3, figure 3.1). Limiting amounts of internal primer are necessary to prevent the generation of an excess of short products. Annealing temperature was 67° and the reaction cycled 18 times with two chases comprising 5 minutes at 67° and 10 minutes at 70° (as described in chapter 2, table 2.2).

PCR products were electrophoresed through a 40cm 2% agarose (Sigma) gel in TBE at 150 volts, alongside ϕX174 x *Hae*III markers until the 118bp marker had run off the bottom of the gel (~15 hours). The DNA was then transferred to a nylon membrane (Amersham hybond-N) and hybridized with a ^{32}P-labelled MS1 *Ava*II minisatellite fragment (see figure 3.1, chapter 3). If all repeat units within the minisatellite were A or C-type, 3 states for each position in the internal map would be possible for genomic DNA: (i) a band in the A track but not the C track, indicating that both alleles have an A-type repeat at this position; (ii) a

Figure 4.2. Internal repeat primers for mapping variation within MS1. Due to the limited length of the MS1 repeat unit, each primer overlaps two variable positions in repeat units preceding the one to be mapped, rendering it necessary to intoduce redundancy at these positions.

TAG sequence: 5′-TCATGCGTCCATGGTCCGGA-3′

A-specific primer: 5′-TCATGCGTCCATGGTCCGGAT$^{A}/_{G}$TCCACCCT$^{A}/_{G}$TCCACCCTA-3′

C-specific primer: 5′-TCATGCGTCCATGGTCCGGAT$^{A}/_{G}$TCCACCCT$^{A}/_{G}$TCCACCCT©-3′

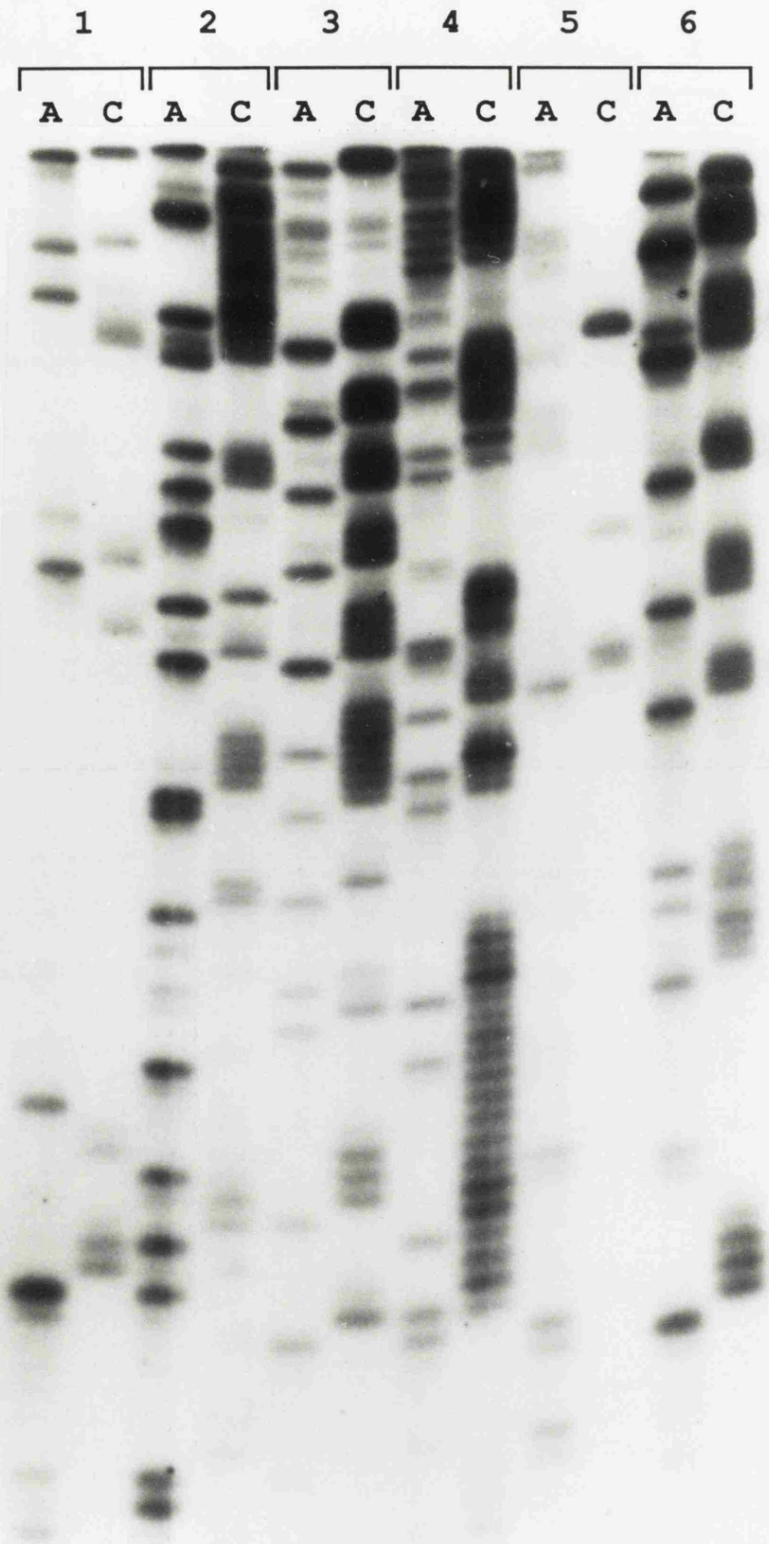
band in the C track but not the A track, indicating that both alleles have a C-type repeat at this position; (iii) a band in both lanes, if one allele is A-type and the other C-type.

Figure 4.3 shows MS1 internal maps from 6 unrelated individuals. It is immediately apparent that MS1, like MS32, has a high degree of internal variation between alleles, with no two individuals having similar internal map profiles. However, many positions in each ladder do not have a band in either the A or C track, implying that priming off many repeats is not possible with either the A or C amplimer, possibly due to further variant repeats within the array. This phenomenon of 'null' repeats is particularly evident in individuals 1 and 5, which appear to have arrays in both alleles consisting predominantly of repeats which are refractory to priming by primer A or C. An alternative explanation for null repeats is that the PCR conditions are such that primer annealing within the array is stochastic, allowing amplification from some repeat units within the array but not others on a random basis. However, duplicate reactions containing the same input genomic DNA gave the same profiles, suggesting that this is not the case (not shown), although the signal from the fainter bands at the bottom of the maps may be increased by adjusting the concentration of the internal primers. In order to identify any further variant repeats within MS1, the 3' ends of alleles containing null repeats were sequenced.

4.5 Identification of 'null' repeat units within MS1 alleles

Internal mapping reactions as described above were performed, but in a 50µl reaction and with the number of cycles increased to 40. This allowed the MVR profiles to be visualized directly in an ethidium stained gel. Bands from the internal map ladders which were positioned above 'null' rungs were gel purified and digested with *Ava*II, which cleaves within the 3' flanking DNA between the end of the minisatellite and primer B (see figure 3.1B, chapter 3) and also cuts within the tag primer sequence (figure 4.2). Following filling in of the 3' overhangs with the klenow fragment of DNA polymerase I, these 3' ends of MS1 alleles were cloned into the *Sma*I site of M13mp18 (Yanisch-Perron *et al.*, 1985) and sequenced. (Attempts to directly sequence double-stranded PCR products failed, possibly as a result of rapid reannealing of the template strand with its complement due to the simple repeated nature of MS1).

Figure 4.3. Internal maps of MS1 from 6 unrelated individuals. 'A' denotes A-type repeats, 'C' denotes C-type repeats. Although there is obviously considerable internal variability at MS1, there are evidently more than two common types of repeat unit, with other variants manifested as gaps in the MVR maps.


1
2
3
4
5
6
A C A C A C A C A C A C

Sequence analysis of independent clones of each allelic 3' end revealed gross rearrangements of the repeat unit arrays, with different clones of the same allele showing different internal structures, presumably due to rearrangements during M13 propagation. Such rearrangements were not apparent when cloning short MS1 alleles into Bluescript vectors (see chapter 3) possibly due to the use of the recombination deficient XL1 Blue (Bullock *et al.*, 1987) as a host instead of the *rec*$^+$ NM522 (Gough & Murray, 1983) used here. Although scrambling of the arrays during cloning rendered comparison with MVR maps and subsequent identification of null repeats unfeasible, two variant repeats other than A and C type could be identified within the arrays: B type (CCCTCTCCA) and K type (CCCTATTCA). Both of these variants had been seen previously within the variable region of the minisatellite array (see chapter 3). It therefore seemed likely that the inhibition of priming from either of these repeat unit types was responsible for null repeats in MVR maps.

4.6 Discussion

PCR MVR mapping has been shown to be an incredibly powerful identification system which allows digitization of individual specific DNA codes, and subsequent storage and non-subjective comparisons of profiles in computer databases, of enormous consequence in forensic science (Jeffreys *et al.*, 1991b). However, few minisatellites meet the criteria required for PCR MVR, which include sufficient internal variability to give a reasonable level of discriminatory power, a repeat unit length which allows resolution of a large number of repeats on agarose gel electrophoresis, an absence of microdeletions/insertions between repeat units and a limited number of well-dispersed variant repeat unit types. MS32 fulfils all of these criteria. (Jeffreys *et al.*, 1991b).

MS1, although an obvious choice for PCR MVR with a very high degree of internal variation (see chapter 3 and figure 4.3), appears to possess at least four common variant repeat unit types. Although it may be technically possible to MVR map all variants by using further internal primers, the length of the MS1 repeat unit (9bp) requires redundancy to be introduced in each of the 20-mer internal primers (see figure 4.2), and

the level of redundancy required to allow priming at all pairwise combinations of the four common variants upstream of the repeat unit of interest is likely to render PCR MVR unfeasible at this locus.

CHAPTER 5

'MICROSATELLITES' - $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ REPEAT LOCI

5.1 Introduction

$(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ repeat sequences or 'microsatellites', where n is typically 15-30, are ubiquitous dispersed repetitive elements of eukaryote genomes (Miesfeld *et al.*, 1981; Hamada *et al.*, 1982; Jeang & Hayward 1983; Tautz & Renz 1984a; Sun *et al.*, 1984) with a copy number of 30,000-100,000 in human DNA (Miesfeld *et al.*, 1981; Hamada & Kakunaga, 1982; Hamada *et al.*, 1982; Hamada *et al.*, 1984a), but which are not found in bacteria (Gross & Garrard, 1986). The function of these $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ blocks is unknown, but it has been proposed that they serve as hot-spots for recombination (Stringer, 1982, 1985; Pardue *et al.*, 1987; Hellman *et al.*, 1988) or gene conversion events (Slightom *et al.*,1980; Flanagan *et al.*, 1984); it has been noted that single-stranded breaks occur at a higher frequency in $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ tracts than in other nucleotide combinations and suggested that these breaks may serve as triggers for recombination (Rogers *et al.*, 1983a). Alternatively, $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ have been implicated in gene regulation (Russell *et al.*, 1983; Young, 1983; Hamada *et al.*, 1984b; Berg *et al.*, 1989), possibly as transcriptional enhancers, although this seem unlikely in view of the high abundance of $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ repeat blocks relative to the estimated frequency of eukaryotic enhancers (Weber *et al.*, 1983). Other suggested functions include a role in chromatin folding (Johnson *et al.*, 1978), telomere formation (Rogers, 1983a; Walmsley *et al.*, 1983), X-chromosome inactivation (Pardue *et al.*, 1987) and homogenization of repetitive gene arrays (Kedes, 1979).

The ability of alternating purine/pyrimidine sequences to form Z-DNA *in vitro* (Arnott *et al.*, 1980; McIntosh *et al.*, 1983; Haniford & Pulleybank, Nordheim & Rich 1983), possibly leading to novel biological function, has given rise to much of this speculation; however, it appears that such tracts are not in the Z-DNA conformation within cells (Tautz & Renz 1984a; Hamada *et al.*, 1984a; Rodriguez-Campos *et al.*, 1986). Rather than having a major biological function, it seems more likely that such sequences arise and are maintained as a consequence of mechanisms such as DNA slippage during replication (Levinson & Gutman, 1987a). Indeed, the occurrence of other classes of dinucleotide repeat sequences with similar frequency to $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ blocks, and of trinucleotide repeats at somewhat lower frequencies, would suggest that all such repeat

motifs arise through the same mechanisms (Tautz & Renz 1984a). It is therefore probable that these sequences are ubiquitous not through evolutionary conservation but rather through frequent independent formation.

It has recently been shown that $(dT\text{-}dG)_n \cdot (dA\text{-}dC)_n$ repeats are polymorphic with respect to length, due to variation in dinucleotide repeat copy number, adding further weight to the argument for generation and contraction/expansion by slippage-type events (see Levinson & Gutman, 1987a). Moreover, by designing PCR primers complementary to unique sequence DNA flanking microsatellites, this polymorphism can be exploited to provide a new source of genetic markers (Litt & Luty 1989, Weber & May 1989, Tautz 1989). It has been estimated that microsatellite markers with a Polymorphism Information Content (PIC) of ≥ 0.5 could yield genetic maps with an average resolution of approximately 0.3cM (Weber 1990).

In order to explore the nature and variability of these loci further, a genomic DNA library was constructed in M13 and screened with a synthetic $(dT\text{-}dG)_n \cdot (dA\text{-}dC)_n$ probe. DNA was prepared from a selection of the resulting positive plaques, and, following sequencing, PCR primers were designed for amplification of three microsatellite loci, one perfect $(dT\text{-}dG)_n \cdot (dA\text{-}dC)_n$ repeat and two imperfect repeats which contained base substitutions within the array. The variability of these loci in great ape DNA was also studied.

5.2 Construction and screening of an M13 library

All variable microsatellites isolated to date are in the size range 15-30 repeats (Weber 1990). In order to establish whether longer dinucleotide repeats with proportionally greater variability could be isolated, the following strategy was employed. High molecular weight DNA from 16 individuals was pooled and digested with *Alu*I, *Hae*III and *Rsa*I, to remove the bulk of the DNA flanking microsatellites. A 400-1000bp size fraction was selected by electroelution from an agarose gel onto dialysis membrane (Yang *et al.*, 1979) and ligated into the *Sma*I site of M13mp18 (Yanisch-Perron *et al.*, 1985). M13 was chosen as a vector to enable clones containing microsatellite inserts to be sequenced and PCR primers

designed without subcloning. The resulting library was screened with a synthetic $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ probe, constructed by annealing and ligating $(dT\text{-}dG)_{10}/(dA\text{-}dC)_{10}$ oligonucleotides and PCR amplifying the resulting concatamers. (Out of register annealing during PCR generated products several kilobases in length). Of a library of ~10,000 plaques, 60 proved positive. If the mean insert size is assumed to be 700bp, this gives an estimate of one microsatellite every 117kb, or 26,000 in the human genome, in reasonable agreement with a previous estimate of 30-60,000 by colony hybridization (Miesfeld *et al.*, 1981). 14 of the positive clones were selected for sequence analysis. (Thanks are due to Annette MacLeod for assistance with clone isolation and sequencing).

5.3 Size constraints and evidence for clustering of simple tandem repeats

The sequences of the microsatellite clones are summarized in figure 5.1. It is immediately apparent that despite attempting to optimise array length, all of the $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ repeat blocks are severely constrained in size, with the exception of the cryptic repeat comprising clone TG6. The longest perfect microsatellites are two runs of 24 dinucleotide repeats (clones TG1 and TG2). Two of the 14 clones possess more than one simple repeat sequence: clone TG23 has a $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ repeat followed by an imperfect $(dA\text{-}dT)_n\cdot(dT\text{-}dA)_n$ repeat 168bp downstream, with several short runs of simple repeats in between; similarly, clone TG1 has a $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ array 176bp downstream of a $(TTCC)_n$ tetranucleotide repeat, implying that certain DNA regions may be prone to the formation of such sequences. Scanning the EMBL database for sequences similar to the DNA flanking the microsatellites revealed no homologous sequences, with the exception of TG1, which was found to be embedded within the non-transcribed spacer of a ribosomal RNA gene (Dickson *et al.*, 1989). Significantly, the number of repeats in both the dinucleotide array and the $(TTCC)_n$ array differ between the two independently determined sequences, suggesting that these regions show length polymorphism (figure 5.1). Three of the microsatellites were chosen for further study: TG10, a perfect run of 17 dinucleotide repeats; TG29, consisting of 20 repeats disrupted in the centre by a G->T/C->A substitution, and TG9, an array of 19 repeats with A->T/C->G transitions at repeat numbers 7 and 15.

Figure 5.1. $(dT\text{-}dG)_n.(dA\text{-}dC)_n$ microsatellites isolated from an M13 library. Tandemly repeated arrays are shown in bold type; N denotes unreadable sequence. Clone TG1 is derived from a ribosomal RNA non-transcribed spacer and is shown in alignment with the previously published sequence of Dickson *et al.* (1989), designated RNTS. Both of the tandemly repeated arrays within this spacer show length variation between the two independently determined sequences. Sequence from which PCR primers were derived is underlined for clones TG9, TG10 and TG29.


TG1: ATGGGTTCTTCTGTGTCAT.GTCACGTTCTATCG.TTGCTTGCCTGCTTGCCTGTTATTTTCCTTCCTTCCTTCCTTCCTTCC-
RNTS: ATGGGTTCTTCTGTGTCATTGTCACGTTC.ATCGTTTGCTTGCCTGCTTGCCTGTTATTTTCCTTCCTTCCTTCCTTCCTTCC-
TG1: TTCCTTCCCTCCTTCCTTCCTTCCCNNNCTCCCTTACTGGCAGGTCTTCCTCTGTCTCTGCCGCCCAGGATCACCCCAACCT-
RNTS: TTCC.....TCCTTCCTTCCTTCCTTCCCTCCCTTACTGGCAGGTCTTCCTCTGTCTCTGCCGCCCAGGATCACCCCAACCT-
TG1: CAAACGCTTTGGACCGACCAAACGGTCGTTCTGCTCTGATCCCTCC.ATCCCATTAC.TGAGACTACAGGCGCG(N)55TGTGTGT-
RNTS: CAAACGCTTTGGACCGACCAAACGGTCGTTCTGCTCTGATCCCTCCCATCCCATTACCTGAGACTACAGGCGCG(N)55TGTGTGT-
TG1: GTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG.....
RNTS: GTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG
TG2: AACCTAAGCCTCAATTTCCTAAATTTGTAAAAATGCCAAGAGTAGTGCTTGAAGCCAATTAGTGGGTTTCAGCACAAATGACTGGT-
TCTATCATAAACTAGGGGCTTTCCATCTCAGCCCACAATATTCCTGTAATTTTTAAAGCAATCATTTATTTGTGCTTTAACCCTTA-
ATCCTTACCACAGTGATTATATTAACATTCTGGCTTTCTCAGGCTCCTACTCATCACTATCA(N)~70(TG)24
TG3: ACAAGAAAGACCAAGAACTGGATTACAATGTTGATACTGAATCAAAAGGATCAAAAAGTAAAATAAGAAACATAAGAATACACATC-
TTAGGGAGGTTTGAATAAAGAGTGAT(TG)13TACAAAAATGATCAAAGAGAATCAAAACTGAAAGGACAATATATATACTTTTTG-
ACCTGAGCCTAAAGTAA
TG4: TTTTATTCACCAGTTCATAAC(GT)23CGGGAAGGGGGGAGGTCCATGAATGACTTCATGTAACCACCACAGTCAAGATACAGAAC-
TAGACTCTCACCATAAGTTCCCCAATAGCTACTCCCTTATAGGC

TG5: TGTCCGCTGGAATAGC **(GT)** $_{21}$AAAGAGAGAGAGGNNAGGGAGGGAGGAAAACACCTCTTCCTTGATGCCTTAATTGATAACTGGT-
CTCTCAAAACCCATGCTACCAGCAGTGCCCAAGAAGTCCCC

TG6: GG (TG) $_3$CATGCAT (TG) $_3$ (CA) $_2$ (TG) $_3$CATT (TG) $_5$CACG (TG) $_3$GGTGCGTGCATGTGGGGTT (TG) $_4$CATGCATT (TG) $_2$CCTG-
TGCC (TG) $_2$GGGG (TG) $_2$ (CA) $_2$ (TG) $_3$CGCAGGTT (TG) $_5$CATGCT (TG) $_2$CGAGTGCATA (TG) $_2$CACG (TG) $_3$CATGCG (TG) $_4$C-
A (TG) $_2$CA (TG) $_2$TATGCA (TG) $_5$TAC (TG) $_4$

TG7: CCTGACACAGTATACTAGCCAGTGACAATCTTGGATGGTATCCCATCTCCAGATCACCAAAAAGTTTTAGCAGTGACAGGGGTTAT-
AAGAACTCCGTTATCATTTCACCACCAACTGAAGAGTGACCAACTCTAAAACTCTATTTTGAATATATGAATCTAGGACTATTCAA-
AACCAATGATCTACATTTCTTG (N) $_{\sim 40}$ **(TG)** $_{14}$

TG8: AAGGCATAAACATGTTTATTNNGAGAAAATATNCTCACCCTAGTGATCCTGTGTAGTTTATAGCAAGAATTTATAGCAAGAATTTT-
CTCTGAAATACAATACCACATATTTTNCATTGTATTATCTCACNTGGTAAAAGGATAGCCCTAATCTTGGAAGA (N) $_{\sim 100}$ **(TG)** $_{16}$

TG9: AGTTATGTGCTAATCTCCACTACTAGCCTGTGAGTTTCTTACAGCAAGACCCTGCCTCTTTGTTCATTTTTTTATATGTTGTAGCC-
CTTTACACAGTACTTTAGGAAGTTTAGGTTATT **(TG)** $_6$**AG (TG)** $_7$**AG (TG)** $_4$**T**ATTGGGAGTGGGGATTTGTGCAGGAAATGAG-
GAAAGAAGGAAAGTG

TG10: GAGAATGGTGGTTATCTAGGTCTTTAGTGGATACAGTACTGGAATTACTACTTCCAAGACTTCTCTTTCATAATTCTGAATTTAAA-
GGCACCAGATCCTTCGTTTAAT **(TG)** $_{17}$AACAGTTAACATTTTACAAAGGGGCTCTACCTGTGGTTATCAGTCCATCACTACATGT-
TTTGTAATGATTTCCTTCACTTGAGCCTGGCTGGTATTGCTTTAAAT

TG11: GCTAATTGCTTGTTGGTTCAAGAGAAATGTTATCTCAGTAGACTAACGTATGGACTTTCTGAGTCAGAAAATGCAAACTATGATGT-

GTAACCTGTTTGCCATTTTTTCTTTGTATCCGCAAACTGACAGCAAAATATATTGCACAAGTATTATAAATCTGTTCAGTTCAAGT-

GCCAGAGAACATTTACTCTGCTTCCTTCATGAGGTAATTCAAAACTTTGTTTTAATATTTATT(N)$_{\sim 50}$**(TG)$_{9}$A(TG)$_{10}$**

TG22: **TC(TG)$_{4}$TA(TG)$_{4}$TC(TG)$_{10}$(TC)$_{8}$G(TG)$_{14}$**CTTATCAGTTTCCCTGAGTGGGGTTACCTTTGGTTAACTAGTAAAATC-

ATTGCTATTTGGACTTTCCAGGCTGAGGATCCAAATCTAAAGTCATTGTAAACCAAGGCATTCACATCATGGGCATGAGATATCTA-

GTGAGTCTTCTTTCACAATTTCCCTTAATTCATTTTCTCCATTTCCTACTGTCATAATCATGATCC

TG23: TTTTTCAAGGTACA**(TG)$_{3}$TATGCG(TG)$_{24}$**AAAGACAGAAGAAAGAGGGAGACCTAAGAAGACTATGAGACACTAAGAGAAAA-

ATTAAGGTAAAAAAGACACACACTTAGAAAAACACACACACACATAGGGAGGAGGGAGGAGGTTAAGACATTTTACTATGTGCTGT-

GAATGGAAACTACAAACCATTTTTG**(AT)$_{3}$GC(AT)$_{4}$AC(AT)$_{3}$(AC)$_{3}$(AT)$_{2}$AC(AT)$_{2}$**

TG29: CCATTAACCA<u>TTTTGCCCTAGCCACACGCT</u>GAGGGTGATGGGTGGGTTGG**(TG)$_{10}$TT(TG)$_{9}$T**CAGGGGATGGGAGGGTGATCA-

AATGAATTAAATTT<u>GGGCTGGTTTTCAGTCTTTG</u>AGGTGGCAATTGTAAAATTTAATAATTTGAATAGATGCCTATTGATTCAAAA-

AACTGAAAAACCTGAAAAGGTACCCAGTGAAAAATGTAATAA

5.4 Analysis of TG10

20-mer PCR primers complementary to the unique sequence flanking the dinucleotide repeat within clone TG10 were synthesized. Figure 5.2A shows the profiles of the amplified products from 24 unrelated individuals typed on a high percentage agarose gel. This locus has a heterozygosity estimated at 74% and five different alleles in Caucasian populations. Allele sizes and frequencies are shown in figure 5.2B.

Amplification of this locus from chimpanzee and gorilla DNA revealed the dinucleotide array to be variable in these species also (figure 5.3) although in the limited sample studied, the gorilla has a slightly different allele size distribution when compared to chimp and human, with somewhat larger PCR products. Amplification from orang-utan DNA was not possible, presumably due to critical primer mismatches.

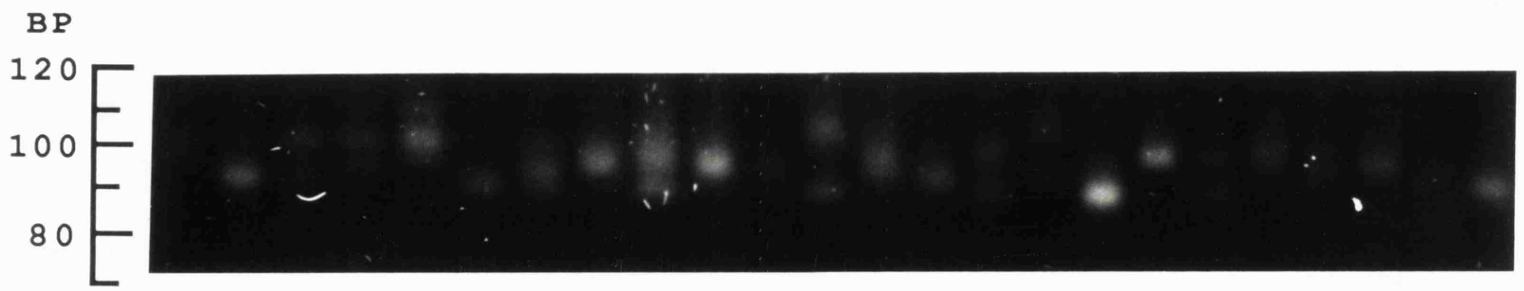
In order to assign TG10 to a human chromosome, the locus was PCR amplified from human/rodent somatic cell hybrid DNAs (provided by Dr. Sue Povey, University College London; see table 2.1) resulting in localization to chromosome 18. This positioning was confirmed by pedigree analysis using two large CEPH families (see figure 5.4) with the closest informative maker being CRI-R397 (Donis-Keller *et al.*, 1987; $z=5.3$ at $\theta=0.071$; linkage analysis was performed using the LINKAGE programs of Lathrop *et al.*, 1985 and the CEPH database version 2 with the assistance of Dr John Armour).

5.5 Analysis of TG29

As for TG10, PCR primers were designed for amplification of TG29, an imperfect $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ with a transversion disrupting the dinucleotide array (figure 5.1). This locus was found to be dimorphic in humans, with a heterozygosity of 38%, the two alleles having frequencies of 0.81 and 0.19 (figure 5.5). However, in great apes, TG29 appears to show a high level of variability, comparable to, if not greater than that of TG10 (figure 5.6).

Figure 5.2. Variability at microsatellite TG10.
A. TG10 alleles PCR amplified from 24 unrelated individuals. 10ng of genomic DNA was cycled 33 times in a 10µl reaction as described in table 2.2. PCR products were electrophoresed through a 4% agarose gel (3 parts Nusieve, 1 part HGT) in TBE with ethidium bromide.
B. Allele frequency distribution at TG10. TG10 shows a typical microsatellite allele frequency distribution (Weber & May, 1989); a limited number of alleles with one (98bp) being particularly common (Frequency = 0.43).

A.



B.

