THE ISOLATION AND CHARACTERIZATION OF HUMAN MINISATELLITE LOCI

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

by

John Anthony Armour, M.A.(*Cantab.*),B.M.,B.Ch.
Department of Genetics
University of Leicester

August 1990

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

J.A.Armour - *Isolation and characterization of human minisatellites.*

Hypervariable minisatellite regions of human DNA are of considerable interest, not only as highly informative genetic systems, but also as intermediately sized regions of tandem repetition. Methods for the isolation of minisatellite loci from the human genome have been investigated, and 23 new hypervariable loci cloned from an ordered array Charomid library. This method not only allows very efficient isolation of human minisatellites, but can also be used to observe the degree of overlap between multi-locus DNA fingerprinting probes. The 23 new loci have a mean heterozygosity level of 71%, and have been characterized and mapped in the genome. The genomic disposition of human minisatellites has been analysed by investigation of cloned examples. The minisatellites studied show a strong tendency to cluster near the ends of chromosomes, and sequence analysis demonstrates a significant excess of dispersed repeat elements in the DNA flanking human minisatellites. Minisatellite variant repeat (MVR) mapping has also been used to investigate the internal structure of minisatellite alleles.

Somatic allele length mutation events have been demonstrated in DNA from colorectal adenocarcinomas, and the mutations observed show many features of general similarity to germline mutation events. A series of human breast tumours has been screened for somatic change, using both multi-locus DNA fingerprinting probes and single-locus minisatellite probes. Somatic change in breast cancers is much less frequent than in colorectal tumours, but some allele losses and mutations have been defined, including a highly unusual mutation, which may be the result of a minisatellite transposition event. Finally, evolution at minisatellite loci has been studied, both by examination of allelic states in current human populations, as well as comparison with non-human primates.

Acknowledgements

My thanks go out to large numbers of people, without whose generous help some or all of this would not be possible. The lion's share goes to Alec for his finely calculated combination of close supervision and *laisser faire*, and the transmission of an enthusiasm and energy stringently controlled by the experimental facts.

To Anita, for her toleration and encouragement, and to Pob, for his critical vetting of the manuscript, go love and thanks.

Many others have been helpful in the struggle; while I name names, I hope those omitted will infer that they have somehow simply slipped my mind. They are: from the Department of Genetics; Nicola Royle, Raymond Dalgleish, Zilla Wong, Andy Collick, Robert Kelly, Ross Hawkins, Esther Signer, Ian Gray, Mark Gibbs, Darren Monckton, Annette MacLeod, Vicky Wilson, Rita Neumann, Ila Patel, and Maureen Hill. Special mention is reserved for Moira Crosier, whose technical help in the last six months of this work has been as uniformly excellent as it has been useful.

Elsewhere, help, advice and collaboration have come from Noreen Murray, Bill Brammar, Doug Higgs, Sue Povey, Steve Jeremiah, Sue Malcolm, Swee Lay Thein, Martin Fey, Ian Talbot, Rosemary Walker, Jenny Varley, Kristin Anderson and Bruce Ponder.

I am grateful to Ian Ridell and the Central Photographic Unit for their expert help in the preparation of figures, and to Barry Holland and Peter Williams for the use of the Department's facilities.

The work described in this thesis was carried out during the tenure of an MRC Training Fellowship.

A note on structure

The centre of gravity of this thesis has been deliberately shifted towards the middle "results and discussion" chapters (numbers 3, 4 and 5). Accordingly, while the initial introductory chapter and final discussion chapter may be a little shorter, introductory remarks and discussion appropriate to each of chapters 3, 4 and 5 are placed with the experimental results, so that, to a large extent, each of these chapters may be read as a self-contained unit.