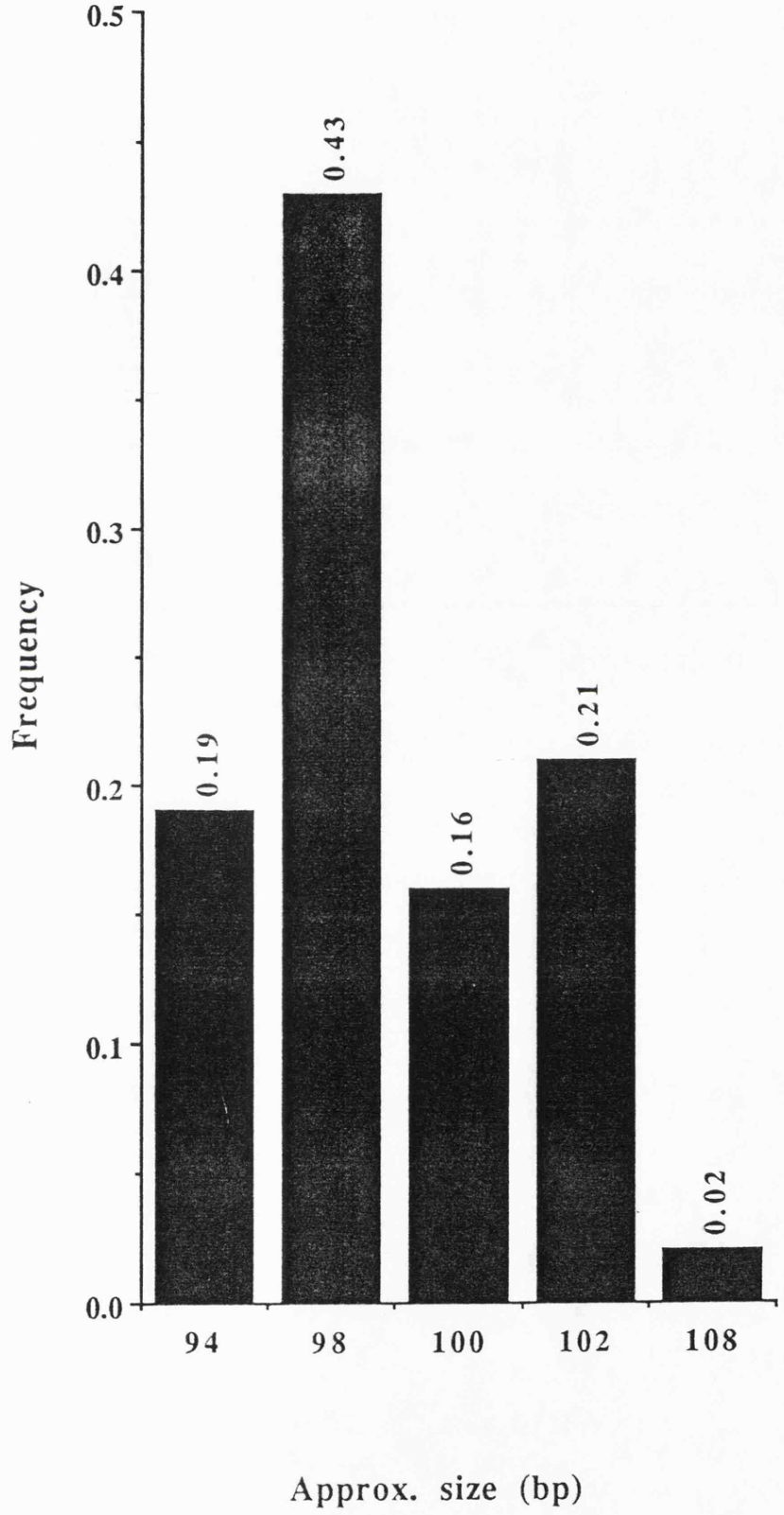

0.5
0.4
0.3
0.2
0.1
0.0
Frequency
0.19
0.43
0.16
0.21
0.02
94
98
100
102
108
Approx. size (bp)

Figure 5.3. TG10 alleles amplified from chimpanzee and gorilla DNA. The TG10 locus was PCR amplified from chimpanzee and gorilla DNA and the products electrophoresed through a 4% agarose gel (3 parts nusieve, 1 part HGT) in TBE with ethidium bromide. TG10 is clearly polymorphic in these species, but with somewhat larger alleles in the gorilla. A third band, migrating behind the two alleles, is evident in heterozygotes; this band is the result of formation of a heteroduplex of the two alleles.

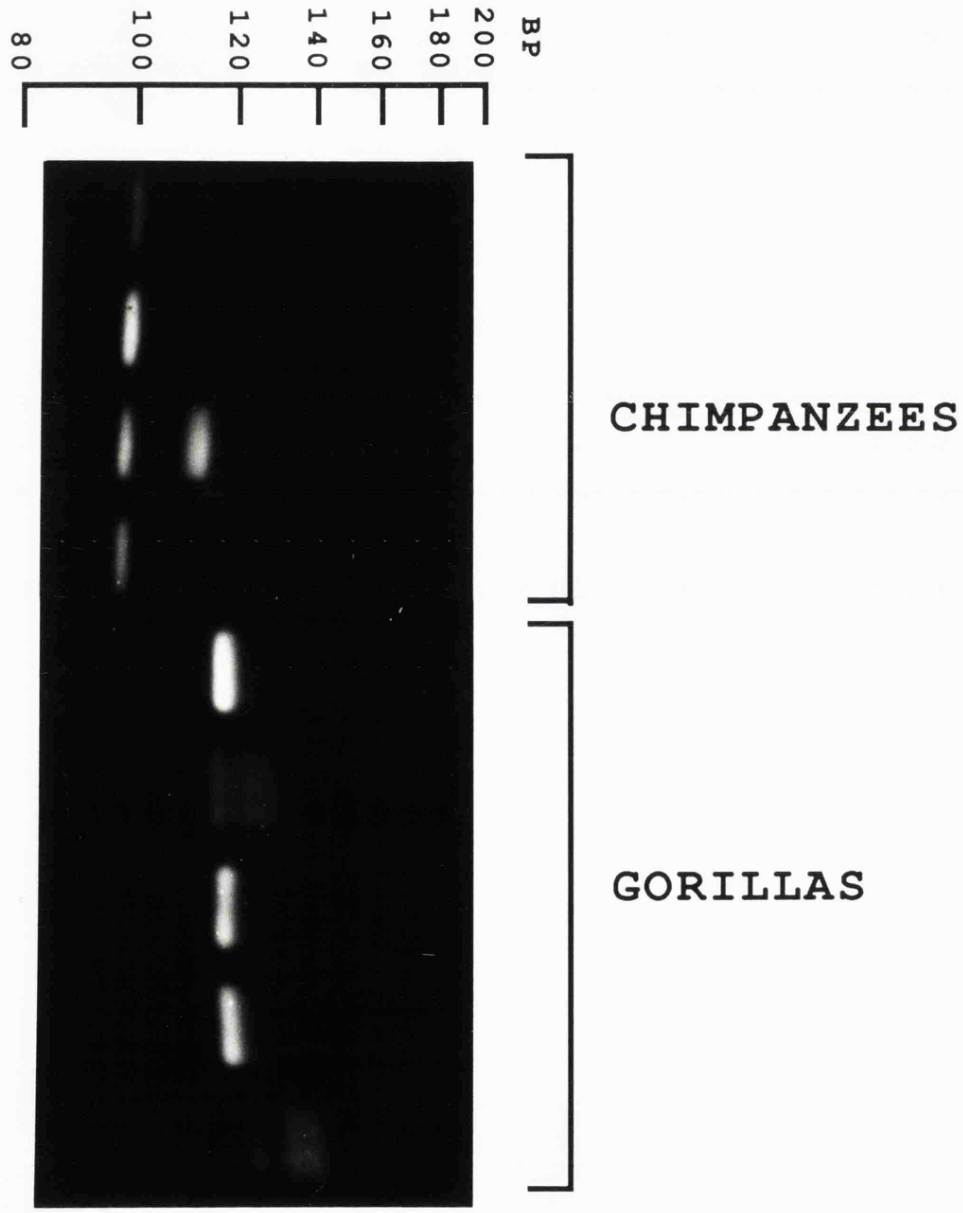
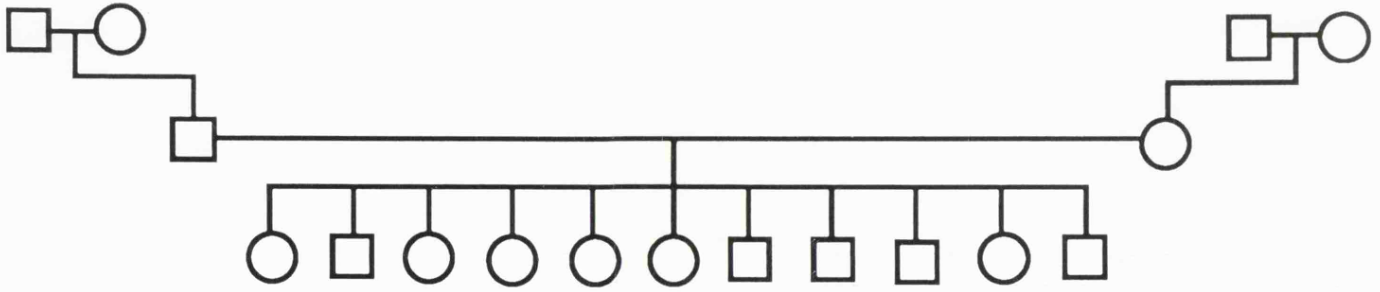
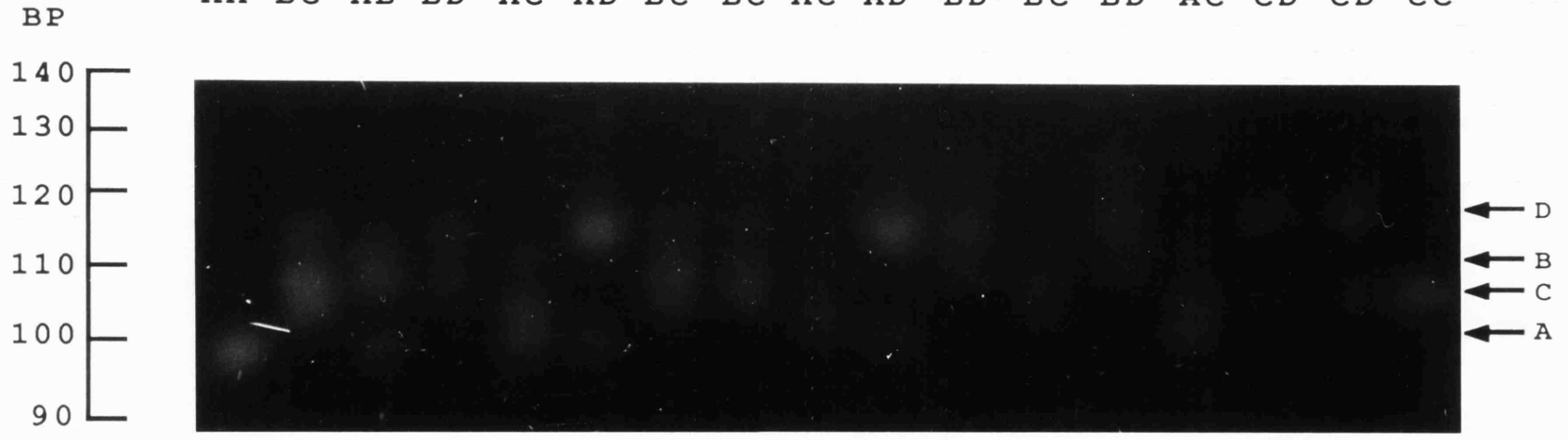

80
100
120
140
160
180
200
BP
CHIMPANZEES
GORILLAS

Figure 5.4. A. Pedigree analysis of TG10. The inheritance of TG10 alleles through 3 generations of a large CEPH family, as determined by agarose gel electrophoresis (3% nusieve, 1% HGT in TBE with ethidium bromide). The genotype of each individual is given above the appropriate lane.
B. A diagrammatic representation of the inheritance pattern, with heteroduplexes shown as dotted lines.

A.



AA BC AB BD AC AD BC BC AC AD BD BC BD AC CD CD CC



B.

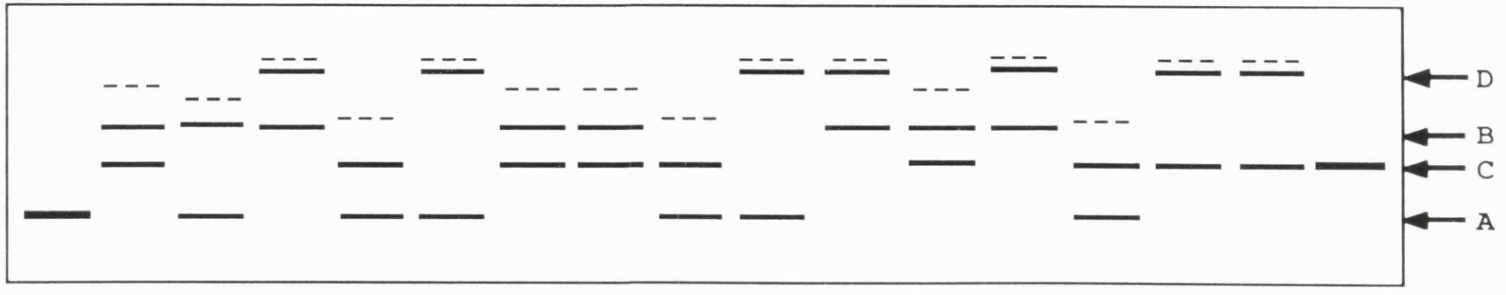


Figure 5.5. Variability at TG29. TG29 alleles were PCR amplified from 24 unrelated individuals and electrophoresed through a 4% agarose gel (3 parts nusieve, 1 part HGT) in TBE with ethidium bromide. The lower allele in this dimorphic system is 132bp whereas the upper is approximately 140bp. The respective allele frequencies are 0.81 and 0.19. The third band present in heterozygotes is due to heteroduplex formation.

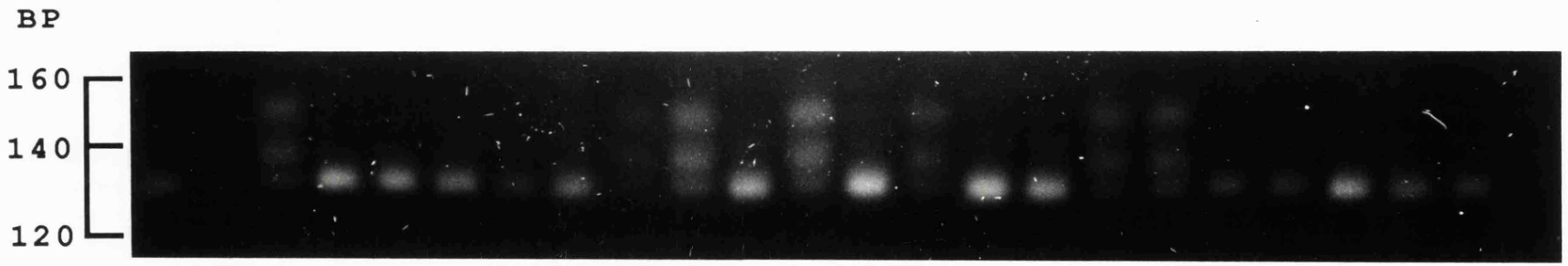
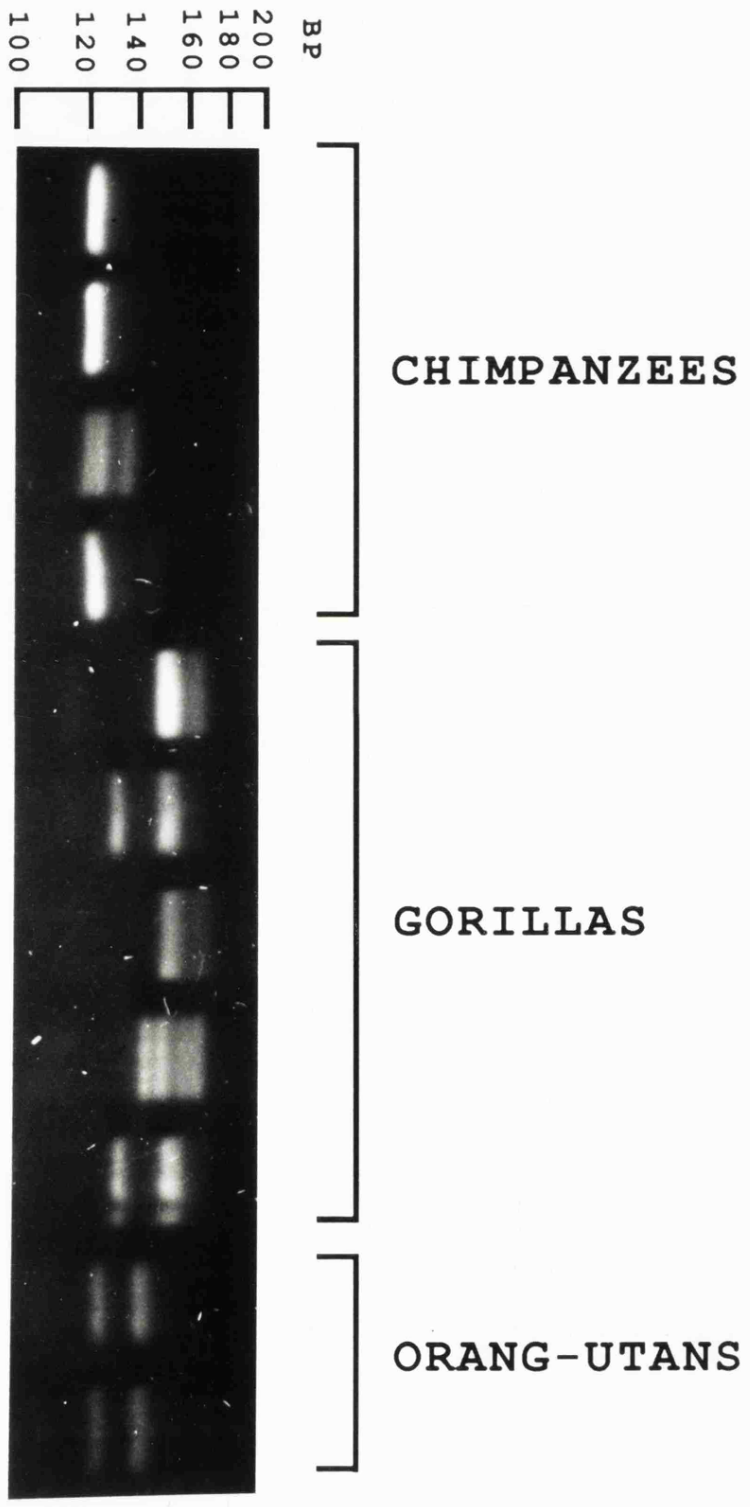

BP
160
140
120

Figure 5.6. PCR amplified TG29 alleles from great ape DNA. TG29 alleles were amplified from chimpanzee, gorilla and orang-utan DNAs and the products electrophoresed through a 4% agarose gel (3 parts nusieve, 1 part HGT). In sharp contrast to the human locus, TG29 appears to be highly variable in these species. Again, heteroduplexes are evident in heterozygotes.


BP
200
180
160
140
120
100
CHIMPANZEES
GORILLAS
ORANG-UTANS

5.6 Analysis of TG9

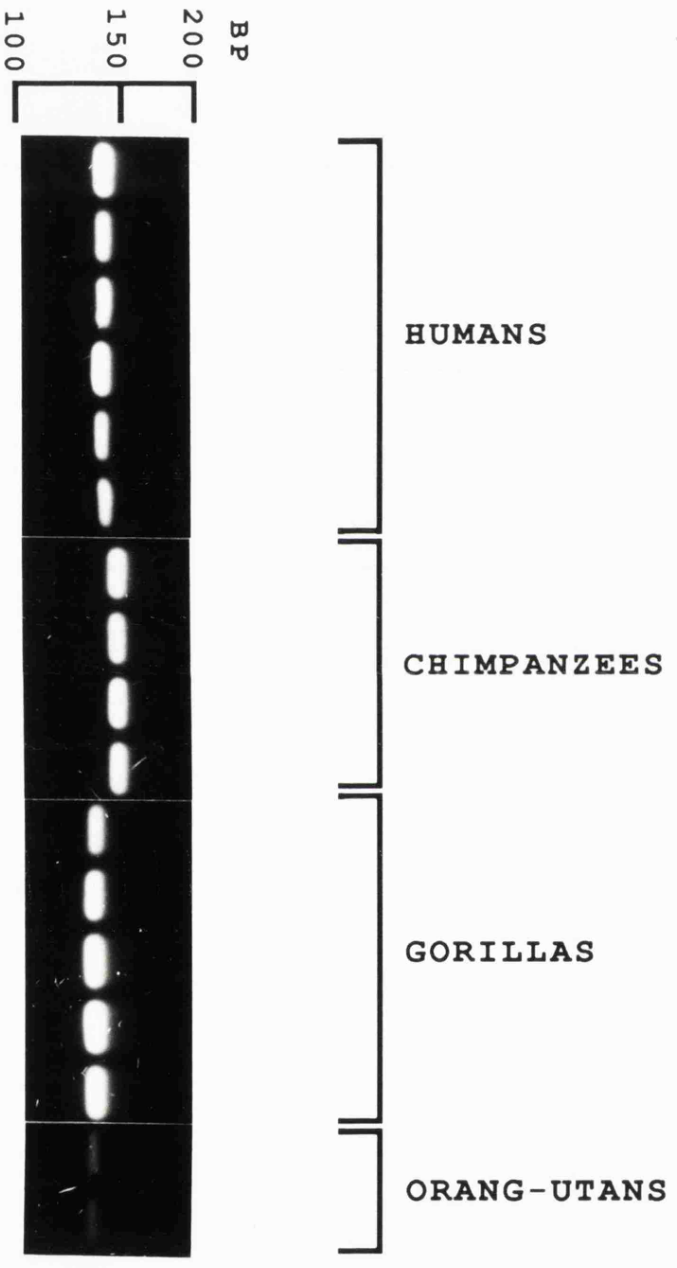
Microsatellite TG 9 consists of a 39bp $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ array with two transversional disruptions at positions 13 and 29 (figure 5.1). PCR amplification of TG 9 from 6 humans, 4 chimpanzees, 5 gorillas and 2 orang-utans showed it to apparently be monomorphic within all four species, but with size variation between species other than gorilla and orang-utan, which both gave a product of the same size (figure 5.7).

5.7 Discussion

It appears that perfect $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ tracts are limited to a maximum length of approximately 30 dinucleotide repeats. This upper limit is possibly due to instability of larger arrays, which may show a bias in favour of collapse rather than further expansion due to an increased level of events such as intrastrand deletion (Walsh, 1987). Alternatively, larger dinucleotide repeat arrays may present a larger target for base substitutional mutation events. If the base substitution rate is greater than the rate of array homogenization, through mechanisms such as DNA slippage during replication, the microsatellite will eventually be lost; swamped by base substitutions. Therefore, if the rate of homogenization is not proportional to array length, short microsatellites will be favoured and larger microsatellites will be lost. In the absence of further studies of mutational processes operating at these loci this remains speculation; however, the existence of long cryptic dinucleotide repeats such as TG6 (figure 5.1) would suggest accumulation of base substitutions in long microsatellites.

The evidence for clustering of simple repeat sequences (two simple repeat arrays found within 200bp of each other in 2/15 clones; figure 5.1) implies that certain regions of the genome are prone to the formation of such sequences. Clustering of simple repeated blocks has also been noted elsewhere (Tautz & Renz, 1984b), and has been documented for more complex minisatellite repeat arrays (Armour *et al.*, 1989). Indeed, minisatellites and microsatellites have sometimes been found in close association with each other (J. Armour and M. Gibbs, personal communication).

Figure 5.7. TG9 alleles, PCR amplified from human and great ape DNA. TG9 was amplified from human, chimpanzee, gorilla and orang-utan DNAs and the resulting products electrophoresed through a 4% agarose gel (3 parts nusieve, 1 part HGT) in TBE with ethidium bromide. Although apparently monomorphic within each species, this locus shows clear inter-species variation, with the exception of gorilla and orang-utan, which appear to share the same sized allele.


BP
100
150
200
HUMANS
CHIMPANZEES
GORILLAS
ORANG-UTANS

Unfortunately, published methods for typing microsatellites are cumbersome, involving end-labelling PCR primers with γ-32P-dATP and resolving the resulting radioactive PCR products on denaturing sequencing-type polyacrylamide gels (Litt & Luty, 1989; Weber & May 1989; Tautz, 1990). This appears to be largely unnecessary, as high percentage Nusieve agarose gels offer an acceptable degree of resolution for assaying the degree of variabilty of microsatellites, and for linkage analysis. However, for purposes of individual identification in, for example, a forensic context, absolute allele sizes and frequencies within the population are required, rendering denaturing polyacrylamide gels unavoidable (see chapter 6).

The informativeness of TG10 in man, with a heterozygosity of 74% and 5 alleles with varying frequencies in the population is typical of microsatellites (Weber *et al.*, 1990). This compares rather unfavourably with the most informative minisatellite loci, which have allele length heterozygosities in excess of 90% and may have more than 40 different length alleles which show a fairly smooth distribution in the population (Wong *et al.*, 1987). However, the high copy number of $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ repeat blocks in eukaryotes (Miesfeld *et al.*, 1981, Hamada & Kakunaga, 1982, Hamada *et al.*, 1982, Hamada *et al.*, 1984a) and their apparent random distribution throughout the genome (Luty *et al.*, 1990) renders them extremely promising markers for linkage analysis and diagnosis of genetic disease (see for example Feener *et al.*, 1990; Richards *et al.*, 1991).

The level of variability of TG10 in the great apes appears to be comparable to that in man, albeit with different allele frequency distributions, implying that this locus has shown a similar degree of variability for at least the last 7 million years, the estimated divergence time of man, chimpanzee and gorilla (Koop *et al.*, 1986). Such persistence is expected of tandem arrays which show a modest level of variability (and presumably reflects a low mutation rate), and is in stark contrast to minisatellites MS1 and MS32, which are highly variable only on the human lineage, and show extreme evolutionary transience (see chapter 3).

The mutational processes by which microsatellites change length to generate and maintain polymorphism is unknown. However, a recent

study of a microsatellite locus within the cystic fibrosis transmembrane conductance regulator (CFTR) gene has shown strong linkage disequilibrium between a group of microsatellite alleles and a haplotype of flanking markers on either side, implying that these alleles are all derived from the same precursor and arose through intra-allelic events such as replication slippage rather than inter-allelic recombination mechanisms (Morral *et al.*, 1991). The low level of variability of TG29 in man compared with an apparently high level in the great apes would suggest that the base substitution disrupting the array in man is preventing generation of new alleles through length change. This is in agreement with studies by Weber (1990) which demonstrate that informativeness of interrupted dinucleotide sequences is significantly lower than that of perfect dinucleotide repeats, and that the degree of informativeness of a given locus is reflected in the number of uninterrupted repeats. By this criterion, TG29, with a run of 10 perfect repeats, would be predicted to have very low or zero informativeness. This lends further evidence to the argument for strand slippage during replication as a mechanism for length change at these loci. The rate of strand slippage has been shown to increase with increasing length for $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ sequences in M13 (Levinson & Gutman, 1987b) and it would appear that the rate of slippage at these sequences jumps as the number of repeats exceeds 10 in man (Weber, 1990). Similarly the apparently monomorphic nature of the imperfect repeat TG9 in man and all species of great ape implies blockage of the slippage mechanism through base substitution. However, the variation between species suggests that length change does occur with 10 or fewer perfect repeats, but at a much lower rate. Such slowly evolving loci may be useful for phylogenetic analysis at the species level (Burke *et al.*, 1991).

CHAPTER 6

IDENTIFICATION OF THE SKELETAL REMAINS OF A MURDER VICTIM USING MICROSATELLITE MARKERS

6.1 Introduction

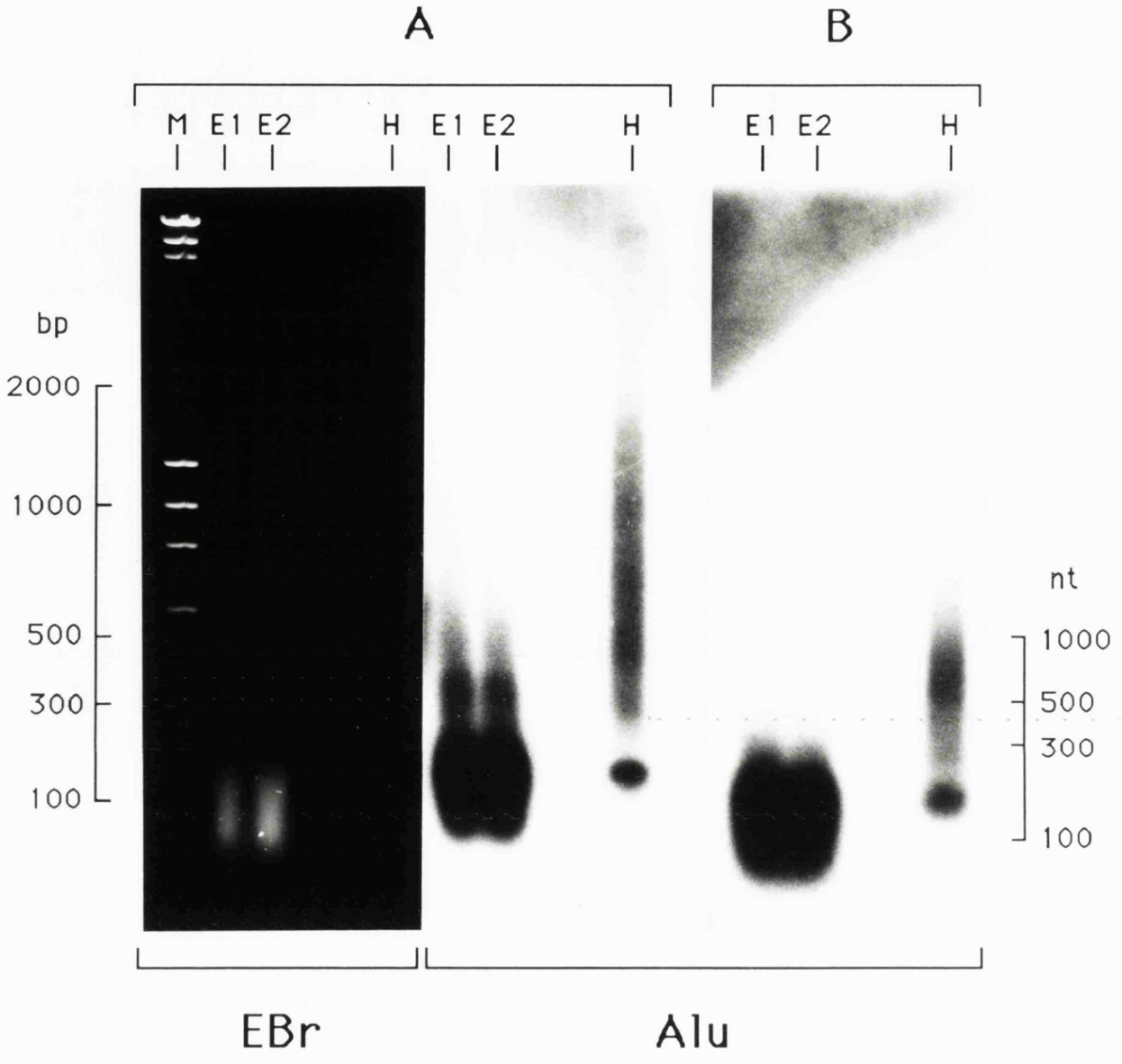
In 1981, a 15-year old Caucasian girl was murdered and her body wrapped in a carpet and buried. The skeletal remains were discovered by chance in 1989. A facial reconstruction allowed the identity of the victim to be provisionally established, and appeared to be confirmed by dental records. To obtain further evidence of identification, bone DNA analysis was attempted.

6.2 Quantity and quality of DNA extracted from the remains

DNA extractions were performed by Dr. Erika Hagelberg of the Institute of Molecular Medicine, University of Oxford, as follows. A 5 gram fragment of femur from the victim was sand-blasted to remove the surface 1-2mm of bone, in an attempt to remove any contaminant DNA introduced by handling, and divided in two. Each sample was pulverized and DNA extracted as described in Hagelberg & Clegg (1991).

To determine the quantity and quality of DNA recovered, aliquots of each bone DNA sample were run through an agarose gel, alongside a known amount of human genomic DNA cut with *Sau*3AI. Native and denatured samples were run, to assess the degree of nicking. Total DNA was visualized by staining with ethidium bromide, and the fraction of the DNA that was of human origin assessed by Southern blot hybridization with a human Alu repeat probe (figure 6.1), prepared by denaturation of human DNA, followed by reannealing to low C_0t and S_1 nuclease digestion (Houck *et al.*, 1979; probe prepared by Mrs. Vicky Wilson). Use of a cloned human Alu probe was avoided, due to the potential risk of contaminating *E. coli* host DNA cross-hybridizing to bacterial DNA in the remains. Scanning laser densitometry and estimation by eye showed that approximately 5μg of DNA were recovered from each extract, but that only about 10% of the DNA was of human origin. It is likely that the remaining 90% was derived from bacterial or fungal contamination of the remains. The human DNA fraction was highly degraded, with virtually all single-stranded DNA fragments smaller than 300 nucleotides.

Figure 6.1. Assessment of the quality and quantity of human DNA recovered from the femur DNA extracts from the victim. 1/5 of each femur extract (E1, E2) was electrophoresed through a 1.5% agarose gel alongside 20ng human genomic DNA digested with *Sau*3A1 (H). DNA was electrophoresed in either the native state (A) or denatured with 0.15M NaOH, 10mM EDTA prior to electrophoresis (B). Total DNA was visualized by staining with ethidium bromide (EBr) and human DNA detected by Southern blot hybridization with a ^{32}P-labelled consensus human Alu probe at low stringency, following transfer to a nylon membrane (Amersham Hybond-N).

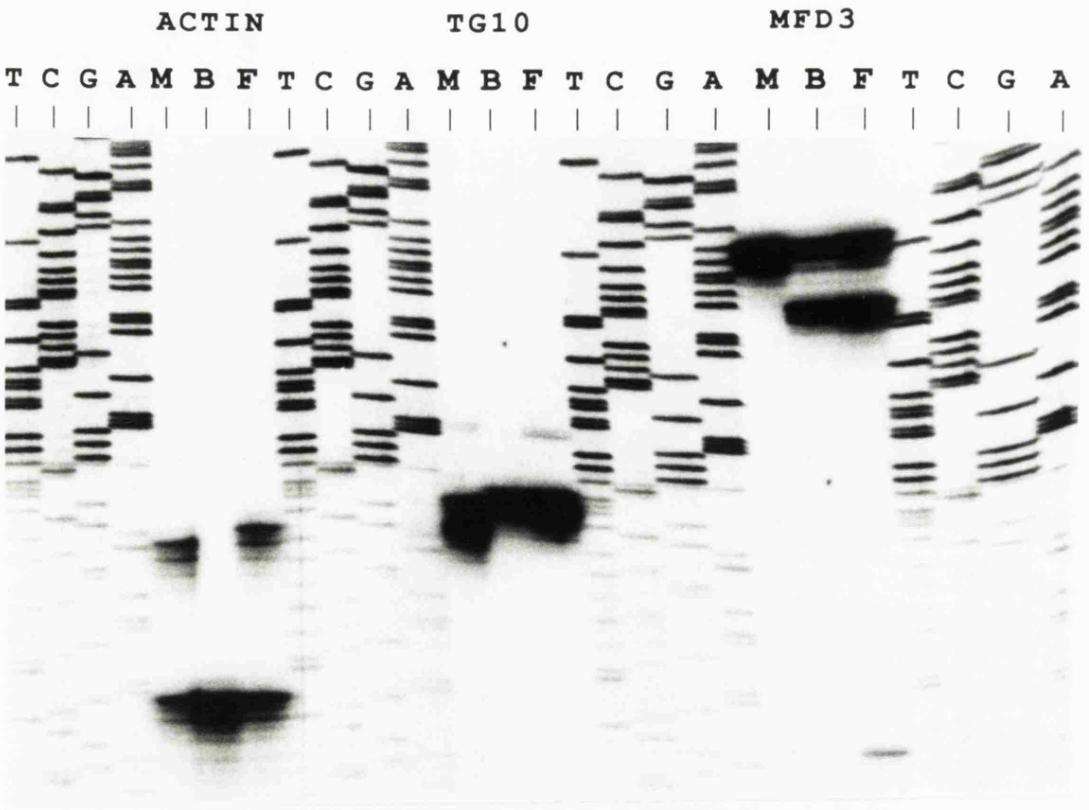

A
B
M E1 E2 H E1 E2 H
E1 E2 H
bp
2000
1000
500
300
100
nt
1000
500
300
100
EBr
Alu

6.3 Identification of the remains

The severe degradation of the DNA recovered rendered DNA typing by Southern blot hybridization with minisatellite markers (Wong *et al.*, 1987) or amplification of minisatellites by PCR (Jeffreys *et al.*, 1988) impractical. Consequently the bone extracts were typed by PCR amplification of much shorter (<200bp) $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ microsatellite loci. Six different loci (TG10, see chapter 5; actin, Litt & Luty, 1989; Mfd3, Mfd5, Weber & May, 1989; Mfd49, Weber *et al.*, 1990a and Mfd64, Weber *et al.*, 1990b) were successfully amplified from the bone DNA samples by Professor Alec Jeffreys and typed on high percentage agarose gels, revealing the victim to be heterozygous at five of the six loci. Amplifications were performed in a laminar flow hood using pipettes, disposable plasticware and reagents that had not been previously exposed to the laboratory enviroment (as described in Jeffreys *et al.*, 1990), to minimize the risk of contamination.

Following the demonstration that microsatellite alleles could be successfully amplified from the bone DNA, identification was attempted by comparing the microsatellite profiles of the victim with those of the presumptive parents, following preparation of the parental DNA from blood samples as described in Jeffreys *et al.*, 1990. Profiles were compared by electrophoresis through high percentage agarose gels and subsequently by typing on denaturing polyacrylamide sequencing-type gels, following re-amplification of PCR products in the presence of a ^{32}P end-labelled primer (figures 6.2 and 6.3). Both methods gave concordant results, but agarose gels gave simpler profiles (figure 6.3). This is thought to be due to resolution of spurious products on polyacrylamide gels, arising from polymerase slippage at CA repeats during amplification (Litt & Luty, 1989) and through non-templated nucleotide addition catalysed by Taq polymerase (Clark, 1988). The results were fully consistent with the femur DNA being derived from an offspring of the presumptive parents; in all cases the skeletal microsatellite alleles could be attributed to either the mother or the father.

Figure 6.2. Comparison of skeletal microsatellite profiles with those of the presumptive parents. For each locus, a 10ng aliquot of DNA was initially amplified for 33 cycles in a 30µl reaction as described in table 2.2. 1µl aliquots of PCR product were then re-amplified for 4 cycles in a 10µl PCR reaction with 1/10 of one of the amplimers ^{32}P-labelled using T4 polynucleotide kinase and γ-^{32}P-dATP. Labelled products were denatured, electrophoresed through a DNA sequencing-type polyacrylamide gel alongside an M13mp18 sequence ladder (TCGA) and visualized by autoradiography. For each locus, the genotype of the bones (B) is compatible with the DNA being from an offspring of the presumptive mother and father (M and F). Allele lengths in table 6.1 were determined from the major PCR product of each allele.


ACTIN
TG10
MFD3
T C G A M B F T C G A M B F T C G A M B F T C G A

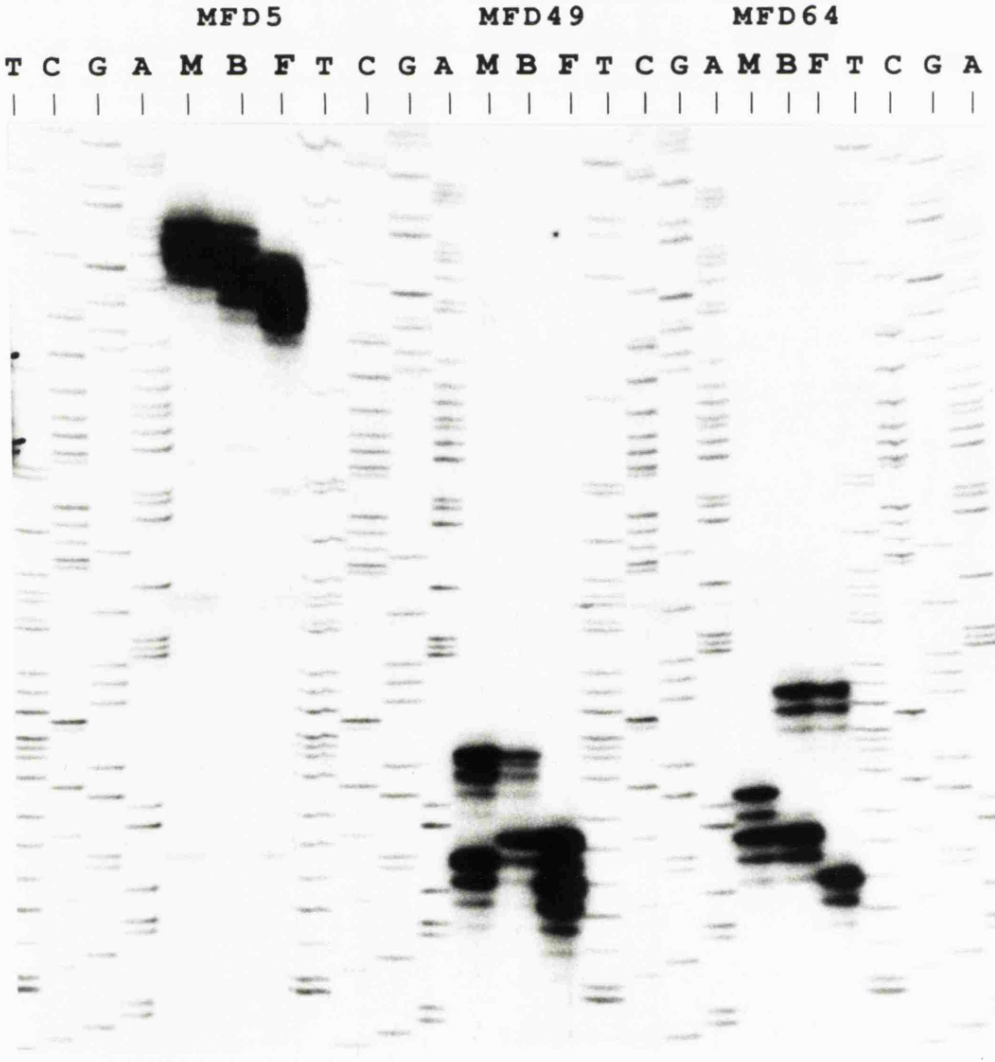
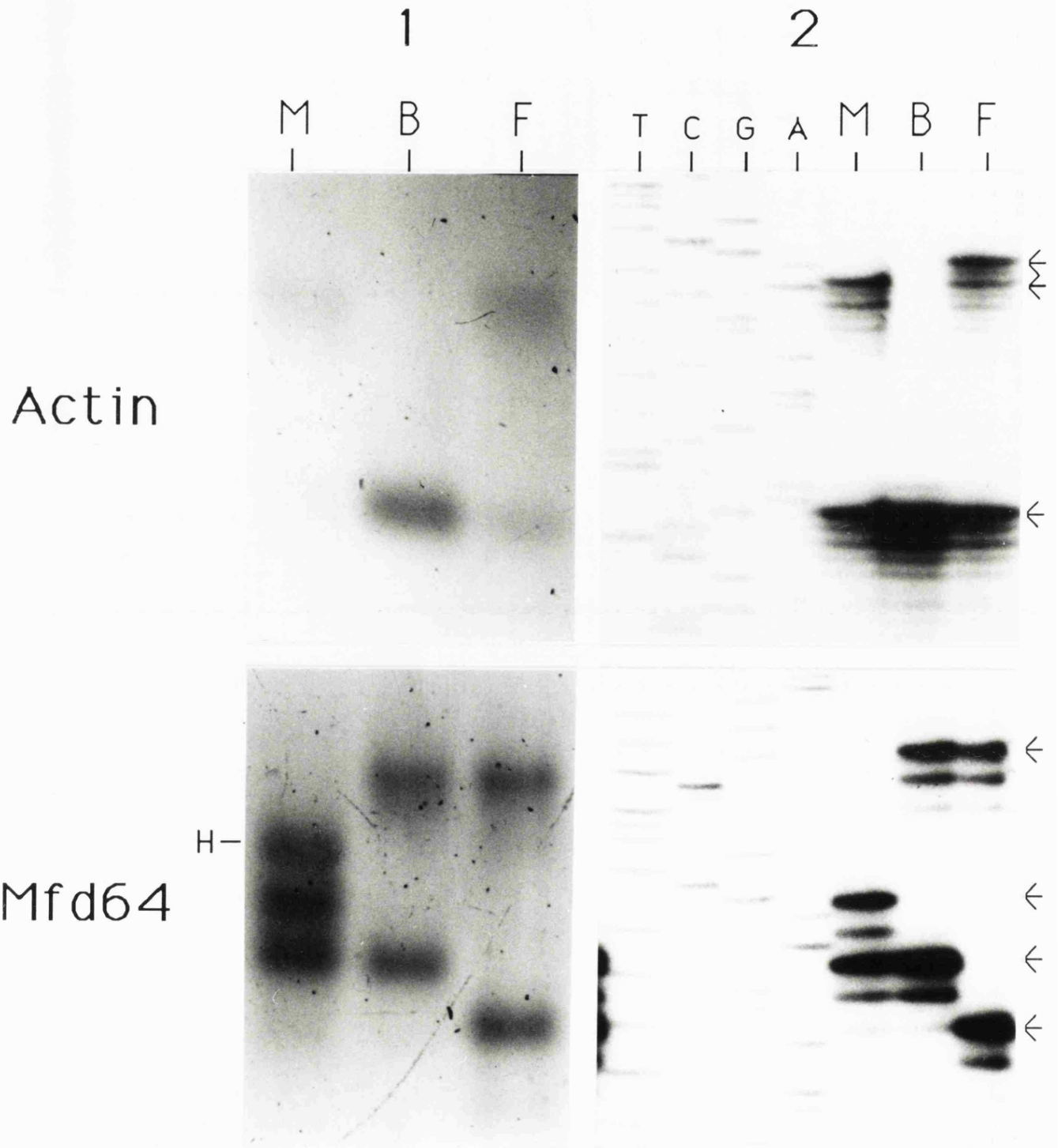

MFD5
MFD49
MFD64
T C G A M B F T C G A M B F T C G A M B F T C G A

Figure 6.3. Comparison of microsatellite typing on high percentage agarose gels and sequencing-type polyacrylamide gels. Alleles at two loci (actin, Mfd64) were PCR amplified and typed either native on a 20cm 4% agarose gel (3 parts nusieve, 1 part HGT) in TBE and detected by ethidium bromide staining, or denatured on a sequencing-type gel following labelling with ^{32}P, as in figure 6.2. Both systems gave concordant results. H = heteroduplex.


1
2
M
B
F
T
C
G
A
M
B
F
Actin
H–
Mfd64

6.4 Statistical evaluation of the positive identification (all statistical analyses performed by Professor Alec Jeffreys)

Assuming Hardy-Weinberg equilibrium at each microsatellite locus and statistical independence of these loci (assumptions which appear to be valid for simple tandem repeat sequences investigated to date - see Chakraborty *et al.*, 1991), the statistical weight of the evidence linking the skeletal remains to the presumptive parents M and F was evaluated using published Caucasian allele frequencies (table 6.1). For each locus, the probability of obtaining the M, F and bone genotypes if the skeletal remains were those of an offspring of M and F was calculated, and conversely the probability of obtaining the precise combination of alleles seen in the presumptive parents and the victim if the skeletal DNA was not derived from an offspring of M and F was determined. If both parents are heterozygous, this likelihood ratio is given by $1/(8q_1q_2)$, where q_1 and q_2 are the population frequencies of the two alleles in heterozygous bone DNA. If one or both parents are homozygous, this ratio becomes $1/(4q_1q_2)$ or $1/(2q_1q_2)$ respectively. If the bone DNA is homozygous for an allele of frequency q the three likelihood ratios are $1/(4q^2)$, $1/(2q^2)$ and $1/(q^2)$ respectively. This ratio in favour of identification varies considerably between loci according to the population frequencies of the femur alleles (table 6.1).

The cumulative likelihood ratio for all six loci is very high ($2x10^8$). However, the population databases for allele frequencies at these loci are currently small, and to compensate for sampling errors in allele frequency estimates, the likelihood ratios were re-evaluated after adjusting all allele frequencies to their upper 95% confidence limits (table 6.1). The resulting cumulative likelihood ratio remains high at $2x10^5$, although it should be noted that major allele frequency differences between the populations represented in the published databases and in the local population relevant in this case (South Wales) could perturb this estimate. However, the data establish with a high degree of confidence that the victim was indeed the daughter of M and F. This PCR data was presented in evidence at a murder trial at Cardiff Crown Court in February 1991, which led to the conviction of the murderer and his accomplice. This was the first occasion where PCR data has been given in evidence at a British court.

Table 6.1. Statistical evaluation of the microsatellite evidence linking the skeletal remains to the presumptive parents. The statistical weight of the evidence linking the skeletal remains to the presumptive parents M and F was evaluated as described in the text. The likelihood ratio in favour of identification varies substantially from locus to locus according to the population frequencies of the femur alleles. The cumulative likelihood ratio for all six loci is very high ($2x10^8$). However, the population databases for allele frequencies at these loci are currently relatively small, and to compensate for sampling errors in allele frequency estimates, the likelihood ratios were re-evaluated after adjusting all allele frequencies to their upper 95% confidence limits. The cumulative likelihood ratio remains high at $2x10^5$. References: (1) Litt & Luty, 1989; (2) chapter 5; (3) Weber & May, 1989; (4) Weber *et al*, 1990a; (5) Weber *et al*, 1990b.

Microsatellite (reference in brackets)	Chromosome	Allele sizes in femur (nt)	Total no. alleles in database	Femur allele frequency	Upper 95% limit of allele frequency	Likelihood ratio offspring: unrelated	Likelihood ratio (95% frequency limits)
Actin (1)	15	72	74	0.11	0.18	21	8
TG10 (2)	18	100 98	58	0.16 0.43	0.25 0.54	2	1
Mfd3 (3)	1	141 131	82	0.35 0.14	0.45 0.22	5	3
Mfd5 (3)	19	153 149	150	0.15 0.03	0.21 0.07	100	34
Mfd49 (4)	15	96 88	120	0.02 0.13	0.05 0.19	58	13
Mfd64 (5)	1	102 88	118	0.01 0.08	0.04 0.13	184	24

6.5 Discussion

There is considerable scientific and public interest in the possibility of DNA typing of ancient remains. The first analysis of ancient DNA resulted from the successful cloning of mitochondrial DNA from a museum specimen of a quagga, a species of horse extinct since 1883 (Higuchi *et al.*, 1984). Subsequently a human Alu DNA sequence was cloned from an Egyptian mummy (Paabo, 1985). With the advent of PCR, the field of ancient DNA analysis has expanded dramatically. Multicopy human mitochondrial DNA sequences have been amplified from extracts retrieved from bones as old as 5500 years (Hagelberg *et al.*, 1989; Hagelberg & Clegg, 1991), and from 7000-year old brain tissue preserved in a peat bog (Paabo *et al.*, 1988). There has even been one report of amplification of chloroplast DNA from a 17-20 million year old *Magnolia* leaf (Golenberg *et al.*, 1990).

However, such is the sensitivity of PCR that it is difficult to exclude the possibility that amplification products result from contaminating material, and not from the tissues themselves. The analysis described here provides very strong evidence that in this case the human DNA retrieved from the bone was that of the skeleton and did not arise through contamination. However, had the profile of the victim not matched that of the presumptive parents, it would be difficult to deem the exclusion authentic given the possibility of inadvertent contamination of the bone DNA, although this problem may be overcome by multiple testing of independent extracts of different bones.

Skeletal remains have previously been identified following bone DNA extraction (Stoneking *et al.*, 1991) but this identification was based on sequence polymorphisms within mitochondrial DNA, which are of limited informativeness; consequently the probability of a false inclusion is relatively high. The analysis presented here provides the first example of the successful identification of skeletal remains using nuclear DNA markers. In addition to the identification of the remains of missing persons, the typing of ancient and degraded forensic DNA samples may have applications in retrospective analysis following potentially unjust convictions, and in the identification of individuals buried in mass graves.

Verification of family relationships in instances of contested wills, following exhumation of the Testate, also becomes a possibility.

More broadly, It may be possible to apply ancient DNA typing to anthropological and phylogenetic questions. However, this will depend heavily on the degree of DNA preservation in archaeological and palaeontological specimens, a criterion which appears to be dictated by burial conditions rather than age for specimens up to a few thousand years old (Hagelberg *et al.*, 1989), although fossil specimens may be expected to contain DNA fragments only a few nucleotides in length, following spontaneous depurination and subsequent phosphodiester bond cleavage (Lindahl & Anderson, 1972; Lindahl & Nyberg, 1972). However, the report of retrieval of 770bp chloroplast DNA sequences from a Miocine *Magnolia* leaf (Golenberg *et al.*, 1990) suggests that DNA may undergo only limited degradation under the appropriate conditions, although others have failed in attempts to repeat these experiments (Sidow *et al.*, 1991). As yet, the vast majority of work on ancient DNA has concentrated on organelle (mitochondrial or chloroplast) genomes, which are present as multiple copies per cell. Whether such analysis can be extended to nuclear DNA remains to be seen, although it would appear to be a viable proposition in at least some cases: 125bp DNA fragments derived from HLA genes have recently been PCR amplified from 7500 year-old human brain tissue preserved in a Florida peat bog (Lawlor *et al.*, 1991).

Practical future applications of ancient DNA analysis should include the establishment of phylogenetic affiliations in anthropological studies, possibly shedding light on areas such as human evolution, the divergence of human races and the origins of contemporary human population structure. The establishment of phylogenetic associations between ancient and modern Japanese populations based on comparisons of mitochondrial DNA sequences from archaeological (200-6000 year-old) specimens and contemporary populations has already been attempted (Horai *et al.*, 1991). Of course, such analysis is not limited to man; phylogenetic affiliation of the extinct marsupial wolf (*Thylacinus*) to Australian rather than South American marsupials has been established from mitochondrial DNA sequences following PCR amplification of DNA from Victorian and early 20th century museum specimens (Thomas *et al.*, 1989). Ultimately,

it may even be possible to study evolutionary processes directly at the DNA level if an evolutionary series of intermediate stages between ancient and modern taxa were available (see Dover's comments in Golenberg, 1991).

CHAPTER 7

POLYMORPHISM AT OTHER SIMPLE TANDEM REPEAT SEQUENCES

7.1 Introduction

Due to the limited length and variability of $(dT\text{-}dG)_n\cdot(dA\text{-}dC)_n$ microsatellites (Weber, 1990; see chapters 5 and 6), it seemed prudent to search for tetranucleotide repeats, in the hope that these sequences would be longer and more variable than dinucleotide repeats, but still remain short enough to be easily typed by PCR. Variability at such sequences was likely, as short trinucleotide sequences in *Drosophila* had already been shown to be variable (Tautz, 1989), and a highly variable mouse minisatellite consisting of the tetranucleotide repeat $(GGCA)_n$ had been characterized (Kelly *et al.*, 1991; Mark Gibbs, personal communication). Many short tetranucleotide repeat arrays which show length variation on PCR amplification have now been isolated (For example see Mercier *et al.*, 1991).

7.2 Screening an M13 genomic library for tetranucleotide repeat sequences

High molecular weight DNA was pooled from 16 unrelated individuals and digested to completion with restriction endonucleases *Alu*I, *Rsa*I and *Hae*III.A 400-1000bp size fraction was selected by gel purification and cloned into the *Sma*I site of M13mp18 (Yanisch-Perron *et al.*, 1985) to generate a library of approximately 10,000 plaques (see chapter 5). Following plaque lifts onto nitrocellulose, this library was screened with a $(GGCA)_n$ repeat probe derived from the mouse minisatellite *Hm2* (Mark Gibbs, personal communication). Two strongly positive clones were detected, and were designated GGCA.1 and GGCA.2. GGCA.2 was later found to contain an insert consisting of a 'midisatellite' repeat sequence, and its characterization is described in chapter 8. Sequence analysis revealed GGCA.1 to consist of approximately 100bp of imperfect $(GGAA)_n$ repeat sequence followed by 9 $(GCAA)_n$ repeats (figure 7.1). Unfortunately, this repeat sequence was immediately preceded by a stretch of DNA extremely rich in A/T residues, which continued to the extreme 5' end of the insert, rendering PCR priming impractical.

The GGCA.1 insert was cut with *Mae*I and *Dra*I to release a 230bp fragment consisting of the repeated sequence (figure 7.1); this fragment was then used as a probe to re-screen the M13 library in the hope that the

Figure 7.1. The sequence of clone GGCA.1, isolated from an M13 library with a $(GGCA)_n$ repeat probe. GGCA.1 was subsequently trimmed with *Dra*I and *Mae*I and the central imperfect repeated motif used as a probe to screen the same library.

GGCA.1:


DraI
GCTTAAGAATAAATGTTAAAATCTAAATAATCCTATTTAAATTTAAAATATTCAATTTGTATTTAAAACCTTTGAAAGAG-
GGAGAGGGAGAGAGTGGCAGAGGGAAGGA(GGAA)2C(GGAA)4A(GGAA)2AGGAAAGAA(GGAA)3AGAA(GGAA)13-
(GGCA)9GGCTTACGGAACCAAACCTGATGGCTTTATTTAGACTAGAGTCTAGTGACACTTGAGAAGAATAATGTAATGT-
TAA
MaeI

$(GGAA)_n$ repeat motif would identify novel polymorphic loci which had not been detected by the pure $(GGCA)_n$ probe. GGCA.1 hybridized strongly to 26 new clones, designated GGAA.1-26, of which 11 were sequenced.

7.3 Sequence analysis of $(GGAA)_n$ repeat loci

Of the 11 clones sequenced, 7 had inserts with long (150-200bp) imperfect $(GGAA)_n$ repeats, each apparently derived from the expanded polyadenosine tail of an immediately adjacent Alu dispersed repeat element (figure 7.2). Within one such clone (GGAA.24) a $(GGAA)_n$ repeat of approximately 200bp was sandwiched between two Alu repeats, and presumably resulted from expansion of the polyadenosine tail of the first. Unfortunately the presence of the Alu element immediately upstream of all of these $(GGAA)_n$ repeats rendered design of a unique sequence 5' flanking PCR primer impossible. Clone GGAA.6 contained an insert with a particularly unusual structure which will be considered further in section 7.6.

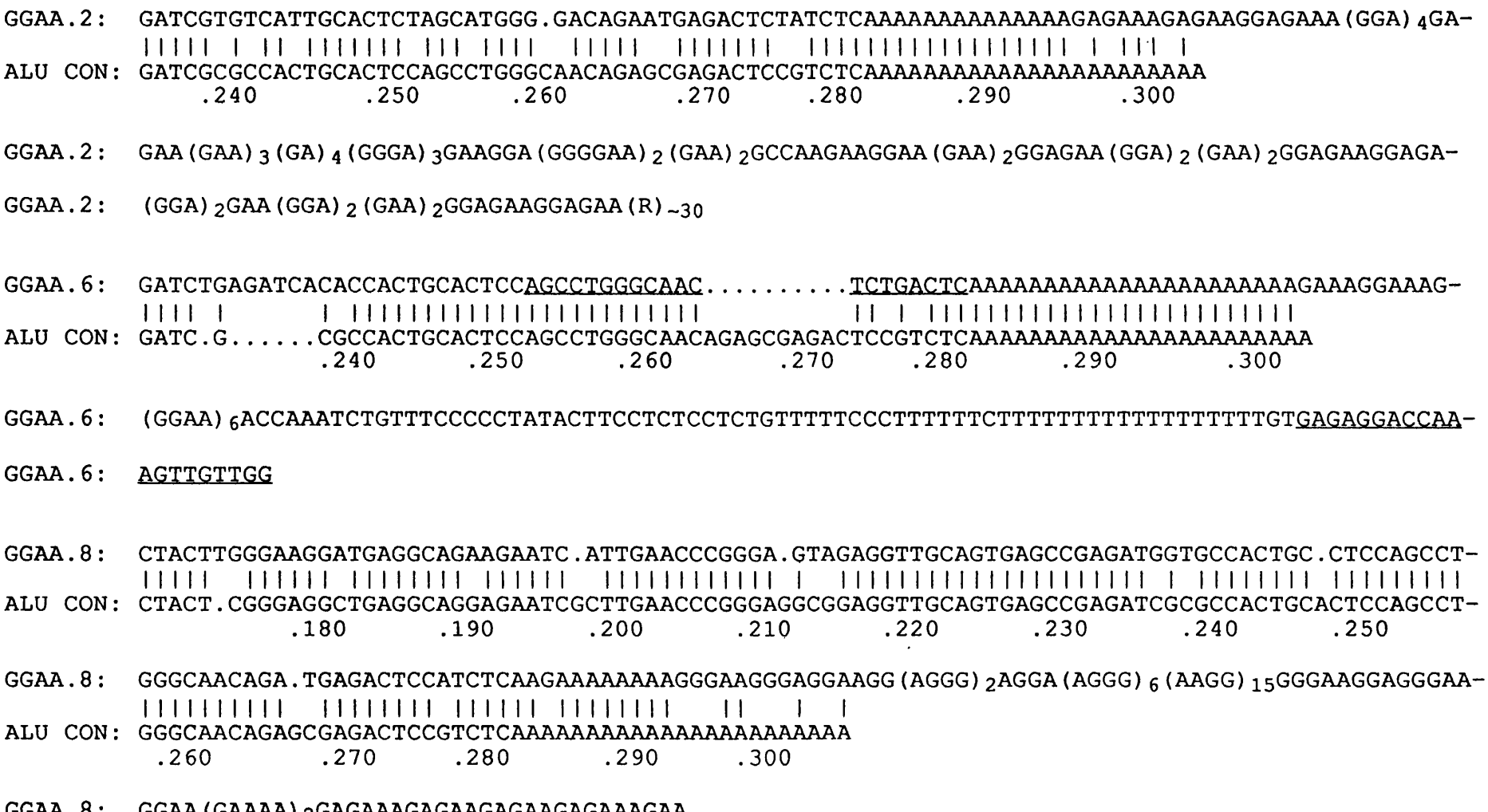
Of the remaining clones, GGAA.19 contained an insert with a considerable $(GGAA)_n$ imperfect repeat (approximately 300bp) which appeared to have expanded from an A/G-rich motif within a Kpn (L1) element (figure 7.3), again making design of unique sequence PCR primers impractical. However, two clones, GGAA.1 and GGAA.7 had inserts with imperfect $(GGAA)_n$ repeat arrays flanked by unique sequence, suitable for PCR priming (figure 7.4).