Abbreviations

ATP	Adenosine 5′-triphosphate
BCIG	5-bromo-4-chloro-3-indolyl-β,D-galactoside
CEPH	Centre d’Etude du Polymorphisme Humain (Paris)
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
HEPES	N-[2-hydroxyethyl]piperazine-N′-[2-ethanesulphonic acid]
IPTG	Isopropyl-1-thio-β,D-galactoside
MVR	Minisatellite Variant Repeat
PCR	Polymerase Chain Reaction
PEG	Polyethylene glycol
SSPE	Saline sodium phosphate/EDTA [150mM sodium chloride, 10mM sodium phosphate,1mM EDTA pH7.7]
SSC	Saline sodium citrate [150mM sodium chloride, 15mM sodium citrate, pH 7.0]
SDS	Sodium dodecyl sulphate
TEMED	N,N,N′,N′-tetramethyl-ethylenediamine
Tris	Tris-(hydroxymethyl)-methylamine [2-amino-(2-hydroxymethyl)-propan-1,3-diol]

Publications arising from this work

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

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

Varley,J.M., Armour,J., Swallow,J.E., Jeffreys,A.J., Ponder,B.A.J., T'Ang,A., Fung,Y-K.T., Brammar,W.J. and Walker,R.A. (1989). The retinoblastoma gene is frequently altered leading to loss of expression in primary breast tumours. *Oncogene* 4,725-729.

Jeffreys,A.J., Wong,Z., Wilson,V., Patel,I., Neumann,R., Royle,N.J. and Armour,J.A.L.(1989). Applications of multilocus and single-locus minisatellite DNA probes in forensic medicine. In Banbury Report 32: DNA Technology and Forensic Sciences (eds. J. Ballantyne, G.Sensabaugh and J.Witkowski; Cold Spring Harbor Laboratory Press,1989) pp.283-295.

Armour,J.A.L, Povey,S., Jeremiah,S. and Jeffreys,A.J.(1990). Systematic cloning of human minisatellites from ordered array Charomid libraries. *Genomics* (in the press).

Submitted

Hannotte,O., Burke,T., Armour,J.A.L and Jeffreys,A.J.(1990). Hypervariable minisatellite sequences in the peafowl *Pavo cristatus*. (to *Genomics*)

Malcolm,S., Nicholls,M., Clayton Smith,J., Robb,S., Armour,J.A.L., Jeffreys,A.J. and Pembrey,M.E.(1990). Angelman syndrome can result from uniparental paternal isodisomy. (to *Nature*)

CONTENTS

CHAPTER 1

INTRODUCTION

I am too much of a sceptic to deny the possibility of anything

T.H.Huxley

Summary

The background to the work on human minisatellites presented in this thesis is outlined with reference to the general structural analysis of the human genome, and to the exploitation of DNA polymorphisms in human genetic analysis. Finally, the starting-points for this work are summarised.

1.1 *GENOMIC ANALYSIS*

1.1.1 *Genome complexity and sequence redundancy*

The human genome appears to be highly uneconomical with its use of genetic material; the approximately 3 x 10^9 base pairs present in a human haploid genome appear to be about two orders of magnitude more than is minimally required to encode the gene products observed (Vogel,1964). The nature of this surplus DNA complement is not only part of a detailed description of the structure of the genome, but may be important in deducing the evolutionary events leading to the structure of the modern human genome.

About half of the excess DNA consists of single-copy DNA sequences, either interrupting coding sequence as introns (Jeffreys and Flavell,1977), or in the large tracts of DNA between coding sequences. The remaining DNA is present in more than one copy in the genome, as shown by the analysis of the reannealing of denatured DNA fragments (Britten and Kohne,1968). This repetitive fraction is discussed further below, and contains both tandemly reiterated sequences and sequences present in multiple copies dispersed around the genome.

1.1.2 *Dispersed repeats*

The much faster reannealing kinetics of 13kb human DNA fragments compared with 600bp fragments (Schmid and Deininger,1975) suggested that many of the larger fragments reannealed more quickly because they contained a short, highly abundant dispersed repeat sequence.

Many of these short interspersed repeats showed significant sequence similarity with dispersed repeats in rodents and primates (Schmid and Jelinek,1982). This most abundant short interspersed repeat element, the Alu element, is about 300bp long, and is present in about 500,000 copies, thereby accounting for some 5% of the entire human genome. The presence of a polyadenylate tract at one end, together with short flanking direct repeats at the insertion site, are features which suggest that these elements may have transposed via ribonucleic acid intermediates.

Many other dispersed repetitive elements have been described in human DNA. These include the L1 (Kpn) elements, which also have structural features suggestive of retrotransposition, and which may contain an open reading frame involved in the transposition mechanism (Demers *et al.*,1986). Other classes of human dispersed repeat include the related "O" and "THE" elements (Sun *et al.*,1984;Paulson *et al.*,1985), as well as a growing class of dispersed repeats defined as such by virtue of multiple human entries in DNA sequence databases (see, for example, Donehower *et al.*,1989).