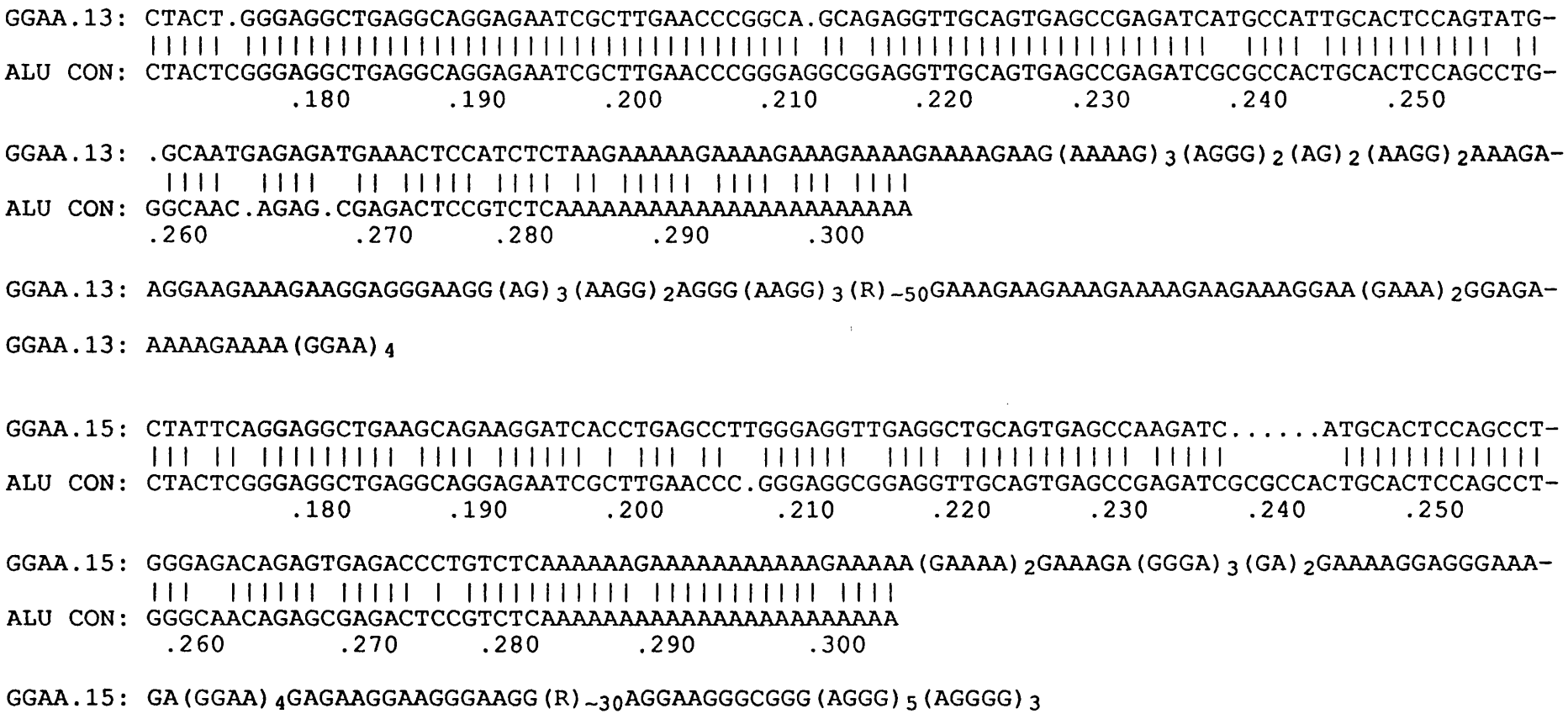
7.4 Characterization of two $(GGAA)_n$ repeat loci, GGAA.1 and GGAA.7

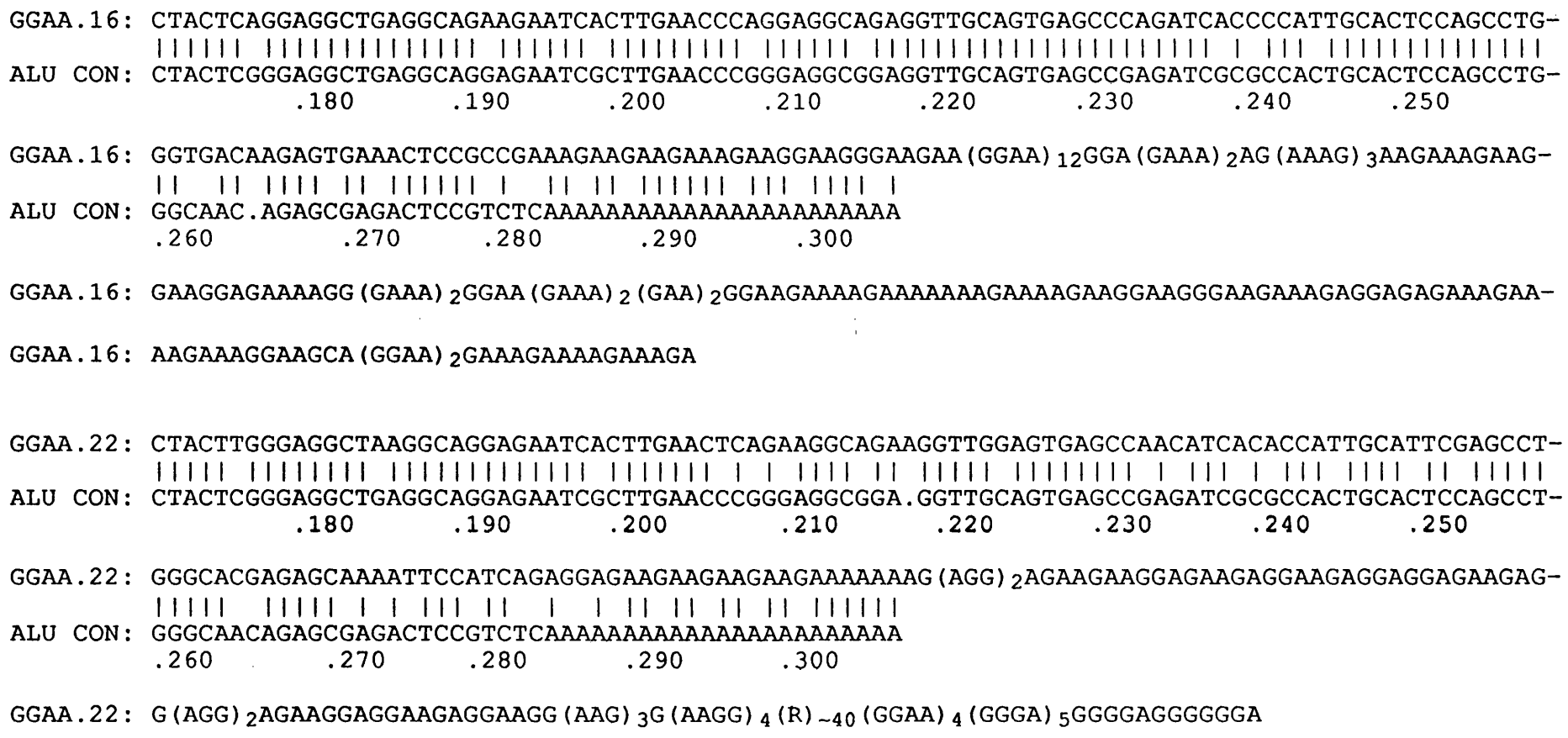
(a) GGAA.1

The tandem repeat array of the GGAA.1 locus was PCR amplified from the DNA of 24 unrelated individuals from the CEPH panel of large families. The resulting products were resolved on a 4% agarose gel (3:1 nusieve: Seachem HGT) run in 1 x TBE. This limited sample revealed 11 resolvable alleles ranging in size from approximately 285 to 560bp, with three apparent divisions of allele sizes: 285-295bp, 370-390bp and 450-560bp. No alleles which were significantly more common than the others were observed; the most common allele had a frequency of approximately

Figure 7.2. Expanded polyadenosine tails of alu repeats (continued over the following 3 pages). Seven of the eleven clones detected by GGCA.1 which were selected for sequencing contained imperfect $(GGAA)_n$ arrays derived from alu polyadenosine tails. Alu con is the alu consensus sequence (Bains, 1986). $(R)_n$ denotes runs of purine residues which could not be read clearly. Clone GGAA.6 has a particularly unusual structure which is described further in section 7.6. The 10bp deletion immediately prior to the polyadenosine tail allowed unique sequence flanking PCR primers (underlined) to be designed for this locus, and it was subsequently found to be variable.


GGAA.2: GATCGTGTCATTGCACTCTAGCATGGG.GACAGAATGAGACTCTATCTCAAAAAAAAAAAAAAAGAGAAAGAGAAGGAGAAA(GGA)4GA-
ALU CON: GATCGCGCCACTGCACTCCAGCCTGGGCAACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAA
.240 .250 .260 .270 .280 .290 .300
GGAA.2: GAA(GAA)3(GA)4(GGGA)3GAAGGA(GGGGAA)2(GAA)2GCCAAGAAGGAA(GAA)2GGAGAA(GGA)2(GAA)2GGAGAAGGAGA-
GGAA.2: (GGA)2GAA(GGA)2(GAA)2GGAGAAGGAGAA(R)~30
GGAA.6: GATCTGAGATCACACCACTGCACTCCAGCCTGGGCAAC.....TCTGACTCAAAAAAAAAAAAAAAAAAAAAAAGAAAGGAAAG-
ALU CON: GATC.G.....CGCCACTGCACTCCAGCCTGGGCAACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAA
.240 .250 .260 .270 .280 .290 .300
GGAA.6: (GGAA)6ACCAAATCTGTTTCCCCCTATACTTCCTCTCCTCTGTTTTTCCCTTTTTTCTTTTTTTTTTTTTTTTTTGTGAGAGGACCAA-
GGAA.6: AGTTGTTGG
GGAA.8: CTACTTGGGAAGGATGAGGCAGAAGAATC.ATTGAACCCGGGA.GTAGAGGTTGCAGTGAGCCGAGATGGTGCCACTGC.CTCCAGCCT-
ALU CON: CTACT.CGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCT-
.180 .190 .200 .210 .220 .230 .240 .250
GGAA.8: GGGCAACAGA.TGAGACTCCATCTCAAGAAAAAAAAGGGAAGGGAGGAAGG(AGGG)2AGGA(AGGG)6(AAGG)15GGGAAGGAGGGAA-
ALU CON: GGGCAACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAA
.260 .270 .280 .290 .300
GGAA.8: GGAA(GAAAA)2GAGAAAGAGAAGAGAAGAGAAAGAA


GGAA.13: CTACT.GGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGCA.GCAGAGGTTGCAGTGAGCCGAGATCATGCCATTGCACTCCAGTATG-
ALU CON: CTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTG-
.180 .190 .200 .210 .220 .230 .240 .250
GGAA.13: .GCAATGAGAGATGAAACTCCATCTCTAAGAAAAAGAAAAGAAAGAAAAGAAAAGAAG(AAAAG)3(AGGG)2(AG)2(AAGG)2AAAGA-
ALU CON: GGCAAC.AGAG.CGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAA
.260 .270 .280 .290 .300
GGAA.13: AGGAAGAAAGAAGGAGGGAAGG(AG)3(AAGG)2AGGG(AAGG)3(R)~50GAAAGAAGAAAGAAAAGAAGAAAGGAA(GAAA)2GGAGA-
GGAA.13: AAAAGAAAA(GGAA)4
GGAA.15: CTATTCAGGAGGCTGAAGCAGAAGGATCACCTGAGCCTTGGGAGGTTGAGGCTGCAGTGAGCCAAGATC.....ATGCACTCCAGCCT-
ALU CON: CTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCC.GGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCT-
.180 .190 .200 .210 .220 .230 .240 .250
GGAA.15: GGGAGACAGAGTGAGACCCTGTCTCAAAAAAGAAAAAAAAAAAGAAAAA(GAAAA)2GAAAGA(GGGA)3(GA)2GAAAAGGAGGGAAA-
ALU CON: GGGCAACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAA
.260 .270 .280 .290 .300
GGAA.15: GA(GGAA)4GAGAAGGAAGGGAAGG(R)~30AGGAAGGGCGGG(AGGG)5(AGGGG)3


GGAA.16: CTACTCAGGAGGCTGAGGCAGAAGAATCACTTGAACCCAGGAGGCAGAGGTTGCAGTGAGCCCAGATCACCCCATTGCACTCCAGCCTG-
ALU CON: CTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTG-
.180 .190 .200 .210 .220 .230 .240 .250
GGAA.16: GGTGACAAGAGTGAAACTCCGCCGAAAGAAGAAGAAAGAAGGAAGGGAAGAA(GGAA)12GGA(GAAA)2AG(AAAG)3AAGAAAGAAG-
ALU CON: GGCAAC.AGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAA
.260 .270 .280 .290 .300
GGAA.16: GAAGGAGAAAAGG(GAAA)2GGAA(GAAA)2(GAA)2GGAAGAAAAGAAAAAAAGAAAAGAAGGAAGGGAAGAAAGAGGAGAGAAAGAA-
GGAA.16: AAGAAAGGAAGCA(GGAA)2GAAAGAAAAGAAAGA
GGAA.22: CTACTTGGGAGGCTAAGGCAGGAGAATCACTTGAACTCAGAAGGCAGAAGGTTGGAGTGAGCCAACATCACACCATTGCATTCGAGCCT-
ALU CON: CTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGCGGA.GGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCT-
.180 .190 .200 .210 .220 .230 .240 .250
GGAA.22: GGGCACGAGAGCAAAATTCCATCAGAGGAGAAGAAGAAGAAAAAAAG(AGG)2AGAAGAAGGAGAAGAGGAAGAGGAGGAGAAGAG-
ALU CON: GGGCAACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAA
.260 .270 .280 .290 .300
GGAA.22: G(AGG)2AGAAGGAGGAAGAGGAAGG(AAG)3G(AAGG)4(R)~40(GGAA)4(GGGA)5GGGGAGGGGGGA

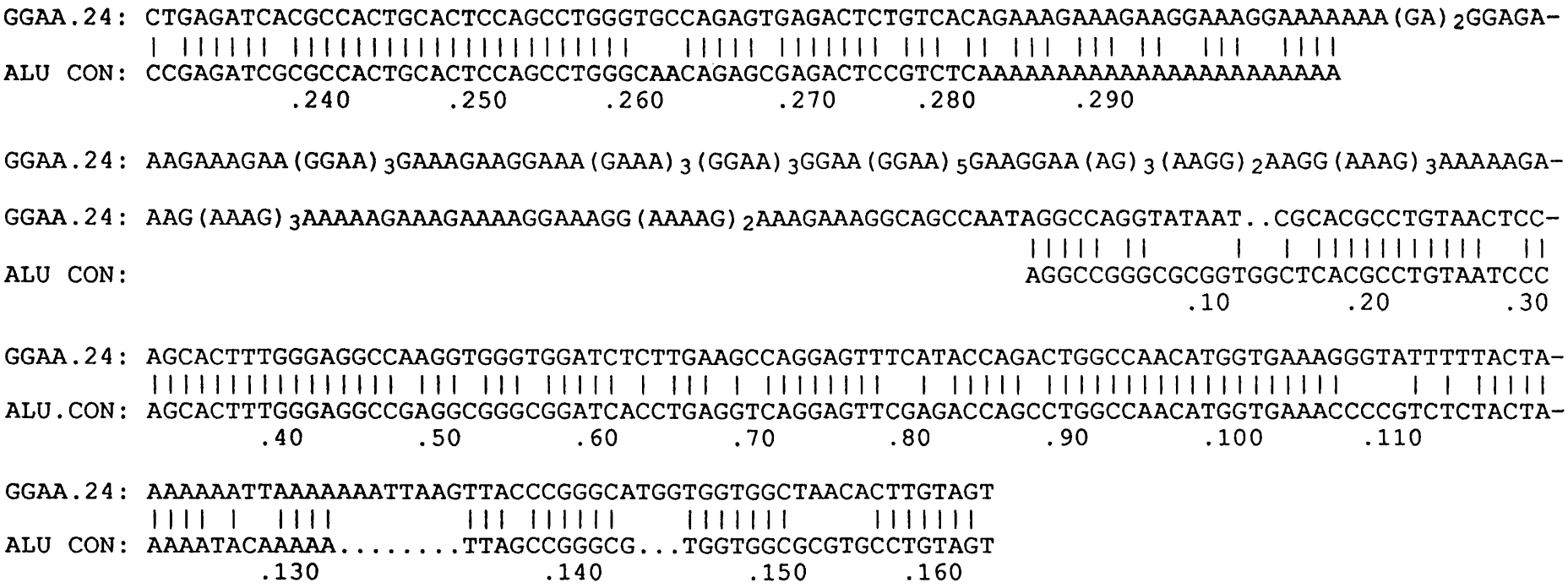

GGAA.24: CTGAGATCACGCCACTGCACTCCAGCCTGGGTGCCAGAGTGAGACTCTGTCACAGAAAGAAAGAAGGAAAGGAAAAAAA(GA)2GGAGA-
ALU CON: CCGAGATCGCGCCACTGCACTCCAGCCTGGGCAACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAA
.240 .250 .260 .270 .280 .290
GGAA.24: AAGAAAGAA(GGAA)3GAAAGAAGGAAA(GAAA)3(GGAA)3GGAA(GGAA)5GAAGGAA(AG)3(AAGG)2AAGG(AAAG)3AAAAAGA-
GGAA.24: AAG(AAAG)3AAAAAGAAAGAAAAGGAAAGG(AAAAG)2AAAGAAAGGCAGCCAATAGGCCAGGTATAAT..CGCACGCCTGTAACTCC-
ALU CON: AGGCCGGGCGCGGTGGCTCACGCCTGTAATCCC
.10 .20 .30
GGAA.24: AGCACTTTGGGAGGCCAAGGTGGGTGGATCTCTTGAAGCCAGGAGTTTCATACCAGACTGGCCAACATGGTGAAAGGGTATTTTTACTA-
ALU.CON: AGCACTTTGGGAGGCCGAGGCGGGCGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCGTCTCTACTA-
.40 .50 .60 .70 .80 .90 .100 .110
GGAA.24: AAAAAATTAAAAAAATTAAGTTACCCGGGCATGGTGGTGGCTAACACTTGTAGT
ALU CON: AAAATACAAAAA.....TTAGCCGGGCG...TGGTGGCGCGTGCCTGTAGT
.130 .140 .150 .160

Figure 7.3. Expansion of a $(GGAA)_n$ repeat motif from within an L1 element. The imperfect $(GGAA)_n$ array of clone GGAA.19 appears to be derived from a short purine rich region within an L1 element. L1 con is the L1 consensus sequence of Singer & Skowronski (1985).

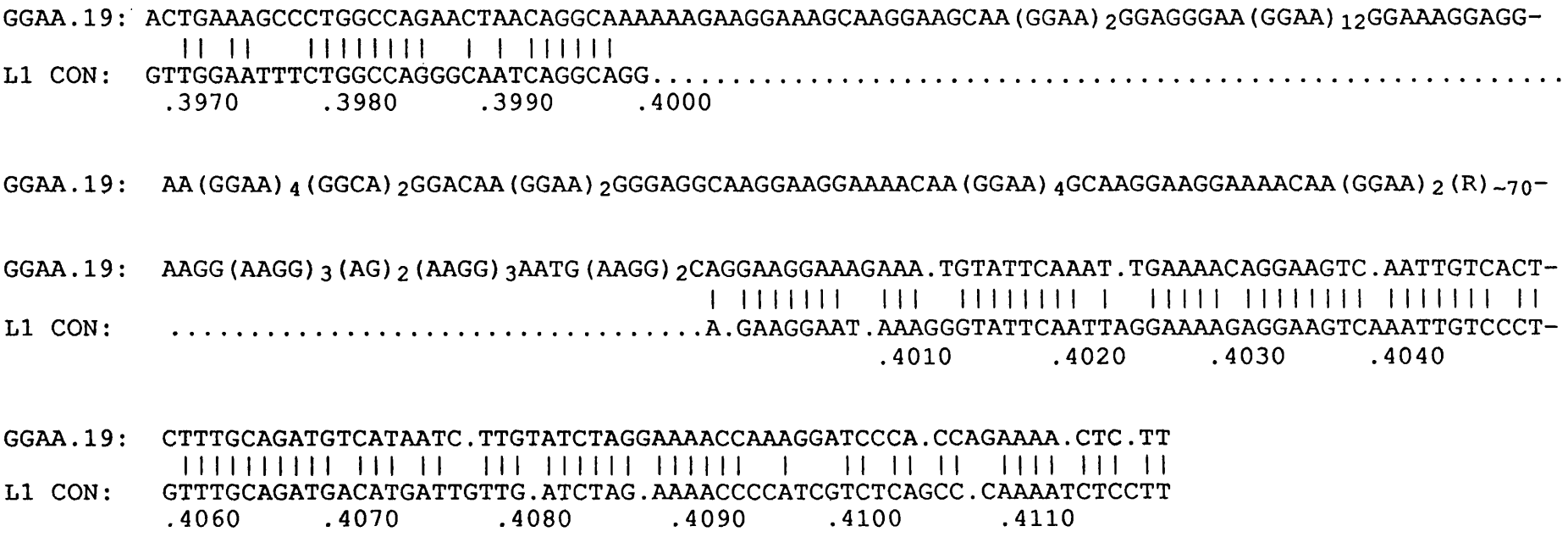

GGAA.19: ACTGAAAGCCCTGGCCAGAACTAACAGGCAAAAAAGAAGGAAAGCAAGGAAGCAA(GGAA)2GGAGGGAA(GGAA)12GGAAAGGAGG-
L1 CON: GTTGGAATTTCTGGCCAGGGCAATCAGGCAGG.....
.3970 .3980 .3990 .4000
GGAA.19: AA(GGAA)4(GGCA)2GGACAA(GGAA)2GGGAGGCAAGGAAGGAAAACAA(GGAA)4GCAAGGAAGGAAAACAA(GGAA)2(R)~70-
GGAA.19: AAGG(AAGG)3(AG)2(AAGG)3AATG(AAGG)2CAGGAAGGAAAGAAA.TGTATTCAAAT.TGAAAACAGGAAGTC.AATTGTCACT-
L1 CON: .....A.GAAGGAAT.AAAGGGTATTCAATTAGGAAAAGAGGAAGTCAAATTGTCCCT-
.4010 .4020 .4030 .4040
GGAA.19: CTTTGCAGATGTCATAATC.TTGTATCTAGGAAAACCAAAGGATCCCA.CCAGAAAA.CTC.TT
L1 CON: GTTTGCAGATGACATGATTGTTG.ATCTAG.AAAACCCCATCGTCTCAGCC.CAAAATCTCCTT
.4060 .4070 .4080 .4090 .4100 .4110

Figure 7.4. GGAA.1 and GGAA.7, imperfect $(GGAA)_n$ repeats flanked by unique sequence. Flanking regions selected for PCR priming are underlined.

GGAA.1:

TCTCTGCATGTTTGTTGCAGT(AAGG)$_4$G(AAGG)$_3$ATGGAAGGAGGGAAGGAGAGATGG(AGGG)$_2$AAGGAAGA(AAGG)$_3$AGAGAAGGAAGTAAAGA-

GGCAGGGAAGGAAGA(AAGG)$_4$AGAG(AAGG)$_2$AAAGAGGGAAAG(AAGG)$_3$TAGGGAGGCGGA(GGAA)$_{10}$(R)$_{-60}$TAGGAAGC(AAGG)$_8$AGGAAA-

AGGGAAGGAGTGAAGAAAG(AGGA)$_2$AGGTGGAGAATAAGGTTGGCTTCCCACACCATAGTCT

GGAA.7:

GATTTTGGAGGGCTACATGGCTTAAGGAGG(AAGG)$_9$AAAGAGGACGG(AGGG)$_4$(AAGG)$_2$G(AGGG)$_2$AGGCAGGGTTTGCTATCCAGCCTGTTAAGA

17%. (figure 7.5). Heterozygosity was estimated to be 92%. PCR amplification of GGAA.1 from human/rodent somatic cell hybrid DNAs (provided by Dr Sue Povey, University College London; described in table 2.1) placed it on chromosome 13, and subsequent kindred studies involving four large CEPH families (see figure 7.6) followed by linkage analysis using the LINKAGE programs (Lathrop *et al.*, 1985) and the CEPH database version 2 (with the assistance of Dr John Armour) further localized it to an interstitial position on the q arm with the most proximal informative marker being D13S2 (Leppert *et al.*, 1986; $z=5.7$ at $\theta=0$).

(b) GGAA.7

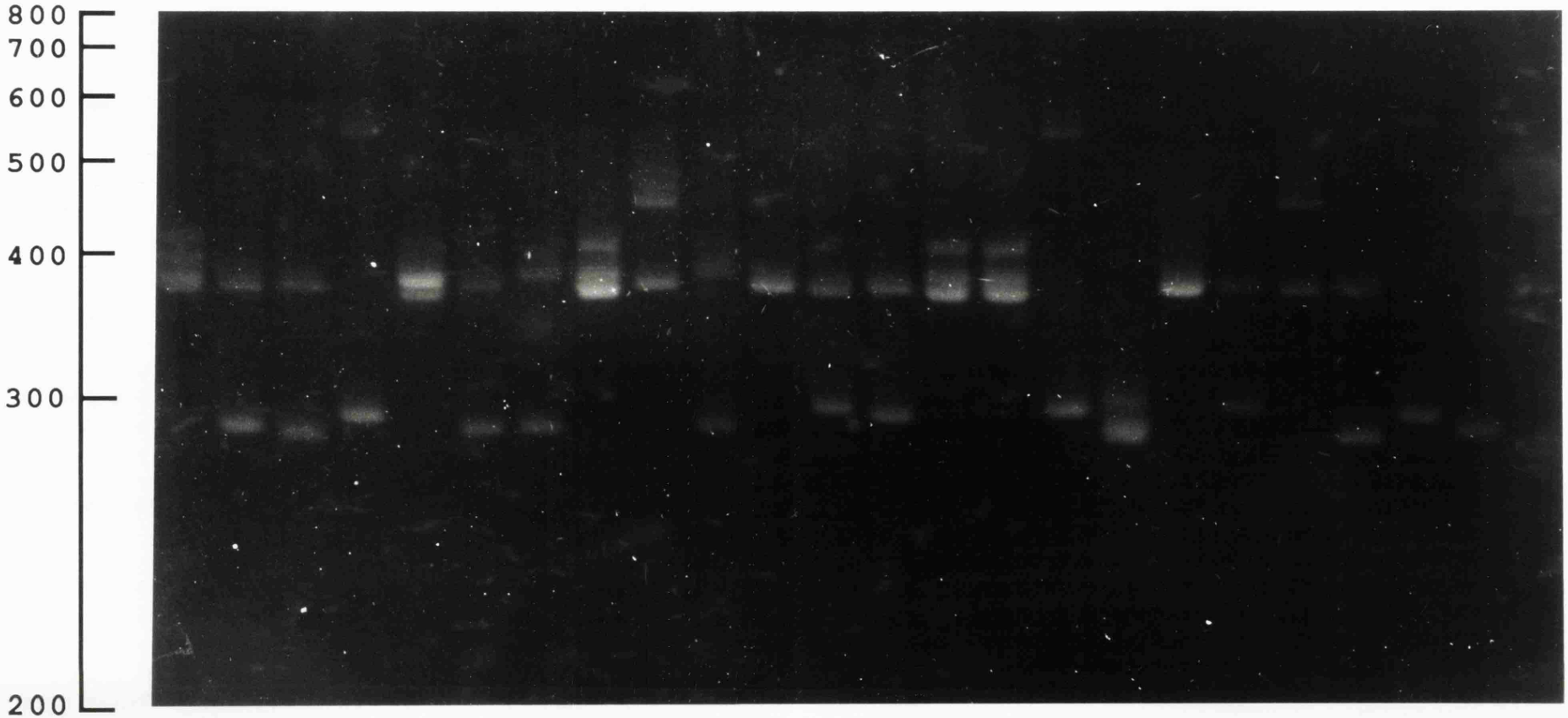
Similarly, PCR primers were designed for amplification of the GGAA.7 repeat array. Analysis of PCR products from 24 unrelated CEPH individuals was carried out as for GGAA.1. GGAA.7 was also found to be highly variable, with eight resolvable alleles ranging smoothly from approximately 147 to 182bp in length (figure 7.7). However, in this instance, a common short (147bp) allele was observed, with a frequency of 43.8%. Heterozygosity was determined to be 88% at this locus. PCR amplification of human/rodent somatic cell hybrid DNAs placed GGAA.7 on chromosome 2, and typing four large CEPH family groups (see figure 7.8) allowed more precise localization to an interstitial position on the short arm with the most proximal informative maker being pEFD122 (O'Connell *et al.*, 1989; $z=3.27$ at $\theta=0.08$).

7.5 GGAA.1 and GGAA.7 in the great apes

GGAA.1 and GGAA.7 were amplified from great ape DNA under the same conditions as for human DNA. GGAA.1 appeared to be highly polymorphic in the chimpanzee but showed minimal variation in the gorilla (figure 7.9A). Alleles from both species were in the size range 280-320bp, the same size as the smallest human alleles. This locus failed to amplify from orang-utan DNA. GGAA.7 was found to be variable in chimpanzee, gorilla and orang-utan (figure 7.9B) with a similar allele size distribution to that seen in man.

Figure 7.5. Variability at GGAA.1.
A. 24 unrelated individuals typed for GGAA.1. 10ng of DNA was cycled 33 times in a 10µl PCR reaction as described in table 2.2. PCR products were electrophoresed through a 4% agarose gel (3 parts nusieve, 1 part HGT) in TBE with ethidium bromide. A third band representing a heteroduplex of the two alleles is evident in heterozygotes.
B. (Overleaf). Allele frequency distribution at GGAA.1. Unlike microsatellites, GGAA.1 appears to have a smooth allele frequency distribution with no common alleles.

A.



B.

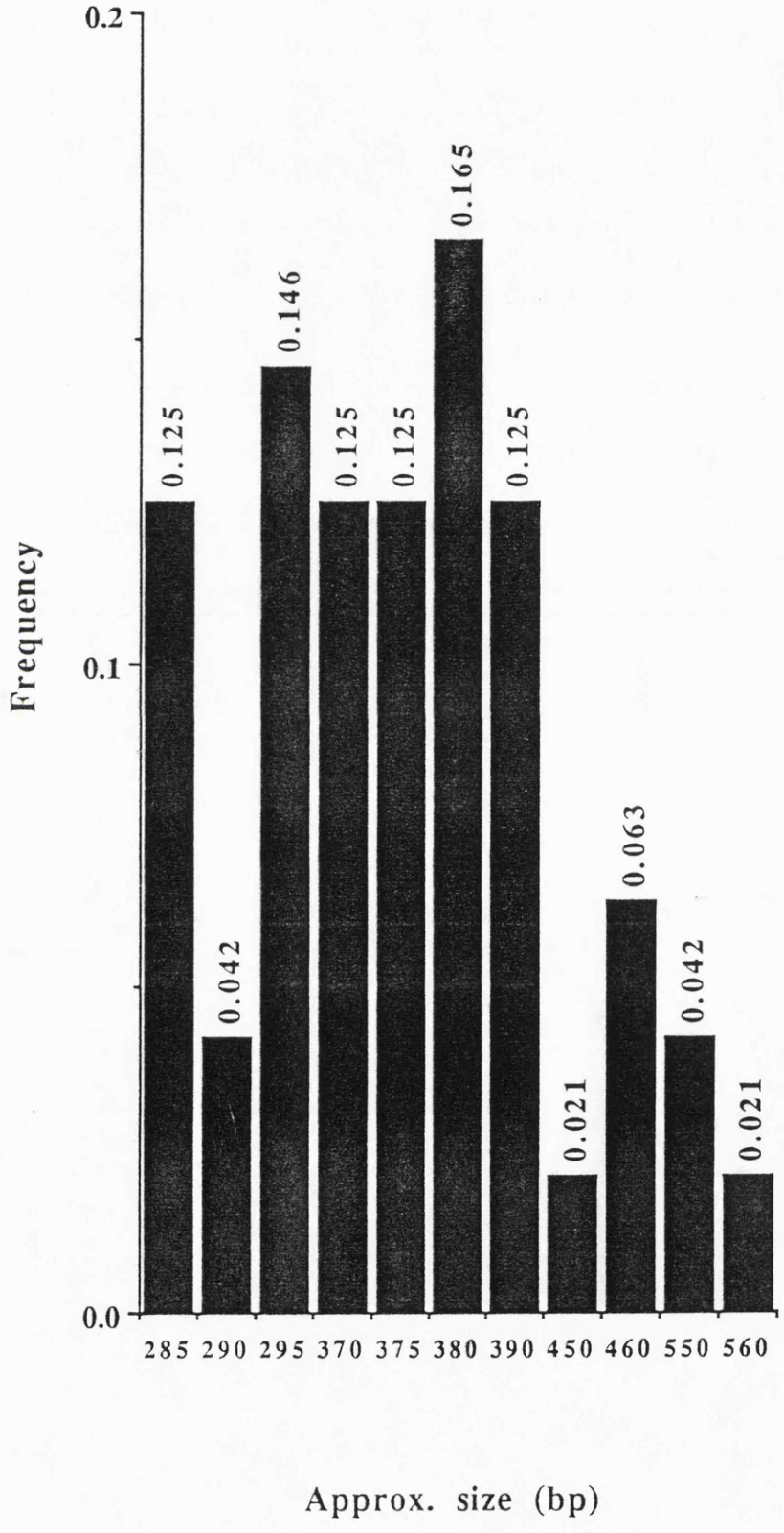

0.2
0.1
0.0
Frequency
0.125
0.042
0.146
0.125
0.125
0.165
0.125
0.021
0.063
0.042
0.021
285 290 295 370 375 380 390 450 460 550 560
Approx. size (bp)

Figure 7.6. Pedigree analysis of GGAA.1 through three generations of a large CEPH family. Experimental procedures were as for figure 7.5. Each allele is indicated on the right, and the genotype of each individual is shown above the appropriate lane.

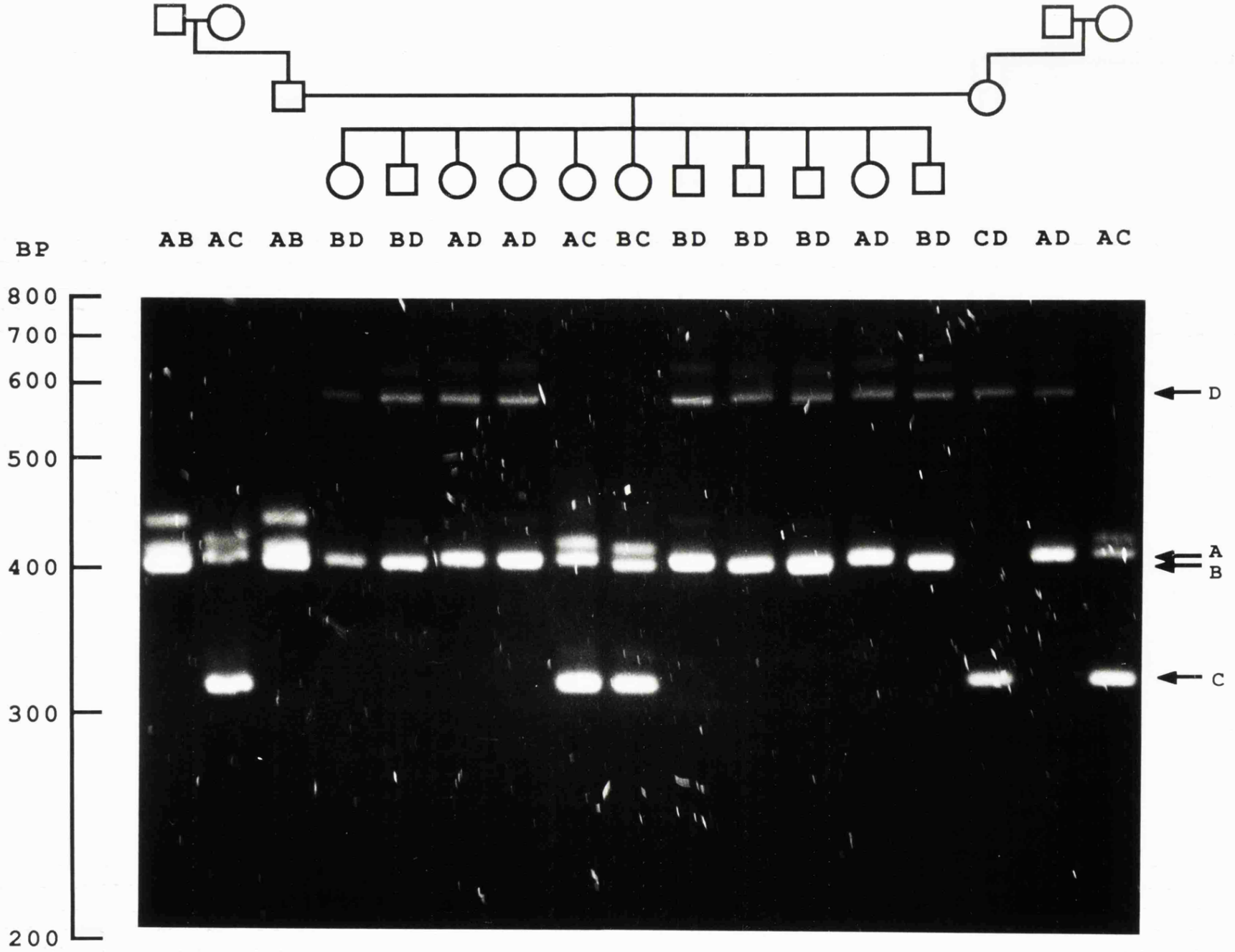
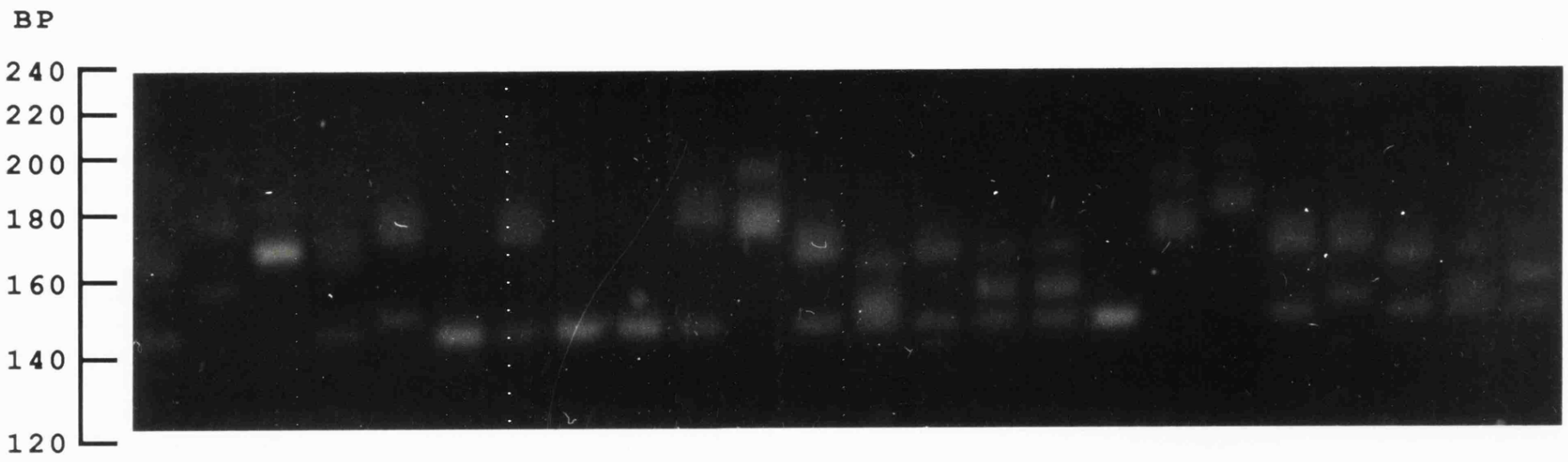

AB AC AB BD BD AD AD AC BC BD BD BD AD BD CD AD AC
BP
800
700
600
500
400
300
200
D
A
B
C

Figure 7.7. Variability at GGAA.7.
A. 24 unrelated individuals typed for GGAA.7. 10ng of DNA was cycled 33 times in a 10µl PCR reaction as described in table 2.2. PCR products were electrophoresed through a 4% agarose gel (3 parts nusieve, 1 part HGT) in TBE with ethidium bromide. Again, heteroduplexes are evident in heterozygotes.
B. (Overleaf). Allele frequency distribution at GGAA.7. Unlike GGAA.1, GGAA.7 has a common small allele (frequency 0.44).

A.



B.

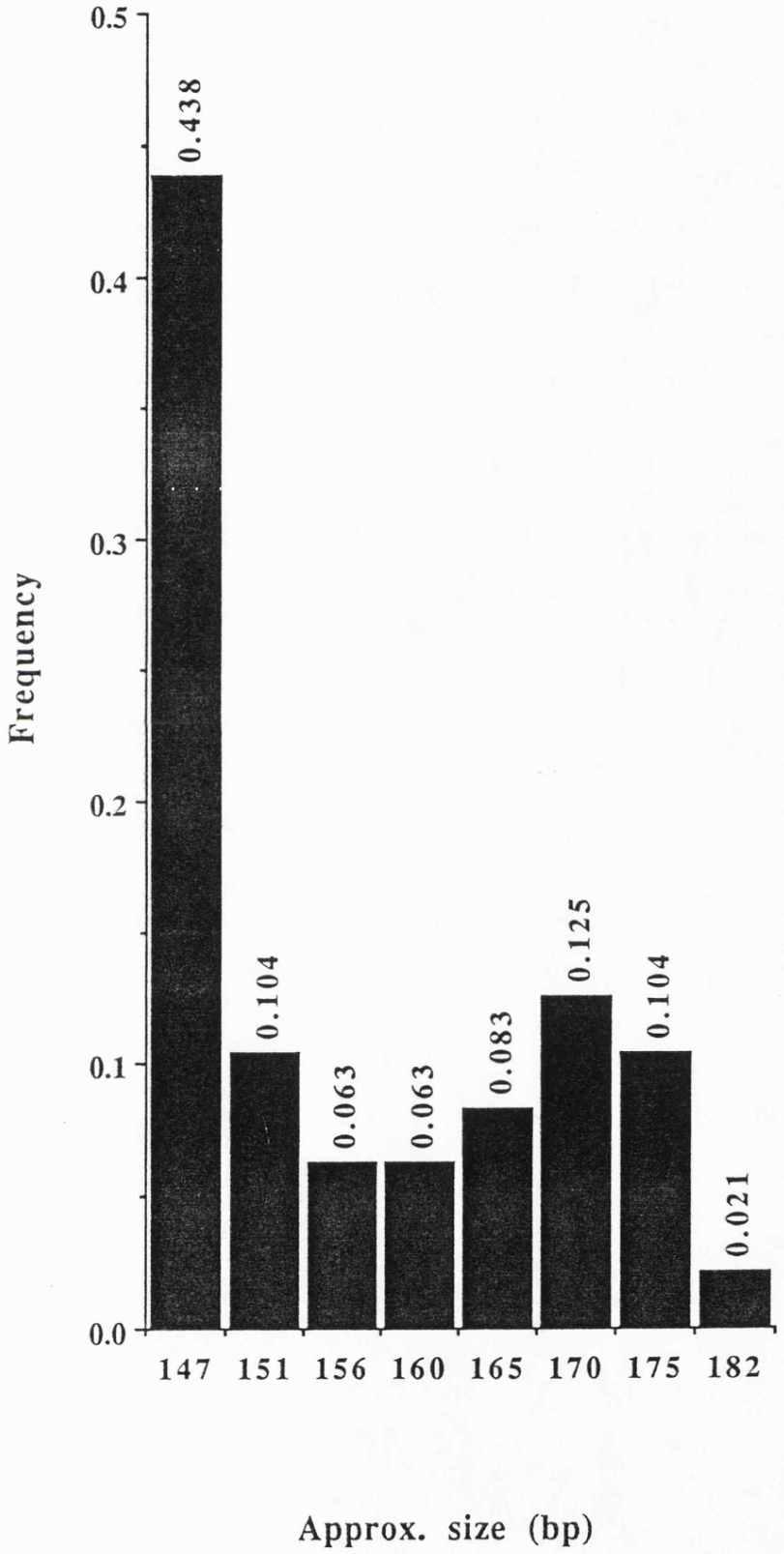

0.5
0.4
0.3
0.2
0.1
0.0
Frequency
0.438
0.104
0.063
0.063
0.083
0.125
0.104
0.021
147
151
156
160
165
170
175
182
Approx. size (bp)

Figure 7.8. Pedigree analysis of GGAA.7 through three generations of a large CEPH family. Experimental procedures were as for figure 7.7. Each allele is indicated on the right, and the genotype of each individual is shown above the appropriate lane.

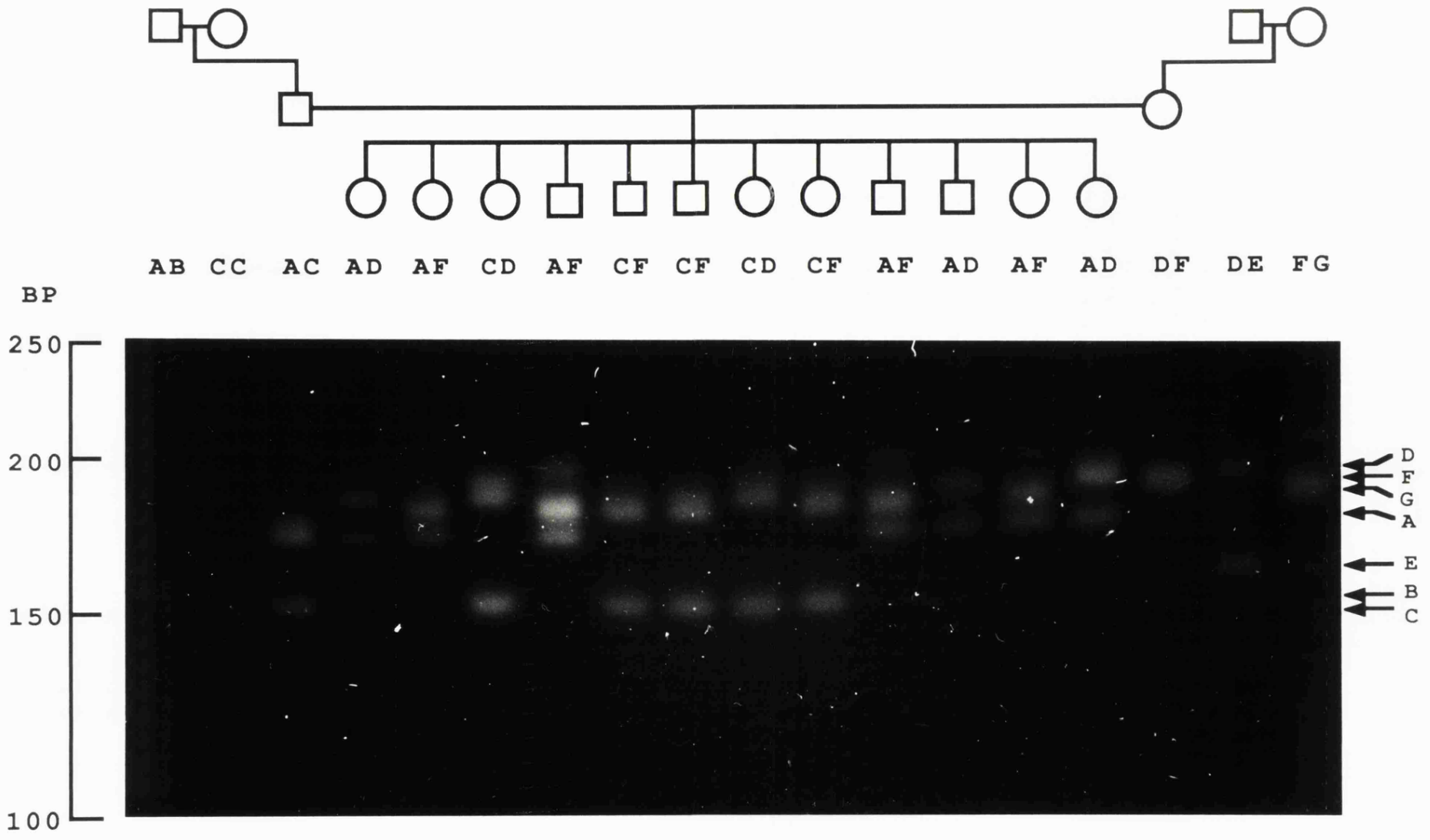
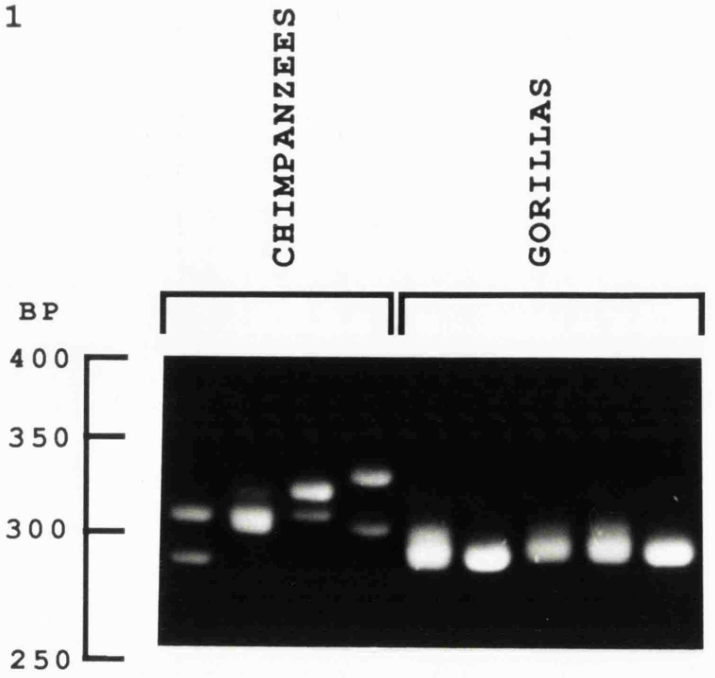
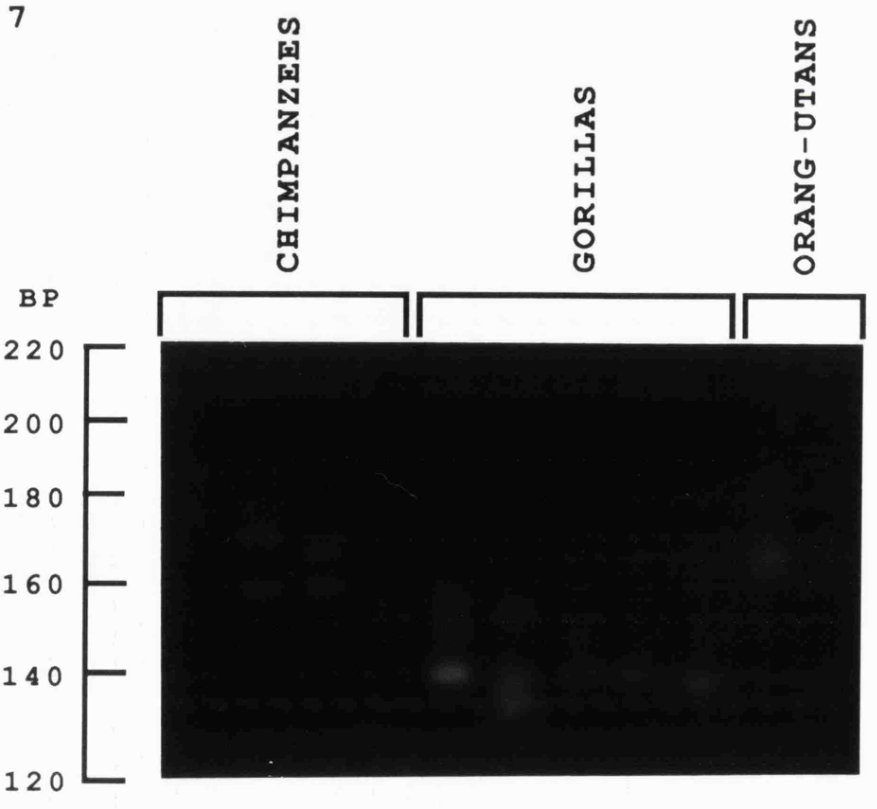

AB CC AC AD AF CD AF CF CF CD CF AF AD AF AD DF DE FG
BP
250
200
150
100
D
F
G
A
E
B
C

Figure 7.9. GGAA.1 and GGAA.7 alleles amplified from great ape DNA. Both loci were amplified and analysed in the same fashion as the human loci.
A. GGAA.1 appears to be highly variable in the chimpanzee, but shows minimal variability in the gorilla. This locus failed to amplify from orang-utan DNA.
B. GGAA.7 is variable in all three species, with an allele size distribution similar to that seen in man.

GGAA.1


CHIMPANZEES
GORILLAS
BP
400
350
300
250

GGAA.7


CHIMPANZEES
GORILLAS
ORANG-UTANS
BP
220
200
180
160
140
120

7.6 Characterization of GGAA.6, an unusually modified polyadenosine tail of an Alu repeat

On sequencing, clone GGAA.6 was found to have a particularly unusual structure: an alu repeat with polyadenosine tail followed by a short $(GGAA)_n$ repeat, which was in turn followed by a TC rich region and a polythymidine tract, ie an imperfect inverted repeat (figure 7.2). This structure is capable of forming a hairpin loop as depicted in figure 7.10. Polyadenosine tails of Alu repeats had previously been shown to vary in the number of residues between individuals (Economou *et al.* 1990) and it therefore seemed likely that this unusual structure, consisting of repeated motifs, would also be variable. Consequently, PCR primers complementary to the flanking sequence were designed. It was possible to design a unique sequence PCR primer from sequence within the Alu repeat upstream of the polyadenosine tail due to large differences from the consensus Alu sequence (Bains, 1986) at the extreme 3' end, comprising of a deletion of approximately 10bp (figure 7.2).

PCR amplification of GGAA.6 from 24 unrelated individuals revealed that the locus was indeed variable, with six alleles and a heterozygosity of 74% (figure 7.11), comparable to a typical $(dA\text{-}dC)_n\cdot(dT\text{-}dG)_n$ microsatellite locus. Due to the close spacing of the alleles on high percentage agarose gels, GGAA.6 alleles were also sized on a denaturing polyacrylamide sequencing-type gel, after end-labelling one of the PCR primers with γ-^{32}P-dATP (figure 7.11B). The allele frequency distribution is dominated by two common alleles, again typical of a microsatellite-like locus. PCR amplification from somatic cell hybrid DNAs placed GGAA.6 on chromosome 12. This was confirmed by linkage analysis using four large CEPH pedigrees (see figure 7.12), which allowed further localization to the proterminal region of the q arm, the closest informative marker being MS43 (Wong *et al.*, 1987; $z=3.16$ at $\theta=0.215$; linkage analysis using the CEPH database was performed with the assistance of Dr John Armour).

PCR amplification of GGAA.6 from great ape DNA revealed an apparently monomorphic locus in each species, the same sized product being shared by chimpanzee and gorilla (roughly the same size as the smallest human allele) and a slightly larger band from the orang-utan

Figure 7.10. Putative hairpin structure of the inverted repeat in clone GGAA.6. The open box represents the Alu element immediately preceding this structure.