1.1.3 *Tandem repeats*

Of the tandemly repeated DNA in the human genome, a considerable fraction is present at sufficiently high copy number to be distinguishable on density gradient centrifugation as fractions of significantly distinct base composition from bulk human DNA (Miklos and John,1979). This highly abundant "satellite" DNA appears to be preferentially localised in heterochromatic chromosomal regions, and shows strong sequence homogeneity within a species, suggesting concerted evolution resulting from homogenization mechanisms within and between chromosomes (Dover,1982).

One well-characterized satellite (the "alphoid") has sequence similarity to a major satellite from the African Green Monkey (Waye and Willard,1986). In humans, it is the dominant sequence in centromeric heterochromatin, and exists in a number of chromosome-specific higher-order repeats.

Many other tandemly repeated sequences are known to exist in the human genome. These include the tandemly-repeated ribosomal RNA genes which account for much of the short arms of the acrocentric chromosomes, and constitute the nucleolar organizing centre. They are organized as 43kb repeats (Arnheim and Southern,1977), and are present in 50-200 copies per haploid genome (Young *et al.*,1976). The telomeres of human chromosomes consist of tandem arrays of a 6bp TTAGGG repeat unit. The terminal repeats are not templated, but synthesised by a specialised "telomerase" enzyme, which has been purified from some ciliates, including *Tetrahymena* (Greider and Blackburn,1989), and counteracts the shortening at the ends of linear chromosomes otherwise consequent upon the replication of one DNA strand by discontinuous synthesis (Watson,1972).

The number of loci containing "minisatellite" arrays (Jeffreys *et al.*,1985a), consisting of tandem repeats of short (9-100bp) repeat units, is still unknown, but the available data suggest that there may be more than 1500 (Jeffreys,1987). Similarly, there appears to be a very large number of simple dinucleotide repeats, of which the commonest appears to be $(AC)_n$ (n=15-30), of which there may be as many as 100,000 interspersed examples (Weber and May,1989).

1.2 *GENETIC ANALYSIS*

1.2.1 *Linkage analysis and linkage mapping*

The linear ordering of genes on the chromosomal DNA of eukaryotes can be determined by linkage analysis. The cosegregation of two markers due to linkage can be quantified by comparing the probability of linkage at a given recombination fraction (θ) with the probability of obtaining the same results under the null hypothesis that $\theta=0.5$. The support provided by segregation data for a putative linkage is conventionally expressed as the $\log_{10}$ of the odds in favour of linkage (at recombination fraction θ) over no linkage. This log-of-the-odds score is usually abbreviated to "LOD", and given the symbol z (Morton,1962).

The pioneering work in linkage mapping, particularly that on *Drosophila* by T.H.Morgan (1910), used carefully-designed crosses between pure-bred mutant stocks. In human linkage studies, however, such selected crosses are not available, and data are gathered by analysis of established pedigrees for the segregation of those polymorphic loci which happen to be informative in those pedigrees. In one important branch of linkage analysis, one polymorphism may be an inherited disorder. However, establishment of a detailed human linkage map requires not only many polymorphic markers spread sufficiently evenly, but also that those markers be variable enough for a high proportion of individuals to be genetically informative.

The ability to distinguish the two alleles at a locus is also of use in other analyses in human genetics. Thus, for example, the loss of an allele from a tumour (Hansen and

Cavenee,1987) is only convincingly demonstrable if the two alleles are distinguishable; such analyses are made much more efficient by the use of highly informative polymorphisms. More recently (*v.i.*), extremely informative markers have made a number of other analyses feasible, including individual identification in forensic work (Gill *et al.*,1985), parentage testing (Jeffreys *et al.*,1985c) and the determination of twin zygosity (Hill and Jeffreys,1985).

1.2.2 *Human polymorphism*

Among the first human biochemical polymorphisms described were many which derived from expressed sequences, the polymorphism often being detected as a protein electrophoretic or immunological variant. While many of these variants are of considerable interest in their own right, for example the haemoglobin variants (Weatherall and Clegg,1976;Lehmann and Kynoch,1976), they have also been of value in providing the first generation of polymorphic markers for human genetic analysis.