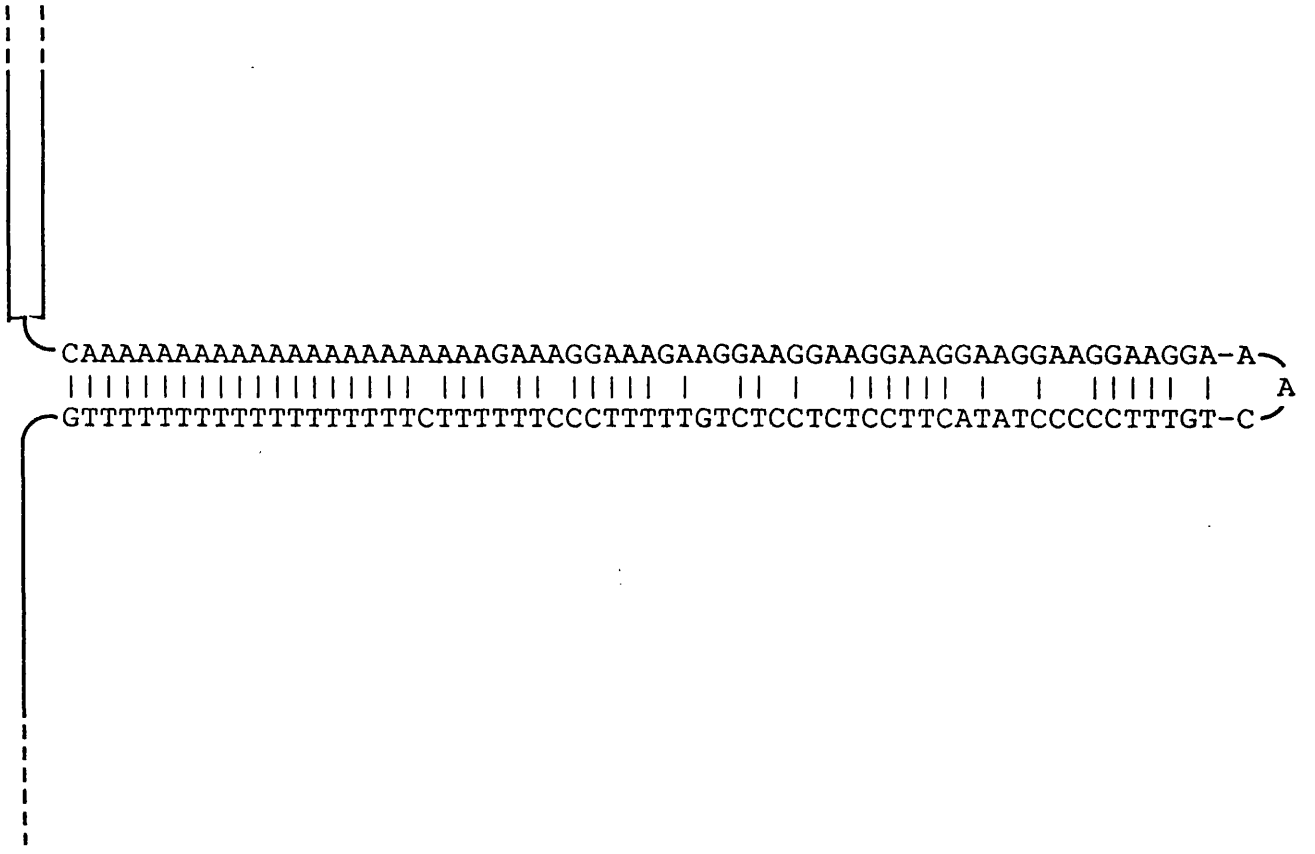
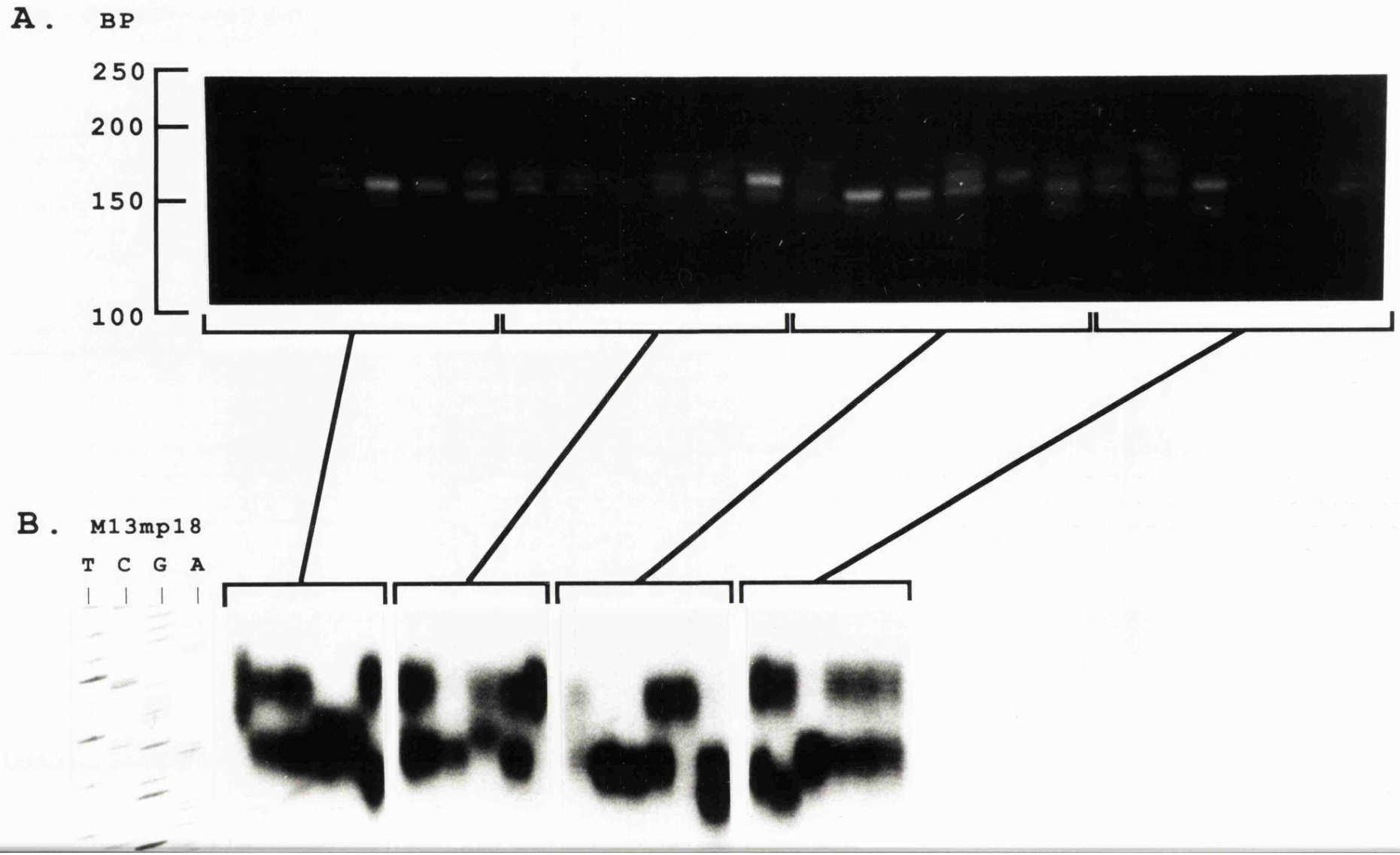

CAAAAAAAAAAAAAAAAAAAAAGAAAGGAAAGAAGGAAGGAAGGAAGGAAGGAAGGA-A
A
GTTTTTTTTTTTTTTTTTTCTTTTTTCCCTTTTTGTCTCCTCTCCTTCATATCCCCCTTTGT-C

Figure 7.11. GGAA.6 alleles, amplified from 24 unrelated individuals and typed native on a 4% agarose gel (A) or denatured on a polyacrylamide sequencing type gel following end-labelling (B). Profiles generated from the same individual by each method are indicated by a connecting line. Heteroduplexes are evident in (A). End-labelled single-stranded products were run alongside an M13mp18 marker ladder. Spurious products, thought to arise through DNA slippage (Litt & Luty, 1989) and terminal transferase activity of Taq polymerase (Clark, 1988) are evident. Approximate allele sizes in C (overleaf) are derived from the major PCR product in B. The allele frequency distribution is clearly dominated by two common alleles (168bp; frequency 0.44 and 175bp; frequency 0.40).


A.
BP
250
200
150
100
B.
M13mp18
T C G A

C .

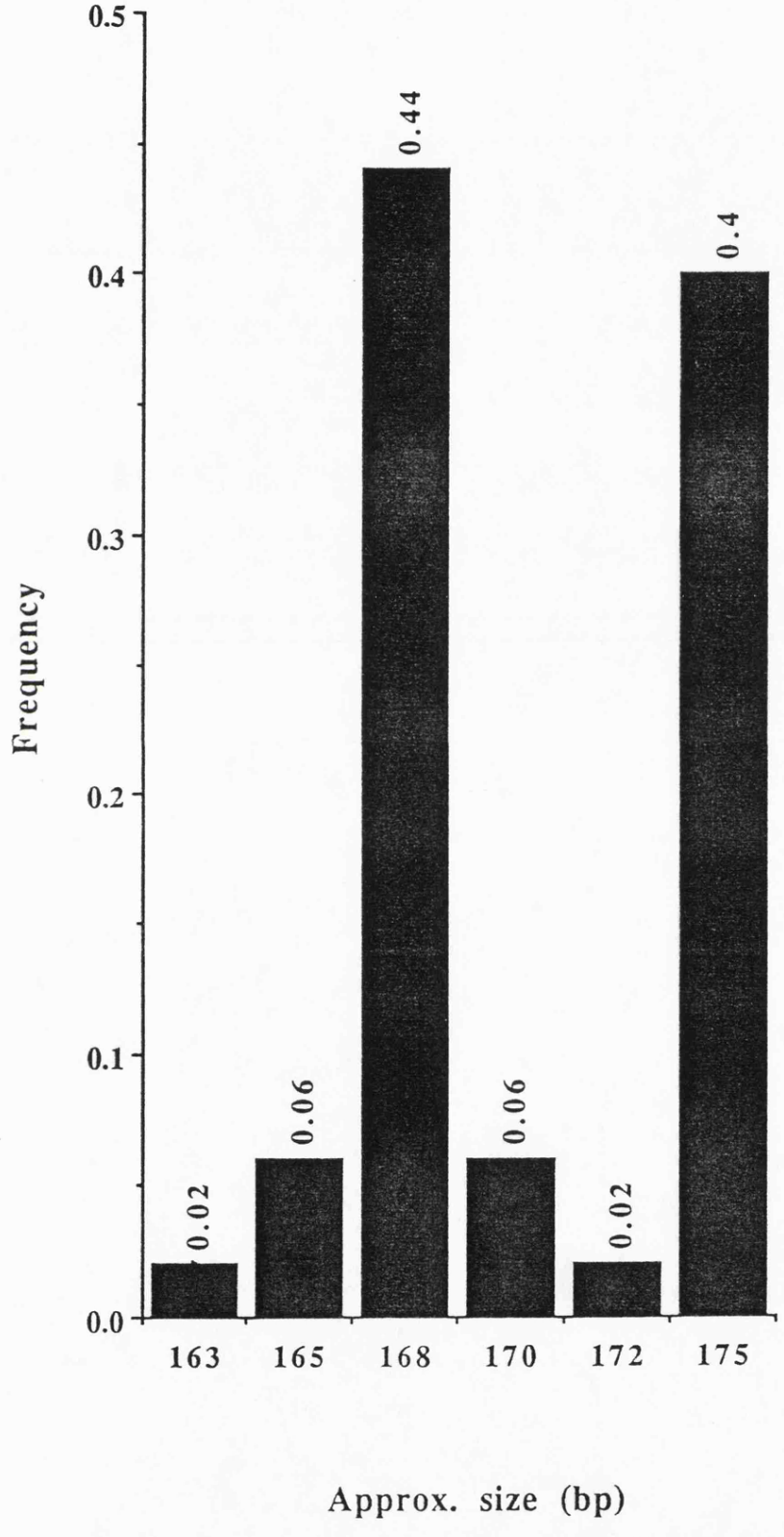
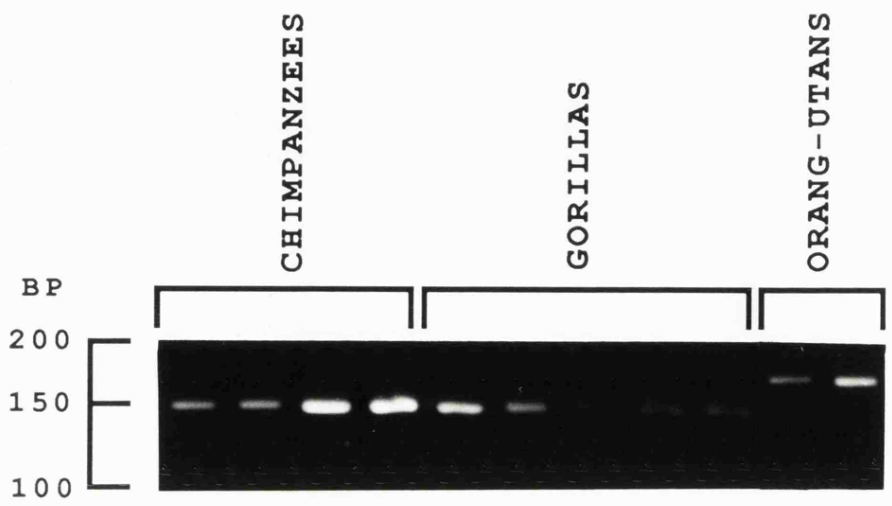

0.5
0.4
0.3
0.2
0.1
0.0
Frequency
0.02
0.06
0.44
0.06
0.02
0.4
163
165
168
170
172
175
Approx. size (bp)

Figure 7.12. Mendelian inheritance of GGAA.6 through 3 generations of a large CEPH pedigree. PCR products were electrophoresed through a 3% nusieve, 1% HGT agarose gel in TBE with ethidium bromide. The genotype of each individual is shown above the appropriate lane. Heteroduplexes can be seen in heterozygotes.


AB BC BB BB BB AB AB BB BB AB BB AB AB BB BB AB AB BB
BP
240
220
200
180
160

Figure 7.13. GGAA.6 in the great apes. Following PCR, alleles were electrophoresed through a 3% nusieve, 1% HGT agarose gel in TBE with ethidium bromide. GGAA.6 appears to be monomorphic from this limited sample. However, although the chimpanzee and gorilla alleles are the same size, sequence analysis revealed differing internal structures (see text).


CHIMPANZEES
GORILLAS
ORANG-UTANS
BP
200
150
100

(figure 7.13). Direct sequence analysis of PCR amplified chimpanzee and gorilla alleles (by the method of Winship, 1989) showed them to have an identical overall structure to the human locus (data not shown). However, although both species had identically sized alleles, a difference in internal structure was evident. The gorilla allele was found to have five perfect GGAA repeats, whereas the chimpanzee allele had only four, a deficiency compensated for by four extra thymidine residues in the polythymidine tract. In view of this and of the limited number of individuals typed, GGAA.6 may be polymorphic in length within these species, and the shared bands seen here may merely represent particularly common alleles.

The GGAA.6 locus has three potentially polymorphic regions: the polyadenosine tail, the $(GGAA)_n$ repeat array and the polythymidine tract. PCR amplification followed by direct sequencing of the two most common human alleles revealed variation in the polyadenosine and polythymidine regions (ie extra adenosine and thymidine residues in the longer allele), and a constant $(GGAA)_6$ array. However, this tetranucleotide array was found to be variable between man, chimpanzee and gorilla (see above); it may therefore be concluded that polymorphism occurs in all three repeated regions at this locus.

7.7 Discussion

Of the 11 sequenced clones detected by the $(GGAA)_n$ repeat probe, 8 had inserts derived from extended Alu polyadenosine tails which had adopted a $(GGAA)_n$ repeated motif, and had expanded to 100-300bp in length (figure 7.2). Consequently, these tandem repeats were unsuitable for a PCR assay of variability, due to the unavailability of unique sequence DNA for PCR priming immediately 5' to each array. With hindsight, this problem could be obviated by a secondary library screen with an Alu consensus probe, and Alu-positive clones disregarded. A library consisting of an ordered array of recombinant phage would allow for a more efficient screening procedure in this case; such a library has recently been described for the rapid isolation of novel minisatellite loci (Armour et al., 1990).

The formation of $(GGAA)_n$ repeats from Alu polyadenosine tails, and the expansion of an array from within a Kpn element (GGAA.19) again raises the question of the relationship between tandem arrays and dispersed repetitive elements. The association of minisatellite tandem repeats with dispersed repeats is well documented (Armour *et al.*, 1989b, 1991; Kelly *et al.*, 1989; see chapter 3) but the significance is not well understood. It is possible that those regions of DNA generally associated with dispersed repeats show relaxed fidelity in DNA replication, allowing the formation of tandem repeat sequences.

Based on a limited sample of 24 unrelated individuals, GGAA.1 and GGAA.7 were found to have 11 and 8 alleles respectively, as resolved by agarose gel electrophoresis (figures 7.4 and 7.6), and heterozygosities estimated at 92 and 88%. This high level of variability is despite the fact that the originally sequenced GGAA.1 and GGAA.7 alleles were comprised mainly of imperfect repeats (figure 7.2) which would be expected to reduce variability by hindering mutation mechanisms such as DNA slippage during replication (see chapter 5). However it is possible that variability is confined entirely to those regions of the array which are comprised of perfect GGAA repeats. The apparently limited variability of GGAA.1 in the gorilla when compared to man and chimpanzee may again reflect evolutionary transience of highly variable tandem repeat loci (see chapter 3). GGAA.7, which shows more modest variability in man is also polymorphic in chimpanzee, gorilla and orang-utan.

These tetranucleotide repeats appear to far more informative than $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ microsatellite markers, which typically have modest heterozygosites and allele distribution frequencies dominated by common alleles (Weber & May, 1989). A further advantage of $(GGAA)_n$ repeats is the greater ease with which alleles can be resolved; although it has been shown that high percentage agarose gels are adequate for resolving microsatellite alleles in many instances (chapters 5 and 6), difficulties arise were alleles differ by a single repeat unit (2bp), and polyacrylamide sequencing-type gels may be necessary in such instances. This is unlikely to be the case for a locus such as GGAA.1, which has well-spaced alleles ranging from 280-560bp. The combination of high variability with with

small allele lengths renders such loci ideal for PCR analysis in forensic cases where DNA is limited or of poor quality (see chapter 6).

Linkage analysis of GGAA.1 and GGAA.7 has shown that unlike minisatellites (Royle *et al.*, 1988), $(GGAA)_n$ arrays do not appear to be localized to the proterminal regions of chromosomes, and may therefore prove to be highly useful linkage markers in genetic analysis. Given that 2 highly polymorphic $(GGAA)_n$ arrays with unique flanking sequence appropriate for PCR amplification were isolated from a library of approximately 10,000 clones, and that the average insert size was 700bp, such arrays, if evenly dispersed, should be found once every 35,000kb; ie >800 copies in the human genome. $(GGAA)_n$ repeats should therefore provide a new set of highly informative markers.

The locus GGAA.6 is a bizarre extended polyadenosine tail comprising an inverted repeat which may be capable of forming a hairpin structure. PCR amplification has revealed this locus to be moderately variable, with a heterozygosity of 76% and 6 alleles, two of which are very common (figure 7.10). This is a typical profile of a polymorphic simple tandem repeat such as a microsatellite (Weber & May, 1989). Sequence analysis of GGAA.6 in the chimpanzee and gorilla revealed the overall structure and polymorphic nature to be ancient, as expected for a locus with a relatively low mutation rate and in contrast to the extreme transience of highly variable minisatellites (see chapter 3).

CHAPTER 8

CHARACTERIZATION OF TWO HUMAN 'MIDISATELLITE' LOCI

8.1 Introduction

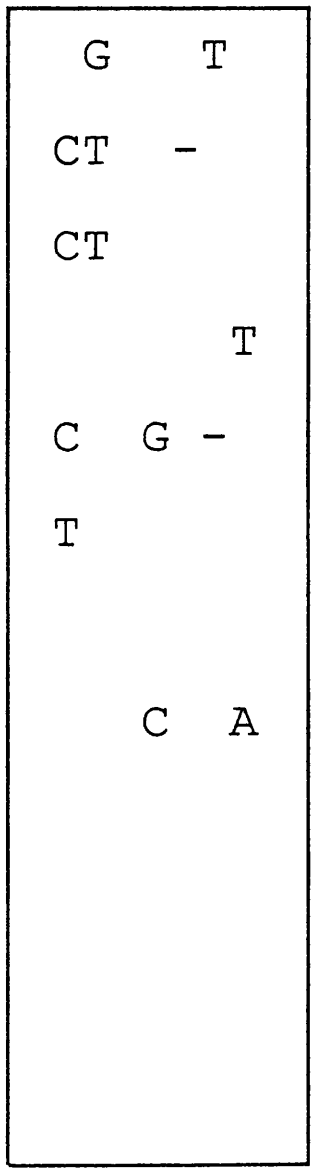
Within the human genome, a hierarchy of tandem repeat structures exist, from microsatellites and simple tandem repeats of less than a kilobase, through minisatellites a few kilobases in length to satellite DNA, which may span several hundred kb. Within this hierarchy, a further class of repeat sequences has been identified: 'midisatellites'. As the name suggests, midisatellites fall between minisatellites and satellite DNA in size range. Two midisatellites have been reported to date, a 40bp tandem repeat of some 250-500kb near the telomere of the short arm of chromosome 1 (Nakamura *et al.*, 1987b) and a more modest 10-50kb array of 61bp tandem repeats located in the pseudoautosomal region of the X and Y chromosomes (Page *et al.*, 1987). Both of these loci exhibit length polymorphism as well as variation in internal repeat sequence, which can be revealed by restriction enzymes that cut some repeat units but not others, producing a complex and variable haplotype of cosegregating fragments. This chapter describes the characterization of two further midisatellite loci, both arrays being detected by the same repeat sequence.

8.2 Isolation of a midisatellite repeat array from an M13 genomic library, and subsequent sequence analysis

Human DNA from 16 individuals was pooled and digested with *Alu*I, *Hae*III and *Rsa*I. The 400-1000bp fraction was gel purified and ligated into the *Sma*I site of M13mp18 (Yanis-Perron *et al.*, 1985) to generate a library of approximately 10,000 clones (see chapter 5). This library was screened with a $(GGCA)_n$ repeat probe as described in chapter 7. Two plaques cross-hybridized strongly to this probe, and were designated GGCA.1 and GGCA.2. GGCA.1 is described in chapter 7. Sequence analysis of clone GGCA.2 revealed an insert consisting of a 9bp tandem repeat sequence with the consensus AGCCAAGCC (presumably derived from an AGCC tetranucleotide repeat), but with a large degree of divergence between repeats (figure 8.1).

Figure 8.1. GGCA.2 tandem repeats. Twelve 9bp tandem repeats from clone GGCA.2 are shown, with base changes from the consensus.

GGCA.2: CONSENSUS: AGCCAAGCC



8.3 Southern transfer analysis of GGCA.2

The original purpose of screening an M13 library with a $(GGCA)_n$ probe was to isolate variable tetranucleotide repeats amenable to typing by PCR (see chapter 7). However, the insert of clone GGAA.2 consisted entirely of repeated sequence and so gave no opportunity to design flanking PCR primers. Consequently, in order to establish the nature of this repeat sequence, a probe was generated by PCR amplification of the GGCA.2 insert using the forward and reverse M13 sequencing primers. The PCR product was gel purified and used to probe a Southern blot of DNA from 3 unrelated individuals cut with 3 different restriction enzymes. Washing at high stringency (0.1x SSC, 0.01% SDS) left a complex pattern of monomorphic and polymorphic bands, varying in size distribution according to the restriction enzyme used (figure 8.2).

To determine whether these bands represented a single locus or multiple scattered loci, pedigree analysis was performed. Aliquots of DNA from a large family consisting of mother, father and 14 offspring were digested with *Hin*fI, which gave the largest number of polymorphic bands, and the resulting Southern blot probed with GGAA.2 (figure 8.3). Two sets of cosegregating polymorphic bands were clearly observed, implying that GGCA.2 detects at least two polymorphic loci. Obviously it was not possible to follow the inheritance pattern of the monomorphic bands. Consequently such fragments may be sections of either the two polymorphic arrays or alternatively may represent one or more completely different loci.

8.4 Chromosomal assignment of the loci detected by GGCA.2

On the assumption that just two loci were detected by GGCA.2, an attempt to determine the chromosomal location of each locus was carried out by somatic cell hybrid analysis. DNAs from a panel of human/rodent hybrid cell lines (described in table 2.1; provided by Dr Sue Povey, University College, London) were digested with *Hin*fI and the resulting Southern blot probed with GGCA.2. This placed one locus firmly on chromosome 2, and the second tentatively on chromosome 15. Subsequent linkage analysis using the LINKAGE programs (Lathrop *et al.*, 1985) and the CEPH database version 2 (with the assistance of Dr John Armour)

Figure 8.2. Multiple DNA fragments detected by GGCA.2 at high stringency. DNA from 3 unrelated individuals (1,2,3) was digested with *Hin*fI, *Hae*III or *Alu*I and electrophoresed through a 0.8% agarose gel in TAE. Following transfer to a nylon membrane (Amersham Hybond-N) hybridization with the GGCA.2 insert, ^{32}P-labelled by random oligonucleotide priming, was performed at high stringency.

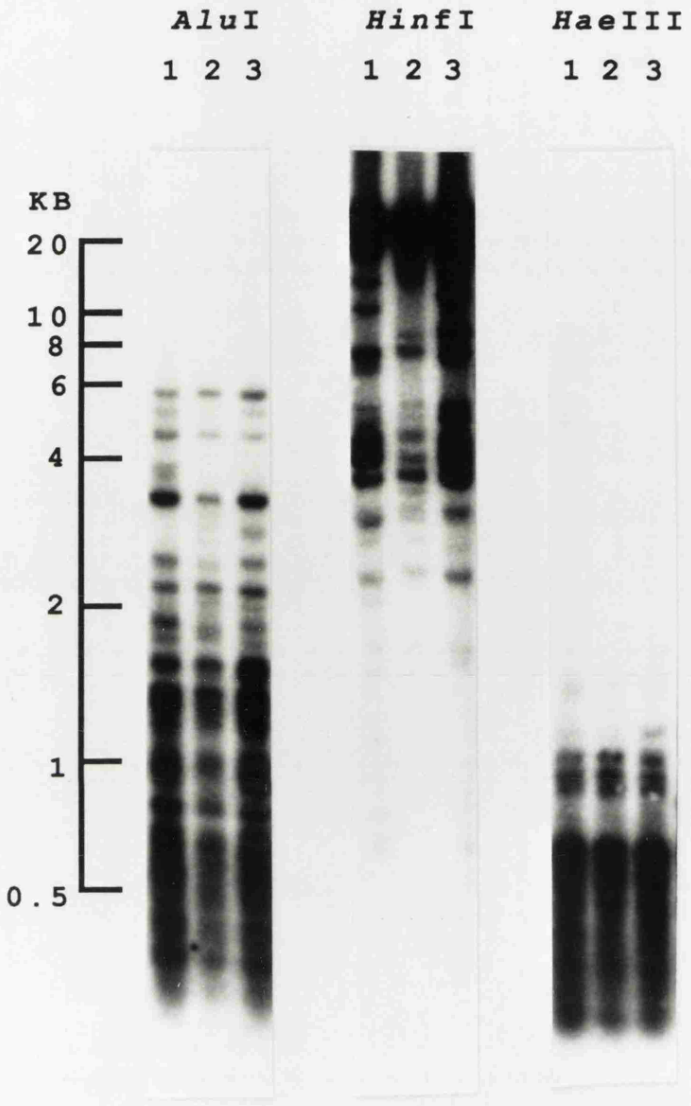
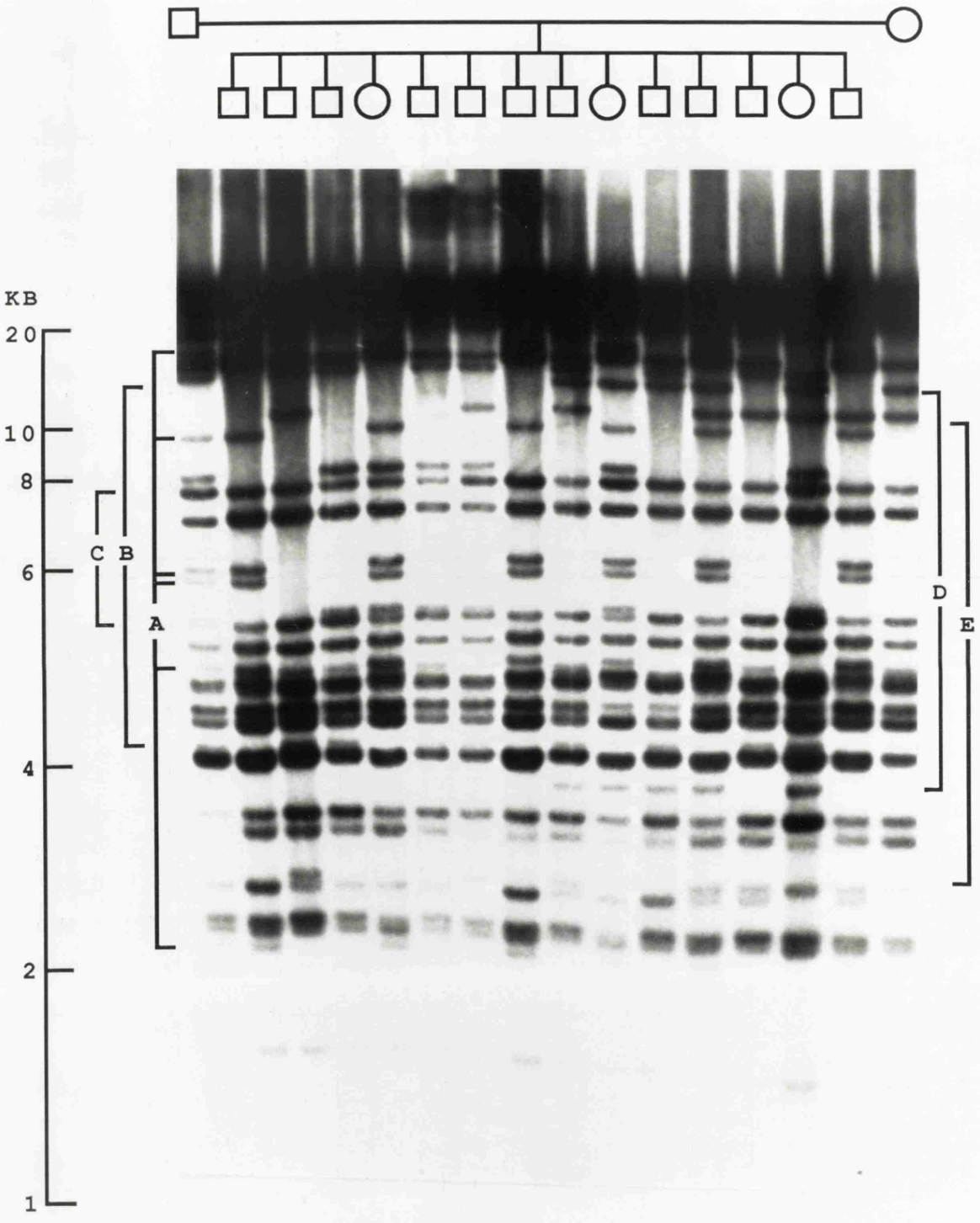

AluI
HinfI
HaeIII
1 2 3
1 2 3
1 2 3
KB
20
10
8
6
4
2
1
0.5

Figure 8.3. Pedigree analysis of the loci detected by GGCA.2 in *Hin*fI digested DNAs from a large CEPH family. Sets of co-segregating fragments are joined by brackets and designated A-E. Paternal haplotypes A and B are allelic, with haplotype C showing a completely different segregation pattern. Similarly, maternal haplotypes D and E do not show allelic segregation. GGCA.2 therefore detects 2 polymorphic loci in human DNA. Monomorphic bands may represent further loci detected by GGCA.2, or may be fragments derived from the variable loci. Pulsed-field gel and somatic cell hybrid analysis suggest that the latter is more likely. Experimental procedures were as for figure 8.2.


KB
20
10
8
6
4
2
1
A
B
C
D
E

confirmed that locus 1 was on chromosome 2, and further positioned it near the centromere, the closest marker being CRI-C13B (Donis-Keller *et al.*, 1987; z=3.61 at θ=0).

8.5 Size estimation by pulsed field gel analysis

A Southern blot of a pulsed-field gel containing human DNA cut with a range of restriction enzymes, kindly provided by Dr. Nicola Royle, was probed with GGCA.2 at high stringency (figure 8.4). The rare cutting enzyme *Sfi*I appears to generate four bands detectable by GGAA.2; two strongly hybridizing bands in the region of 50-100kb, and two much fainter bands at ~150 and ~200kb. It is therefore possible that the lower bands represent alleles of one of the midisatellite loci, whereas the the upper, fainter bands represent the second locus. Alternatively, both loci may be in the size range 50-100kb, the two upper bands representing partial digestion products.

8.6 The nature of these loci in other primates

In order to explore the structure of these midisatellite loci in other primates, DNA samples from great apes and Old World monkeys were digested with *Hin*fI, subjected to agarose gel electrophoresis and the resulting Southern blot probed with GGCA.2 at high stringency (figure 8.5). As expected, the chimpanzee and gorilla showed complex banding patterns similar to those seen in humans, but surprisingly the orang-utan DNA gave no signal whatsoever, even after long exposure times. The Old World monkey DNAs either gave no signal (celebes macaque, baboon and colobus monkey), or showed 1-3 low (<2kb) molecular weight bands (rhesus macaque, iris macaque and mangabey).

8.7 Discussion

Probe GGCA.2 detects two large (50-200kb) tandem arrays in human DNA, which are highly polymorphic in internal structure. The relationship between these two midisatellites is not known; given the simple nature of the repeat unit (AGCCAAGCC) it is possible that the two sequences arose independently. Alternatively, one array may have originated from the other, perhaps via a DNA-mediated transposition

Figure 8.4. Pulsed-field gel analysis of the two midisatellite loci. Blot kindly provided by Dr. Nicola Royle. (3μg aliquots of digested DNA were electrophoresed for 26 hours at 200V through a 1.5% agarose gel in TBE at 14°. Pulse time was 25 seconds. Following electrophoresis, DNA was transferred to Hybond-N). At high stringency, GGCA.2 apparently detects two strongly hybridizing bands in the range 50-100kb and two much fainter bands at ~150 and ~200kb in DNA digested with *Sfi*I (arrowed). The varying intensity of these bands may indicate that they represent the two midisatellite loci, with two alleles of different length at each locus. Alternatively, both loci may be in the size range 50-100kb, the two upper bands representing partial digestion products.

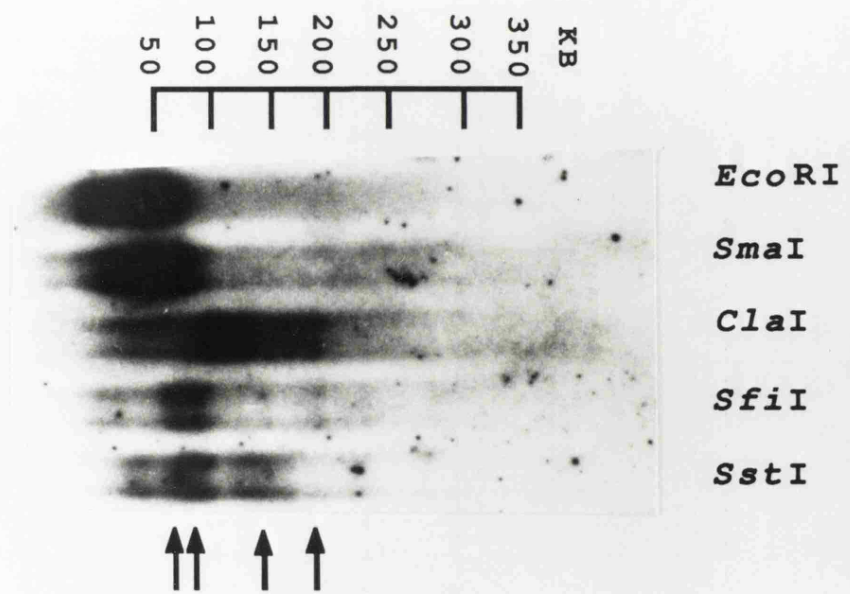
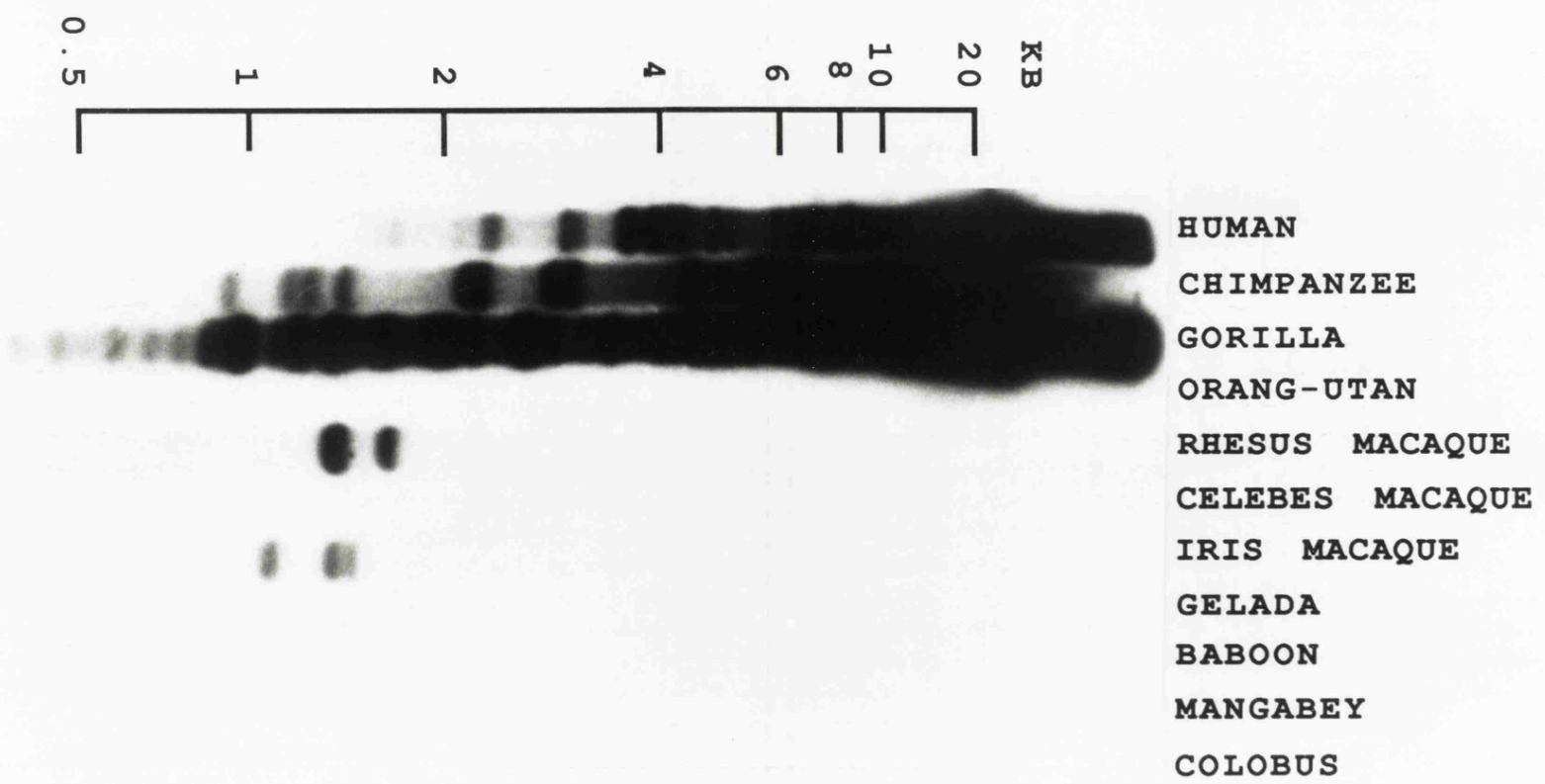

KB
350
300
250
200
150
100
50
EcoRI
SmaI
ClaI
SfiI
SstI

Figure 8.5. Cross-hybridization of GGCA.2 to *Hin*fI digested DNA from a range of Old World primates. As in man, GGCA.2 generates a complex profile comprising multiple fragments from chimpanzee and gorilla DNA at high stringency. However, no cross-hybridizing fragments are apparent in the orang-utan. Old World monkeys have very simple profiles consisting of one, two or three low molecular weight bands, or show no cross-hybridizing fragments. Experimental procedures were as for figure 8.2.


0.5
1
2
4
6
8
10
20
KB
HUMAN
CHIMPANZEE
GORILLA
ORANG-UTAN
RHESUS MACAQUE
CELEBES MACAQUE
IRIS MACAQUE
GELADA
BABOON
MANGABEY
COLOBUS

event similar to that which is thought to have given rise to a second locus detected by minisatellite probe MS29 (Wong *et al.*, 1990).

One or both midisatellite loci appear to be present in the chimpanzee and gorilla, but surprisingly no signal is obtained from orang-utan DNA, suggesting that both loci are absent, or that the repeat sequences have diverged to the extent where they are no longer detectable by GGCA.2. Absence of these arrays in the orang-utan would imply either loss along the orang-utan lineage or incredibly rapid expansion along the lineage leading to the man/chimpanzee/gorilla clade. If the latter is true, the maximum age of both arrays would be approximately 15 million years, the estimated divergence point of the orang-utan and the African apes. Support for such recent expansion comes from the profiles generated by GGCA.2 in Old World monkeys - very simple patterns consisting of one two or three low molecular weight bands, or no signal at all. It is possible that these bands represent a ground state comprising a few repeats, from which the midisatellite loci in humans, chimpanzees and gorillas has expanded, but in the absence of further evidence this remains speculation.

Similar studies of the midisatellite locus on chromosome 1 (Tynan & Hoar, 1989) gave remarkably similar results to those presented here; a large polymorphic locus in man and the African apes, but simple low molecular weight profiles in the orang-utan, gibbon and Old World monkeys. Thus these midisatellite loci may represent an extreme example of the evolutionary transience of highly polymorphic tandem arrays (see chapter 3). In view of this, the suggestion by Nakamura *et al.* (1987) that the chromosome 1 midisatellite may have an important function, such as chromosome recognition at meiotic pairing, seems highly unlikely.

CHAPTER 9

OVERALL DISCUSSION

9.1 Factors affecting variability and mutation rate at minisatellite loci

Two obvious parameters to consider with regard to the factors affecting variability (which ultimately reflects mutation rate) at minisatellite loci are repeat unit sequence and array length. The first batch of myoglobin related minisatellites to be isolated contained an almost invariant 15bp G-rich 'core' sequence embedded within each repeat unit of the tandem array, this core sequence bearing similarity to the χ recombination signal of *E. coli* and giving rise to speculation that minisatellites were eukaryotic recombination hotspots (see below and Introduction). The core sequence therefore seemed to facilitate high levels of variability. However, minisatellites subsequently isolated with both core probes and a variety of other G-rich probes, from both human DNA and other genomes, show limited homology to the core, but can still display high levels of variability (see Dover, 1989). Furthermore, AT-rich hypervariable minisatellites have also been isolated (Stoker *et al.*, 1985; Knott *et al.*, 1986). On this evidence, it would appear that repeat unit sequence plays a limited role in dictating the level of variability at minisatellite loci. However, many different mutational processes appear to be operating at minisatellites (see below), and the presence of a core sequence may promote specific mechanisms, or allow such mechanisms to operate; a view supported by the identification of core binding proteins (Collick & Jeffreys, 1990; Wahls *et al.*, 1991).

The importance of overall array length in determining the degree of variability at minisatellites is also unclear, although as a general rule of thumb, shorter minisatellites tend to be less variable. However, this is not always the case: minisatellite MS228B is far more variable than would be expected for a locus with all alleles shorter than 6kb (~350 repeat units) and most less than 3kb (Armour *et al.*, 1989). Furthermore, MS8 and MS31, despite similar allele size distributions and length of repeat units, show widely differing degrees of variability. While MS31 has far in excess of 40 alleles (David Neil, personal communication), MS8 has just 9, four of which have significant population frequencies (Wong *et al.*, 1987). This presumably reflects a lower mutation rate at MS8 which allows alleles to become common through genetic drift; higher mutation rates do not allow allele fixation but instead render the array extremely unstable and transient (see chapter 3). MS8 would therefore be expected

to be reasonably persistent in evolutionary terms. The midisatellite loci described in chapter 8 would appear to represent an extreme example of evolutionary transience, and may therefore be expected to have a relatively high mutation rate, although no data exist to confirm this. In conclusion, there seems to be no clear correlation between array length and variability or mutation rate.

In contrast, array length has clearly been shown to be an important factor in microsatellite variability (Weber, 1990), probably indicating that the mutational mechanism operating at dinucleotide repeat loci (generally thought to be DNA slippage during replication; see Levinson & Gutman, 1987) is different to that operating at minisatellites. However, perfect dinucleotide repeats are very constrained in length, possibly because large arrays accumulate base substitutions which cannot be removed quickly enough by array homogenization (see chapter 5). This leads to a third factor which may possibly influence mutation rate and variability at minisatellite loci: their location in the genome. It is known that 'compositional isochores' with differing base substitution rates exist in human DNA (Wolfe *et al.*, 1988), and it is clear that the base substitution rate around MS32 is elevated compared to the general rate in non-coding DNA (chapter 3). This might be expected to lower variability at MS32 as a result of erosion of repeat unit homogeneity, which may in turn hamper mechanisms such as sister chromatid exchange, as is thought to have occurred in the stable 5' and 3' haplotypes of variant repeat units in MS1 (chapter 3). However, the rapid expansion of MS32 over a very short evolutionary time period (≤6 million years) is unlikely to have allowed sufficient time for the accumulation of base substitutions, even with this elevated rate.

Mutation events at MS32 are polarized towards one end of the array, implying that the mutational process is influenced or governed by a neighbouring cis-acting element (Jeffreys *et al.*, 1990, 1991b). If this is a general phenomenon at minisatellite loci, it will help to explain the apparent lack of correlation between either array length or repeat unit sequence and variability and mutation rate. At the time of writing, studies involving transgenic mice harbouring MS32 repeat arrays are underway, and will hopefully give an insight into the effect of surrounding sequence

on minisatellite mutation rates (A. Jeffreys & A. Collick, personal communication).

9.2 Minisatellites as hotspots for meiotic recombination

The possible involvement of minisatellites in meiotic recombination (Jeffreys *et al.*, 1985a; see introduction) has been brought into question recently, mainly due to observations leading to the conclusion that the majority of events which generate new length minisatellite alleles are intra- rather than inter-allelic, and are likely to result from sister chromatid exchange or DNA slippage during replication (Wolff *et al.*, 1988, 1989; Jeffreys *et al.*, 1990; see introduction). The notion of minisatellites as mediators of meiotic recombination is also dealt something of a blow by the apparent lack of conservation of hypervariable arrays even between closely related species (chapter 3); it might be expected that functions as important as synapsis and meiotic recombination would be governed by DNA motifs which are reasonably well conserved between very closely related species. Furthermore, the documented expansion of at least three minisatellites from within dispersed repeat elements (Armour *et al.*, 1989; see chapter 3; Kelly *et al.*, 1989; Armour *et al.*, 1991) and their general association with dispersed and other tandemly repeated elements (Armour *et al.*, 1989) would seem to suggest that minisatellites accumulate in regions of DNA where constraints on the large scale organization of DNA sequence are relaxed, allowing retroposon insertion and tandem array expansion, possibly under the influence of cis-acting elements as described above.

Although minisatellite expansion may be the result of cis-acting DNA elements, initial expansion would still appear to be largely stochastic; MS32 is a highly variable minisatellite only in man, and the computer simulation data presented in chapter 3 imply that very few lineages will ever reach a target repeat copy number currently observed at MS32 in contemporary human populations, under a model where gains and losses of repeat units via sister chromatid exchange are equally likely, and where mutation rate is proportional to the number of repeat units. However, in the light of more recent data, this model appears to be an over-simplification. Recent experiments have demonstrated that interallelic recombination, although not a major mechanism in the

generation of new length alleles, is certainly significant at the MS32 locus, and possibly represents a 700-fold enhancement over the mean recombination frequency of the human genome (Jeffreys *et al.*, 1991b). Furthermore, there appears to be a bias in favour of gain of repeat units, and polarization of the mutation process towards one end of the array implies that mutation rate is not proportional to the number of repeat units.

Other mammalian recombination hotspots are generally poorly characterized; the most extensively studied are those of the mouse major histocompatibility complex. Four hotspots have been identified at this locus (reviewed in Steinmetz *et al.*, 1987) and two of these have been characterized at the molecular level, one within the Eβ gene and the other between the $A\beta_3$ and $A\beta_2$ genes (Steinmetz *et al.*, 1986; Uematsu *et al.*, 1986). Perhaps significantly, both hotspots are in the vicinity of a tetranucleotide repeat array; $(CAGG)_{7-9}$ in the case of the Eβ associated hotspot, and $(CAGA)_{4-6}$ in the $A\beta_3/A\beta_2$ hotspot. However, there is no direct evidence to suggest that these elements are involved in (or responsible for) the elevated level of recombination. Array length does not change following a recombination event, and in those cases where recombination products have been analysed at the DNA level, the breakpoint has been found to be upstream of the tandem array. A further candidate for promoting recombination has been identified at the $A\beta_3/A\beta_2$ hotspot; a 36bp stretch of DNA which has homology to a conserved portion of endogenous mouse retroviral LTR-like elements (Uematsu *et al.*, 1986). The frequency of recombination of plasmid-borne LTRs *in vitro* has been shown to be greatly reduced following deletion of this sequence, which interacts with two or more nuclear proteins (Edelmann *et al.*, 1989). Perhaps significantly, MS32 has expanded from within a retroviral LTR-like element (Armour *et al.*, 1989b; see chapter 3), although the MS32-associated element bears no homology to this murine element. Other known recombination hotspots reside in the pseudoautosomal region of the human sex chromosomes (reviewed in Burgoyne *et al.*, 1986) and possibly in the subtelomeric region of the long arm of the X chromosome, at the position of the fragile X site (Szabo *et al.*, 1984). The pseudoautosomal region is known to contain several tandem repeat arrays (Simmler *et al.*, 1987; Page *et al.*, 1987), and the

fragile X site has recently been characterized at the molecular level and shown to consist of a $(CGG)_n$ repeat array (Kremer *et al.*, 1991).

The association of tandem repeat arrays with recombination hotspots is compelling; however, tandem repeats are not a prerequisite for elevated levels of recombination. A 1.9kb fragment from a region of DNA upstream of the β-globin gene, thought to be a recombination hotspot, showed significant enhancement of recombination compared with a neighbouring fragment of similar size on transfection into yeast cells. Although this fragment contained a $(CA)_n$ microsatellite tandem repeat, deletion of the microsatellite had no effect on the recombinational activity (Treco *et al.*, 1985). The circumstantial association of tandemly repeated arrays with recombination hotspots, coupled with the polarity of mutation events observed at MS32 may therefore suggest that minisatellites arise and are maintained as hypervariable arrays as a result of flanking DNA motifs which act to stimulate recombination. If true, this would imply that minisatellites are the identifiable consequence of recombination hotspots, rather than being true recombination hotspots themselves.

9.3 Simple tandem repeats as genetic markers

$(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ microsatellites are potentially very useful markers for genetic linkage analysis. Although variability is low relative to minisatellites, microsatellites are a far better prospect for linkage mapping than RFLPs. It has been estimated that about 12,000 $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ repeats with PIC values of ≥ 0.5 exist in the human genome (Weber, 1990); the maximum PIC value for a diallelic RFLP is 0.375. Given their apparent random distribution (Luty *et al.*, 1990) and the ease with which they may be isolated, microsatellites should make a significant contribution to the human genetic linkage map. The ubiquity of microsatellites in eukaryotic genomes means that their potential as linkage markers is not limited to man. Microsatellites are already proving extremely useful in linkage mapping in the mouse (Love *et al.*, 1990), and have recently been used to identify two genes which confer susceptibility to insulin-dependent diabetes in mice (Todd *et al.*, 1991). Microsatellite markers have also recently been instrumental in the mapping of two genetic loci associated with blood pressure regulation in rats with hereditary hypertension (Hilbert *et al.*, 1991; Jacob *et al.*, 1991).

Identification of the gene defects responsible for such polygenic disorders in animal models should assist in unravelling the pathogenesis of the equivalent multifactorial disorders in man (see Avner, 1991). The potential of microsatellites also is being explored in several commercially important species, including poultry (Bruford & Burke, 1991) and fish (Bentzen *et al.*, 1991), with the overall aim of linking such markers to desirable quantitative traits. Furthermore, low variability at some microsatellite loci may render them useful for establishing distant relationships between, for example, populations and species (see Burke, 1991).

The most frequently used system for typing microsatellites is end labelling one of the PCR amplimers with γ-^{32}P-dATP and resolving the resulting labelled microsatellite alleles on sequencing-type denaturing polyacrylamide gels (Weber & May, 1989; Litt & Luty, 1989; Tautz, 1989). This procedure is cumbersome and time consuming, and although necessary for determination of absolute allele sizes for allele frequency database construction in, for example, a forensic context (see chapter 6), appears to be largely unnecessary for linkage analysis, as an adequate degree of resolution can frequently be achieved from high percentage agarose gels (chapter 5). Only in instances where alleles differ by a single repeat unit (2bp) are acrylamide gels required, as 4bp differences can comfortably be resolved on high percentage agarose gels (see chapter 6, figure 6.3). Resolving alleles on agarose gels simplifies and quickens microsatellite typing considerably, and removes the need for radioactive labelling. Similarly, larger and more variable tri- and tetranucleotide repeat arrays, also amenable to typing by agarose gel electrophoresis, should provide a useful source of linkage markers, although frequent association with dispersed repeats may render careful avoidance of retroposon associated arrays necessary, in order that unique sequence PCR amplimers may be designed.

In forensic science, microsatellites have one advantage over minisatellites; limited length. This allows severely degraded DNA samples, where the average single stranded DNA length may only be a few hundred nucleotides, to be typed. Unfortunately though, relatively low heterozygosities, coupled with allele frequency distributions dominated by common alleles, significantly weaken the statistical power of

microsatellites. This problem may be overcome by using more informative tetranucleotide repeat arrays, which are still typically less than a kilobase in length. However, the use of simple tandem repeat loci in forensic science is likely to be superseded by minisatellite variant repeat mapping (see introduction and chapter 4) which can be successfully applied to degraded DNA samples, providing enough information for database interrogation (Jeffreys *et al.*, 1991b).

9.4 Concluding remarks

Despite the apparent clustering of minisatellites in the proterminal regions of chromosomes (Royle *et al.*,1988; Nakamura *et al.*, 1988; Armour *et al.*, 1990), tandem repeats in general appear to be widely scattered across much of the human genome; it has been suggested that there are very few long tracts of truly unique sequence in human DNA, and that most regions are either obviously tandemly repeated, or consist of 'cryptic' tandem repeats which are scrambled and scattered with base substitutions, rendering them difficult to distinguish from true unique sequence DNA (Tautz *et al.*, 1986; Dover, 1989). Tandem repetitiveness would therefore appear to be a common rather than exceptional state for human DNA, possibly due to mutational biases which preferentially generate simple tandem repeat sequence in the absence of selection. Given that much of this tandemly repeated DNA will show length or internal variation, and coupled with novel methods for the detection of base substitutions such as single stranded conformational polymorphism (SSCP; Orita *et al.*, 1989), it seems likely that polymorphism will be detectable in almost any tract of DNA in the human genome.

References

Adams, S.W., Kaufman, R.E., Kretschmer, K., Harrison, M. & Nienhuis, A.W. (1980). A family of long reiterated DNA sequences, one copy of which is next to the human β-globin gene. *Nucl. Acids Res.* **8:** 6113-6128.

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

Allshire, R.C., Gosden, G.R., Cross, S.H., Cranston, G., Rout, D., Sugawara, N., Szostack, J.W., Fantes, P.A. & Hastie, N.D. (1989). *Nature* **332:** 656-659.

Arnheim, N. & Southern, E.M. (1977). Heterogeneity of the ribosomal genes in mice and men. *Cell* **11:** 363-370.

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

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

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

Armour, J.A.L., Crosier, M. & Jeffreys, A.J. (1991). Undetected 'null' alleles at human minisatellite loci. *Genomics* in press.

Arnott, S., Chandrasekaran, R., Birdsall, D.L., Leslie, A.G.W. & Ratcliffe, R.F. (1980). Left-handed DNA helices. *Nature* **283:** 743-745.

Avner, P. (1991). Sweet mice, sugar daddies. *Nature* **351:** 519-520.

Bains, W. (1986). The multiple origins of human Alu elements. *J. Mol. Evol.* **23:** 189-199.

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

Benton, W.D. & Davis, R.W. (1977). Screening λgt recombinant clones by hybridization to single plaques *in situ.. Science* **196:** 180-182.

Bentzen, P., Harris, A.S. & Wright, J.M. (1991). Cloning of hypervariable minisatellite and simple sequence microsatellite repeats for DNA fingerprinting of important aquacultural species of salmonids and tilapias. In *DNA fingerprinting: approaches and applications.* pp 243-262. Ed: T. Burke, G. Dolf, A.J. Jeffreys & R. Wolff; Birkhauser Verlag, Bern.

Berg, D.T., Walls, J.D., Reifel-Miller, A.E. & Grinnell, B.W. (1989). E1A-induced enhancer activity of the poly(dG-dT)·poly(dA-dC) element (GT element) and interactions with a GT-specific nuclear factor. *Mol Cell. Biol.* **9:** 5248-5243.

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

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

Bonner, T.I., O'Connell, C. & Cohen, M. (1982). Cloned endogenous retroviral sequences from human DNA. *Proc. Natl. Acad. Sci. USA* **79:** 4709-4713.

Botstein, D., White, R.L., Skolnick, M.H. & Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32:** 314-331.

Britten, R.J. & 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., Stout, D.B. & Davidson, E.H. (1989). The current source of human Alu retroposons is a conserved gene shared with Old World monkeys. *Proc. Natl. Acad. Sci. USA* **86:** 3718-3722.

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

Brown, S.D.M. & Dover, G. (1981). Organization and evolutionary progress of a dispersed repetitive family of sequences in widely separated rodent species. *J. Mol. Biol.* **150:** 441

Bruford, M.W. & Burke, T. (1991). Hypervariable DNA markers and their applications in the chicken. In *DNA fingerprinting: approaches and applications*. pp 230-242. Ed: T. Burke, G. Dolf, A.J. Jeffreys & R. Wolff; Birkhauser Verlag, Bern.

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

Bullock, W.O., Fernandez, J.M. & Short, J.M. (1987). XL1 Blue: a high efficiency plasmid transforming *rec*A *Escherichia coli* strain with beta-galactosidase selection. *Biotechniques* **5:** 376-379.

Burgoyne, P. (1986). Mammalian X and Y cross-over. *Nature* **319:** 258-259.

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

Burke, T., Hanotte, O., Bruford, M.W. & Cairns, E. (1991). Multilocus and single locus minisatellite analysis in population biological studies. In

DNA fingerprinting: approaches and applications. pp 154-168. Ed: T. Burke, G. Dolf, A.J. Jeffreys & R. Wolff; Birkhauser Verlag, Bern.

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

Chakraborty, R., Fornage, M., Gueguen, R. & Boerwinkle, E. (1991). Population genetics of hypervariable loci: analysis of PCR based VNTR polymorphism within a population. In *DNA fingerprinting: approaches and applications*. pp 127-143. Ed: T. Burke, G. Dolf, A.J. Jeffreys & R. Wolff; Birkhauser Verlag, Bern.

Chapman, B.S. & Wilson, A.C. (1982). Variable structure of IVS 1 in human and ape ζ-globin genes. *J. Cell. Biochem. (suppl.)* **6:** 257.

Chapman, B.S., Vincent, K.A. & Wilson, A.C. (1981). Extensive polymorphism and evolution in ζ-globin genes. J. *Cell. Biochem. (suppl.)* **5:** 200.

Chapman, B.S., Vincent, K.A. & Wilson, A.C. (1986). Persistence or rapid generation of DNA length polymorphism at the ζ-globin locus of humans. *Genetics* **112:** 79-92.

Charlesworth, B., Langley, C.H. & Stephan, W. (1986). The evolution of restricted recombination and the accumulation of repeated DNA sequences. *Genetics* **112:** 947-982.

Church G.M. & Gilbert, W. (1984). Genomic sequencing. *Proc. Natl. Acad. Sci. USA* **81:** 1991-1995.

Clark, J.M. (1988). Novel non-templated nucleotide addition reactions catalysed by prokaryotic and eukaryotic DNA polymerases. *Nucl Acids Res.* **16:** 9677-9686.

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

Croce, C.M. & Koprowski, H. (1974). Somatic cell hybrids between mouse peritoneal macrophages and SV40 transformed human cells. I. Positive control of the transformed phenotype by the human chromosome 7 carrying the SV40 genome. *J. Exp. Med.* **140:** 1221-1229.

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

Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T. & Schmid, C.W. (1981). Base sequence studies of 300 nucleotide renatured repeated human DNA clones. *J. Mol. Biol.* **151:** 17-33.

Denison, R.A. & Weiner, A.M. (1982). Human U1 RNA pseudogenes may be generated by both DNA and RNA mediated mechanisms. *Mol. Cell. Biol.* **2:** 815-828.

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

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., Botstein, 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. & Abrahamson, J. (1987). A genetic linkage map of the human genome. *Cell* **51:** 319-377.

Dickson, K.R., Braaten, D.C. & Schlessinger, D. (1989). Human ribosomal DNA: conserved sequence elements in a 4.3kb region downstream from the transcription unit. *Gene* **84:** 197-200.

Dodd, B.E. (1985). DNA fingerprinting in matters of family and crime. *Nature* **318:** 506-507.

Doolittle, W.F. & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284:** 601-603.

Dover, G.A. (1980). Ignorant DNA. *Nature* **285:** 618-620.

Dover, G.A. (1989). DNA fingerprints: victims or perpetrators? *Nature* **342:** 347-348.

Economou, E.P., Bergen, A.W. & Antonarakis, S.E. (1989). Novel DNA polymorphic system: variable poly A tract 3' to AluI repetitive elements. *Am. J. Hum. Genet. (suppl.)* **45:** A138.

Edelmann, W., Kroger, B., Goller, M. & Horak, I. (1989). A recombination hotspot in the LTR of a mouse retrotransposon identified in an *in vitro* system. *Cell 57: 937-946.*

Edwards, Y.H., Parkar, M., Povey, S., West, L.F., Parrington, J.M. & Solomon, E. (1985). Human myosin heavy chain genes assigned to chromosome 17 using a human cDNA clone as a probe. *Ann. Hum. Genet.* **49:** 101-109.

Edwards, Y.H., Barlow, J.H., Konialis, C.P., Povey, S. & Butterworth, P.H.W. (1986). Assignment of the gene determining human carbonic anhydrase, CA1, to chromosome 8. *Ann. Hum. Genet.* **50:** 123-129.

Evans, J.P. & Palmiter, R.D. (1991). Retrotransposition of a mouse L1 element. *Proc. Natl. Acad. Sci. USA* **88:** 8792-8795.

Feener, C.A., Boyce, F.M. & Kunkel, L.M. (1991). Rapid detection of CA polymorphisms in cloned DNA: application to the 5' region of the dystrophin gene. *Am J. Hum Genet.* **48:** 621-627.

Fisher, R.A. (1935). The detection of linkage with dominant abnormalities. *Ann. Eugen.* **6:** 187-201.

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

Flanagan, J.C., Le Franc, M.P. & Rabbitts, T.H. (1984). Mechanisms of divergence and convergence of the immunoglobulin α1 and α2 constant region gene sequences. *Cell* **36:** 681-688.

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

Gaubatz, J. & Cutler, R.G. (1975). Hybrdization of ribosomal RNA labelled to high specific radioactivity with dimethyl sulphate. *Biochemistry* **14:** 760-765.

Giblett, E.R. (1977). Genetic polymorphism in human blood. *Ann. Rev. Genet.* **11:** 13-28.

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

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

Golenberg, E.M., Giannasi, D.E., Clegg, M.T., Smiley, C.J., Durbin, M., Henderson, D. & Zurawski, G. (1990). Chloroplast DNA sequence from a Miocine *Magnolia* species. *Nature* **344:** 656-658.

Golenberg, E.M. (1991). Amplification and analysis of Miocine plant fossil DNA. *Phil Trans. R. Soc. Lond. B.* **333:** 419-427.

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

Goodman, M., Tagle, D.A., Fitch, D.H.A., Bailey, W., Czelusniak, J., Koop, B.F., Benson, P. & Slightom, J.L. (1990). Primate evolution at the

DNA level and a classification of the hominoids. *J. Mol. Evol.* **30:** 260-266.

Goodfellow, P.N., Banting, G., Trowsdale, J., Chambers, S. & Solomon, E. (1982). Introduction of a human X-6 translocation chromosome into a mouse teratocarcinoma: investigation of control of *HLA*-A, B, C expression. *Proc. Natl. Acad. Sci. USA* **79:** 1190-1194.

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

Gross, D.S. & Garrard, W.T. (1986). The ubiquitous potential Z-forming sequence of eukaryotes (dT-dGn.(dC-dA)n is not detectable in the genomes of eubacteria, archaebacteria or mitochondria. *Mol. Cell. Biol.* **6:** 3010-3013.

Gough, J.E & Murray, N.E. (1983). Sequence diversity among related genes for recognition of specific targets in DNA molecules. *J. Miol. Biol.* **166:** 1-19.

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

Hagelberg, E. & Clegg, J. (1991). Isolation and characterization of DNA from archaeological bone. *Proc. R. Soc. Lond. B.* **244:** 45-50.

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

Hamada, H., Petrino, M.G. & 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.

Hamada, H., Petrino, M.G., Kakunaga, T., Seidman, M. & Stollar B.D. (1984a). Characterization of genomic Poly(dT-dG)·Poly(dC-dA) sequences: structure, organization and conformation. *Mol. Cell. Biol.* **4:** 2610-2612.

Hamada, H., Seidman, M., Howard, B.H. & Gorman, C.M. (1984b). Enhanced gene expression by the Poly(dT-dG)·Poly(dC-dA) sequence. *Mol. Cell. Biol.* **4:** 2622-2630.

Hanahan, D. (1983). Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* **166:** 557-580.

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

Haniford, D.B. & Pulleybank, D.E. (1983). Facile transition of poly[d(TG)d(CA)] into a left-handed helix in physiological conditions. *Nature* **302:** 632-634.

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

Harris, H. & Hopkinson, D.A. (1972). Average heterozygosity per locus in man: an estimate based on the incidence of enzyme polymorphisms. *Ann. Hum. Genet.* **36:** 9-19.

Hastie, N.D. & Allshire, R.C. (1989). Human telomeres: fusion and interstitial sites. *Trends Genet.* **5:** 326-331

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

Hatlen, L. & Attardi, G. (1971). Proportion of the HeLa cell genome complementary to transfer RNA and 5s RNA. *J. Mol. Biol.* **56:** 535-553.

Hellman, L., Steen, M., Sundvall, M. & Petterson, H. (1988). A rapidly evolving region in the immunoglobulin heavy chain loci of rat and mouse: postulated role of $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ sequences. *Gene* **68:** 93-100.

Henderson, A.S., Warburton, D. & Atwood, K.C. (1972). Location of ribosomal DNA in the human chromosome complement. *Proc. Natl. Acad. Sci. USA* **69:** 3394-3398.

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

Higuchi, R., Bowman, B., Friedberger, M. Ryder, O.A. & Wilson, A.C. (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature* **312:** 282-284.

Hilbert, P., Lindpaintner, K., Beckmann, J.S., Serikawa, T., Soubrier, F., Dubay, C., Cartwright, P., De Gouyon, B., Julier, C., Takahasi, S., Vincent, M., Ganten, D., Georges, M. & Lathrop, G.M. (1991). Chromosomal mapping of two genetic loci associated with blood-pressure regulation in hereditary hypertensive rats. *Nature* **353:** 521-529.

Hill, A.V.S & Jeffreys, A.J. (1985). Use of minisatellite probes for determination of twin zygosity at birth. *Lancet* **ii:** 1394-1395.

Horai, S., Kondo, R., Murayama, K., Hayashi, S., Koike, H. & Nakai, N. (1991). Phylogenetic affiliation of ancient and contemporary humans inferred from mitochondrial DNA. *Phil. Trans. R. Soc. Lond. B.* **331:** 409-417.

Houck, C.M., Rinehart, F.P. & Schmid, C.W. (1979). A ubiquitous family of repeated DNA sequences in the human genome. *J. Mol. Biol.* **132:** 289-306.

Hulten, M. (1974). Chiasmata distribution in in the normal human male. *Hereditas* **76:** 55-78.

Jacob, H.J., Lindpaintner, K., Lincoln, S.E., Kusumi, K., Bunker, R.K., Mao, Y-P., Ganten, D., Dzau, V.J. & Lander, E.S. (1991). Genetic mapping of a gene causing hypertension in the stroke-prone hypertensive rat. *Cell* **67:** 213-224.

Jeang, K.T. & Hayward, G.S. (1983). A cytomegalovirus DNA sequence containing tracts of tandemly repeated CA dinucleotide hybridizes to highly repetitive dispersed elements in mammalian cell genomes. *Mol. Cell. Biol.* **3:** 1389-1402.

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

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

Jeffreys, A.J., Brookfield, J.F.Y. & 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. & Ponder, B.A.J. (1986). DNA 'fingerprints' and segregation analysis of multiple markers in human pedigrees. *Am. J. Hum. Genet.* **39:** 11-24.

Jeffreys, A.J. (1987). Highly variable minisatellites and DNA fingerprints. *Biochem. Soc. Trans.* **15:** 309-317.

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

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

Jeffreys, A.J., Hillel, J., Hartley, N., Bulfield, G., Morton, D.B., Wilson, V., Wong, Z & Harris, S. (1987b). The implications of hypervariable

DNA regions for animal identification: hypervariable DNA and genetic fingerprints. *Anim. Genet.* **18:** 141-142.

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

Jeffreys, A.J., Neumann, R. & 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:** 478-485.

Jeffreys, A.J., Turner, M. & Debenham, P. (1991a). 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., MacLeod, A., Tamaki, K., Neil, D.L. & Monckton, D.G. (1991b). Minisatellite repeat coding: a digital approach to DNA typing. *Nature* in press.

John, B. & Miklos, G.L.G. (1988). The eukaryote genome in development and evolution. Allen & Unwin, England.

Johnson, D. & Morgan, R. (1978). Unique structures formed by pyimidine-purine DNAs which may be four-stranded. *Proc. Natl. Acad. Sci. USA* **75:** 1637-1641.

Jones, C., Kao, F.T. & Taylor, R.T. (1980). Chromosomal assignment of the gene for folylpolyglutamate synthetase to human chromosome 9. *Cytogenet. Cell Genet.* **28:** 181-194.

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

Kan, Y.W. & 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.

Kazazian Jr, H.H., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G. & Antonarakis, S.E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332:** 164-166.

Kedes, L.H. (1979). Histone genes and histone messangers. *Ann. Rev. Biochem.* **48:** 837-870.

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

Kielty, C.M., Povey, S. & Hopkinson, D.A. (1982). Regulation of expression of liver specific enzymes. III. Further analysis of a series of rat hepatoma and human somatic cell hybrids. *Ann. Hum. Genet.* **49:** 101-109.

Kilpatrick, M.W., Klysik, J., Singleton, C.K., Zarling, D.A., Jovin, T.M., Hanau, L.H., Erlanger, B.F. & Wells, R.D. Intervening sequences in human fetal globin genes can adopt left-handed Z-helices. *J. Biol. Chem.* **259:** 7288-7274.

Kipling, D. & Cooke, H.J. (1990). Hypervariable ultra-long telomeres in mice. *Nature* **347:** 400-402.

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

Knott, T., Wallis, S., Pease, R., Powell, L & Scott, J. (1986). A hypervariable region 3' to the human apolipoprotein B gene. *Nucl. Acids Res.* **14:** 9215.

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

Krayev, A.S., Kramerov, D.A., Skryabin, K.G., Ryskov, A..P., Bayev, A.A. & Georgiev, G.P. (1980). The nucleotide sequence of the ubiquitous repetitive DNA sequence complementary to the most abundant class of mouse fold-back DNA. *Nucl. Acids Res.* **8:** 1201-1215.

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

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

La Mantia, G., Pengue, G., Maglione, D., Parnuti, A., Pascucci, A. & Lania, L. (1989). Identification of new human repetitive sequences: characterization of the corresponding cDNAs and their expression in embryonal carcinoma cells. *Nucl. Acids Res.* **17:** 5913-5921.

Lathrop, G.M., Lalouel, J-M., Julier, C. & Ott, J. (1985). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *Am. J. Hum. Genet.* **37:** 482-498.

Lawlor, D.A., Dickel, C.D., Hauswirth, W.W. & Parham, P. (1991). Ancient HLA genes from 7500 year-old archaeological remains. *Nature* **349:** 785-788.

Lehrman, M.A., Goldstein, J.L., Russell, D.W. & Brown, M.S. (1987). Duplication of seven exons in LDL receptor gene caused by Alu-Alu recombination in a subject with familial hypercholestrolaemia. *Cell* **48:** 827-835.

Leppert, M., Cavanee, W., Callahan, P., Holm, T., O'Connell, P., Thompson, K., Lathrop, G.M., Lalouel, J-M. & White, R. (1986). A

primary genetic linkage map of chromosome 13q. *Am. J. Hum. Genet.* **39:** 425-427.

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

Levinson G. & Gutman, G.A. (1987). High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucl. Acids Res.* **15:** 5323-5338.

Lewin, B. (1975). Units of transcription and translation: sequence components of heterogenous nuclear RNA and messanger RNA. *Cell* **4:** 77-93.

Lipman, D.J. & Pearson, W.R. (1985). Rapid and sensitive protein similarity searches. *Science* **227:** 1435-1441.

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

Love, J.M., Knight, A.M., McAleer, M.A. & Todd, J.A. (1990). Towards construction of a high resolution map of the mouse genome using PCR-analysed microsatellites. *Nucl. Acids Res.* **18:** 4123-4130.

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

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

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

Manuelis, L. (1978). Chromosomal localization of complex and simple repeated human DNAs. *Chromosoma* **66:** 23-32.

Martin, M.A., Bryan, T., Rasheed, S. & Khan, A.F. (1981). Identification and cloning of endogenous retroviral sequences present in human DNA. *Proc. Natl. Acad. Sci. USA* **78:** 4892-4896.

McIntosh, L.P., Grieger, I., Ecksein, F., Zarling, D.A., van de Sande, J.H. & Jovin, T.M. (1983). Left-handed helical conformation of poly[dA-m^5C)·d(G-T)]. *Nature* **304:** 83-86.

Mercier, B., Gaucher, C. & Mazurier, C. (1991). Characterization of 98 alleles in 105 unrelated individuals in the F8VWF gene. *Nucl. Acids Res.* **19:** 4800.

Miesfeld, R., Krystal, M. & 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.

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

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

Morgan, T.H. (1910). Sex-limited inheritance in *Drosophila. Science* **32:** 120-122.

Morral, N., Nunes, V., Casals, T. & Estivill, X. (1991) CA/GT microsatellite alleles within the cystic fibrosis transmembrane conductance regulator (CFTR) gene are not generated by unequal crossing over. *Genomics* **10:** 692-698.

Morton, N.E. (1955). Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7:** 277-318.

Moyzis, R.K., Buckingham, J.M., Cram, L.S., Dani, M., Deaven, L.L., Jones, M.D., Meyne, J., Ratliff, R.L. & Wu, J.R. (1988). A highly

conserved repetitive DNA sequence, $(TTAGGG)_n$, at the telomeres of human chromosmes. *Proc. Natl. Acad. Sci. USA* **85:** 6622-6626.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. & 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. & White, R. (1987b). Characterization of a human 'midisatellite' sequence. *Nucl. Acids Res.* **15:** 2537-2547.

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

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

Nicholls, R.D., Fischel-Ghodsian, N. & Higgs, D.R. (1987). Recombination at the human α-globin gene cluster: sequence features and topological constraints.*Cell* **49:** 369-378.

Noda, M., Kurihara, M. & Takano, T. (1982). Retrovirus-related sequences in human DNA: detection and cloning of sequences which hybridize with the long terminal repeat baboon endogenous virus. *Nucl. Acids Res.* **10:** 2865-2878.

Nordheim, A. & Rich, A. (1983). The sequence $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ forms left-handed Z-DNA in negatively supercoiled plasmids. *Proc. Natl. Acad. Sci. USA* **80:** 1821-1825.

Norman, C. (1989). Main case deals blow to DNA fingerprinting. *Science* **246:** 1556-1558.

O'Connell, P., Lathrop, G.M., Nakamura, Y., Leppert, M.L., Lalouel, J-M. & White, R. (1989). Twenty loci form a continuous linkage map of markers for human chromosome 2. *Genomics* **5:** 738-745.

Orgel, L.E. & Crick, F.H.C. (1980). Selfish DNA: the ultimate parasite. *Nature* **284:** 604-607.

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

Paabo, S. (1985). Molecular cloning of ancient Egyptian mummy DNA. *Nature* **340:** 465-467.

Paabo, S., Gifford, J.A. & Wilson A.C. (1988). Mitochondrial DNA sequences from a 7000-year old brain. *Nucl. Acids Res.* **16:** 9775-9787.

Page, D.C., Beiker, K., Brown, L.E., Hinton, S., Leppert, M., Lalonel, J.-M., Lathrop, M., Nystrom-Lahti, M., de la Chapelle, A. & White, R. (1987). Linkage, physical mapping and DNA sequence analysis of pseudoautosomal loci on the human X and Y chromosomes. *Genomics* **1:** 243-256.

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

Paulson, K.E., Deka, N., Schmid, C.W., Misra, R., Schindler, C.W., Rush, M.G., Kadyk, L. & Leinwand, L. (1985). A transposon-like element in human DNA. *Nature* **316:** 359-361.

Richards, R.I, Shen, Y., Holman, K., Kozman, H., Hyland, V.J., Mulley, J.C. & Southerland, G.R. (1991). Fragile X syndrome: diagnosis using highly polymorphic microsatellite markers. *Am. J. Hum. Genet.* **48:** 1051-1057.

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

Rodriguez-Campos, A., Ellison, M.J., Pe'rez-Graw, L. & Azorin, F. (1986). DNA conformation and chromatin organization of a $d(CA/GT)_{30}$ sequence cloned in SV40 minichromosomes. *EMBO. J.* **5:** 1727-1734.

Rogers, J. (1983). CACA sequences - the ends or the means? *Nature* **305:** 101-102.

Rogers J. H. (1985). The origin and evolution of retroposons. *Int. Rev. Cytol.* **93:** 187-279.

Rogers, J. (1986). The origin of retroposons. *Nature* **319:** 725.

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

Rosenberg, H., Singer, M. & Rosenberg, M. (1978). Highly reiterated sequences of simiansimiansimiansimiansimian. *Science* **200:** 394-402.

Royle, N.J., Clarkson, R.E., Wong, Z. & Jeffreys, A.J. (1987). Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. *Genomics* **3:** 352-360.

Russell, D.W., Smith, M., Cox, D., Williamson, V.M. & Young, E.T. (1983). DNA sequences of 2 yeast promoter-up mutants. *Nature* **304:** 652-654.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mulis, K.B. & Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239:** 487-491.

Saito, I. & 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.

Saffer, J.D. & Lerman, M.I. (1983). Unusual class of Alu sequences containing a potential Z-DNA segment. *Mol. Cell. Biol.* 3: 960-964.

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

Sharp, P.A. (1983). Conversion of RNA to DNA in mammals: Alu-like elements and pseudogenes. *Nature* **301:** 471-472.

Shen, M.R., Batzer, M.A. & Deininger, P.L. (1991). Evolution of the master Alu gene(s). *J. Mol. Evol.* **33:** 311-320.

Short, J.M., Fernandez, J.M., Sorge, J.A. & Huse, W.D. (1988). λZAP: a bacteriophage λ expression vector with *in vitro* excision properties. *Nucl. Acids Res.* **16:** 7583-7600.

Sibley, C.G. & Ahlquist, J.E. (1987). DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. *J. Mol. Evol.* **26:** 99-121.

Simmler, M-C., Johnsson, C., Petit, C., Rouyer, F., Vergnaud, G. & Weissenbach, J. (1987). Two highly polymorphic minisatellites from the pseudoautosomal region of the human sex chromosomes. *EMBO. J.* **6:** 963-969.

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

Singer, M.F. & Skowronski, J.(1985). Making sense out of LINES: long interspersed repeat sequences in mammalian genomes. *Trends Biochem. Sci.* **10:** 119-121.

Skowronski, J. & Singer, M.F. (1985). Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. *Proc. Natl. Acad. Sci. USA* **82:** 6050-6054.

Sidow, A., Wilson, A.C. & Paabo, S. (1991). Bacterial DNA in Clarkia fossils. *Phil. Trans. R. Soc. Lond. B.* **333:** 429-433.

Slightom, J.L., Blechl, A.E. & Smithies, O. (1980). Human fetal Gγ and Aγ globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* **21:** 627-638.

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

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

Solomon, E., Swallow, D., Burgess S. & Evans, L. (1979). Assignment of the human acid α-glucosidase gene (αGLU) to chromosome 17 using somatic cell hybrids. *Ann. Hum. Genet.* **42:** 273-281.

Solomon, E., Hiorns, L., Dalgleish, R., Tolstoshev, P, Crystal, R. & Sykes, B. (1983). Regional localization of the human α2(1) collagen gene on chromosome 7 by molecular hybridization. *Cytogenet. Cell Genet.* **35:** 64-66.

Steele, P.E., Rabson, A.B., Bryan, T. & Martin, M.A. (1984). Distinctive termini characterize two families of human endogenous retroviral sequences. *Science* **225:** 943-947.

Steffensen. D.M., Duffey, P. & Prensky, W. (1974). Localization of 5S ribosomal RNA genes on human chromosome 1. *Nature* **252:** 741-743.

Steinmetz, M., Stephan, D. & Fischer-Lindahl, K. (1986). Gene organization and recombinational hotspots in the murine major histocompatability complex. *Cell* **44:** 895-904.

Steinmetz, M., Uematsu, Y & Fischer-Lindahl, K. (1987). Hotspots of homologous recombination in mammalian genomes. *Trends Genet.* **3:** 7-10.

Stephan, W. (1986). Recombination and the evolution of satellite DNA. *Genetics* **115:** 553-547.

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

Stoker, N.G., Cheah, K., Griffin, J., Pope, F. & Solomon, E. (1985). A highly polymorphic region 3' to the human type II collagen gene. *Nucl. Acids Res.* **13:** 4613-4622.

Stoneking, M., Hedgecock, D., Higuchi, R.G., Vigilant, L. & Erlich, H.A. (1991). Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotides. *Am J. Hum. Genet.* **48:** 370-382.

Stringer, J. (1982). DNA sequence homology and chromosomal deletion at a site of SV40 integration. *Nature* **296:** 363-366.

Stringer, J. (1985). Recombination between poly[d(GT)·d(CA)] sequences in simian virus 40-infected cells. *Mol. Cell. Biol.* **5:** 1247-1259.

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.

Sun, L., Paulson, K.E., Schmid, C.W., Kadyk, L. & Leinwand, L. (1984). Non-alu family interspersed repeats in human DNA and their transcriptional activity. *Nucl. Acids Res.* **12:** 2699-2690.

Swallow, D.M., Povey, S., Parkar, M., Andrews, P.W., Harris, H., Pym, B. & Goodfellow, P. (1986). Mapping of the gene coding for the liver/bone/kidney isozyme of alkaline phosphatase to chromosome 1. *Ann. Hum. Genet.* **50:** 229-235.

Sykes, B. & Solomon, E. (1978). Assignment of a type 1 collagen structural gene to human chromosome 7. *Nature* **272:** 548-549.

Szabo, P., Purello, M., Rocchi, M., Archidiacono, N., Alhadeff, B., Filippi, G., Toniolo, D., Martini, G., Luzzatto, L. & Siniscalo, M. (1984). Cytological mapping of the human glucose-6-phosphate dehydrogenase gene distal to the fragile-X site suggests a high rate of

meiotic recombiation across this site. *Proc. Natl. Acad. Sci. USA* **81:** 7855-7859.

Tabor, S. & Richardson, C.C. (1987). DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl. Acad. Sci. USA* **84:** 4767-4771.

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

Tautz, D. & Renz, M. (1984b). Simple sequences of *Drosophila virilis* isolated by screening with RNA. *J. Mol. Biol.* **172:** 229-235.

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

Tautz, D. (1989). Hypervariability of simple sequences as a general source of polymorphic DNA markers. *Nucl. Acids Res.* **17:** 6463-6471.

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

Thomas, R.H., Schaffner, W., Wilson, A.C. & Paabo, S. (1989). DNA phylogeny of the extinct marsupial wolf. *Nature* **340:** 465-467.

Todd, J.A., Aitman, T.J., Cornall, R.J., Ghosh, S., Hall, J.R.S., Hearne, C.M., Knight, A.M., Love, J.M., McAleer, M.A., Prins, J-B., Rodrigues, N., Lathrop, M., Pressey, A., DeLarato, N.H., Peterson, L.B. & Wicker, L.S. (1991). Genetic analysis of autoimmune type 1 diabetes mellitus in mice. *Nature* **351:** 542-547.

Treco, D., Thomas, B. & Arnheim, N. (1985). Recombination hotspots in the human b-globin gene cluster: meiotic recombination of human DNA fragments in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **5:** 2629-2638.

Tyler-Smith, C. & 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.

Tynan, K.M. & Hoar, D.I. (1989). Primate evolution of a human chromosome 1 hypervariable repetitive element. *J. Mol. Evol.* **28:** 212-219.

Uematsu, Y., Kiefer, H., Schulze, R., Fischer-Lindahl, K. & Steinmetz, M. (1986). Molecular characterization of a meiotic recombinational hotspot enhancing homologous equal crossing-over. *EMBO. J.* **5:** 2123-2129.

Ullu, E. & Tschudi, C. (1984). Alu sequences are processed 7SL RNA genes. *Nature* **312:** 171-172.

Van Heyningen, V., Bobrow, M., Bodmer, W.F., Gardiner, S.E., Povey, S. & Hopkinson, D.A. (1975). Chromosome assignment of some human enzyme loci: mitochondrial malate dehydrogenase to 7, mannose phosphate isomerase and pyruvate kinase to 15 and probably, esterase D to 13. *Ann. Hum. Genet.* **38:** 295-303.

Vanin, E.F. (1984). Processed pseudogenes. Characteristics and evolution. *Biochem. Biophys. Acta* **782:** 231-241.

Varesco, L., Thomas, H.J., Williams, S., Fennell, S.J., Hockey, A., Searle, S., Bodmer, W.F., Frischauf, A-M. & Solomon, E. (1989). Clones from a deletion encompassing the adenomatous polyposis coli gene (APC): Human Gene Mapping 10. *Cytogenet. Cell Genet.* **51:** 1098.

Vassart, G., Georges, M., Monsieur, R., Brocas, H., Lequarre, A.S. & 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. *Nucl. Acids Res.* **17:** 7623-7630.

Vasseur, M., Candamine, H. & Duprey, P. (1985). RNAs containing B2 repeated sequences are transcribed in the early stages of mouse embryogenesis. *EMBO J.* **4:** 1749

Vidaud, M., Vidaud, D., Siguret, V., Lavergne, J.M. & Goossens, M. (1988). Mutational insertion of an Alu sequence causes haemophilia B. *Am. J. Hum. Genet.* **45:** A226.

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

Wahls, W.P., Wallace, L.J. & Moore, P.D. (1990). Hypervariable minisatellite DNA is a hotspot for homologous recombination in human cells. *Cell* **60:** 95-103.

Wahls, W.P., Swenson, G. & Moore, P.D. (1991). Two hypervariable minisatellite DNA binding proteins. *Nucl. Acids Res.* **19:** 3269-3274.

Walmsley, R.M., Szostak, J.W. & Petes, T.D. (1983). Is there left-handed DNA at the ends of yeast chromosomes? *Nature* **302:** 84-86.

Walsh, J.B. (1989). Persistence of tandem arrays: implications for satellite and simple sequence DNA. *Genetics* **115:** 553-567.

Waye, J.S. & Willard, H.F. (1986). Structure, organization and sequence of alpha-satellite DNA from human chromosome 17: evidence for evolution by unequal crossing-over and an ancestral pentamer repeat shared with the human X chromosome. *Mol. Cell. Biol.* **6:** 3156-3165.

Waye, J.S., Durfy, S.J., Pinkel, D., Kenwick, S., Patterson, M., Davies, K.E. & Willard, H.F. (1987). Chromosome-specific alpha-satellite DNA from human chromosome 1: hierarchical structure and genomic organization of a polymorphic domain spanning several hundred kilobase pairs of centromeric DNA. *Genomics* **1:** 43-51.

Weber, F., de Villiers, J. & Schaffner, W. (1984). An SV40 'enhancer trap' incorporates exogenous enhancers or generates enhancers from its own sequences. *Cell* **36:** 983-992.

Weber, J.L. & 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. (1990). Informantiveness of $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ polymorphisms. *Genomics* **7:** 524-530.

Weber, J.L., Kwitek, A.E. & May, P.E. (1990a). Dinucleotide repeat polymorphism at the D15S87 locus. *Nucl. Acids Res.* **18:** 4640.

Weber, J.L., Kwitek, A.E. & May, P.E. (1990b). Dinucleotide repeat polymorphism at the D1S103 locus. *Nucl. Acids Res.* **18:** 2199.

Wetton, J.H., Carter, R.E., Parkin, D.T., & Walters, D. (1987). Demographic study of a wild house sparrow population by DNA fingerprinting. *Nature* **327:** 147-149.

Willard, C., Wong, E., Hess, J.F., Shen, C.J., Chapman B., Wilson, A.C. & Schmid, C.W. (1985). Comparison of human and chimpanzee ζ-1 globin genes. *J. Mol. Evol.* **22:** 309-315.

Willard, H.F. & Waye, J.S. (1987). Hierarchical order in chomosome-specific alpha satellite DNA. *Trends Genet.* **3:** 192-198.

Winship, P.R. (1989). An improved method for directly sequencing PCR amplified material using dimethyl sulphoxide. *Nucleic Acids Res.* **17:** 1266.

Wolfe, K.H., Sharp, P.M. & Li, W. (1989). Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283-285.

Wolff, R., Nakamura, Y. & White, R. (1988). Molecular characterization of a spontaneously generated new allele at a VNTR locus: no exchange of flanking sequence. *Genomics* **3:** 347-351.

Wolff, R.K., Plaetke, R., Jeffreys, A.J., & White, R. (1989). Unequal crossing-over between homologous chromosomes is not the major

mechanism involved in generation of new alleles at VNTR loci. *Genomics* **5:** 382-384.

Wolff, R., Nakamura, Y., Odelberg, S., Shiang, R & White, R. (1991) Generation of variability at VNTR loci in human DNA. In *DNA fingerprinting: approaches and applications.* pp 20-38. Ed: T. Burke, G. Dolf, A.J. Jeffreys & R. Wolff; Birkhauser Verlag, Bern.

Wong, Z., Wilson, V., Patel, I., Povey, S., & Jeffreys, A.J. (1987). Characterization of a panel of highly variable minisatellites cloned from human DNA. *Ann. Hum. Genet.* **51:** 269-288.

Wong, Z., Wilson, V., Jeffreys, A.J. & Thein, S.L. (1986). Cloning a selected fragment from a human DNA 'fingerprint': isolation of an extremely polymorphic minisatellite. *Nucl. Acids Res.* **14:** 4605-4616.

Wong, Z., Royle, N.J. & Jeffreys, A.J. (1991). A novel human DNA polymorphism resulting from transfer of DNA from chromosome 6 to chromosome 16. *Genomics* **7:** 222-234.

Wyman, A. & White, R. (1980). A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* **77:** 6754-6758.

Wyman, A.R., Wolfe, L.B. & Botstein, D. (1985). Propogation of some human DNA sequences in bacteriophage vectors requires mutant *Escherichia coli* hosts. *Proc. Natl. Acad. Sci. USA* **82:** 2880-2884.

Wyman, A.R., Mulholland, J. & Botstein, D. (1986). Oligonucleotide repeats involved in the highly polymorphic locus D14S1. *Am. J. Hum. Genet.* **39:** A226.

Yanisch-Perron, C., Viera, J., & Messing, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33:** 103-119.

Young, B.D., Hell, A. & Birnie, G.D. (1976). A new estimate of human ribosomal gene number. *Biochem. Biophys. Acta* **454:** 539-548.