Many such systems have now been defined, including no fewer than 21 commonly polymorphic blood group antigens (Race and Sanger,1975). Very many of these, however, are of limited variability, with only two or three allelic states, and are thus of relatively low informativeness in genetic analyses. One important exception is the human leucocyte antigen (HLA) system (Albert *et al*.1984). The extreme variablility of this system, which mediates the rejection of allografts in tissue transplantation, is well documented. For example, the class I antigens are encoded by three tightly linked loci (A,B and C) on human chromosome 6. Each of these has a large number (from 9 at C to more than 30 at B) of allelic states, many of which are of low population frequency. This results in extreme variability of HLA haplotypes to give a highly variable and informative genetic system.

1.2.3 *DNA polymorphisms*

The advent of recombinant DNA technology allowed the investigation of variation in the human genome at the DNA sequence level. DNA sequence variants which created or destroyed sites for restriction endonucleases could be detected by Southern blot hybridization, and early studies at the human β-globin cluster suggested that such restriction fragment length polymorphisms could be very widespread in the human genome (Kan and Dozy,1978;Jeffreys,1979). However, although such restriction site dimorphisms have been of great value in the establishment of human genetic linkage maps (Donis-Keller *et al.*,1987), their utility is limited by their modest informativeness; a dimorphic system in Hardy-Weinberg equilibrium can have a frequency of heterozygotes of at most 50%.

It was soon clear, however, that a number of human loci were much more informative than could be afforded by restriction site dimorphisms. The first to be described was a highly polymorphic sequence, initially isolated as a random single-copy sequence from a human genomic library. This detected a locus at which there was a large number of allelic states, and at which most people in the population were heterozygotes (Wyman and White,1980). Other examples of multiallelic polymorphic loci were soon discovered, including loci 3′ to the human α-globin gene (Higgs *et al.*,1981), in a ζ-globin intron and between the ζ and pseudo-ζ globin genes (Goodbourn *et al.*.1983), and in the non-coding DNA at the insulin (Bell *et al.*,1982) and H-*ras* (Capon *et al.*,1983) genes. The common feature of these systems was the presence of a length-variable tandemly-repeated region, at which allele length was determined by the number of tandem repeats. At many of these loci, the existence of a large number of rare alleles resulted in an extremely informative locus.

The most informative of these loci are of exceptional value in human genetics. Since nearly all the population will be heterozygous at these loci, they provide a quantum leap in the efficiency of linkage mapping and other analyses over dimorphic systems. Furthermore, previously impracticable applications, such as individual identification in forensic work (Wong *et al.*,1987) and the affected sib-pair method for linkage analysis (Suarez *et al.*,1978), become feasible, and mutation and evolutionary processes may be directly studied at the most unstable loci (Jeffreys *et al.*,1988a,1990a).

The analysis and isolation of hypervariable minisatellites was given great impetus by the discovery that some tandemly repeated probes, derived originally from a tandem repeat sequence in the first intron of the human myoglobin gene, detected a large number of highly polymorphic loci in human DNA (Jeffreys *et al.*,1985a). Each of the two most useful probes found, 33.6 and 33.15, usually detects between 10 and 20 multiallelic polymorphic loci in human DNA (Jeffreys et *al.*,1986), and the composite profile obtained by detecting them all simultaneously, the "DNA fingerprint", is so variable as to be individual-specific (Jeffreys *et al.*,1985b). This individual-specificity, combined with the simple inheritance of the loci, led to a wide variety of applications, including the determination of parenthood and family relationships (Jeffreys *et al.*,1985c), forensic testing (Gill *et al.*,1985) and the determination of twin zygosity (Hill and Jeffreys,1985). Furthermore, the loci detected are genetically dispersed around the genome (Jeffreys *et al*.1986), and allow, in a single test, a screen at a large number of loci for linkage to autosomal dominant inherited disease (Jeffreys *et al.*,1986) and for somatic change in tumour tissues (Thein *et al.*,1987).

While DNA fingerprinting has phenomenal resolving power in many analyses (Jeffreys *et al.*,1985c), it has the considerable disadvantage that the loci detected are anonymous, in the sense that the locus contributing any one band on a DNA fingerprint cannot be deduced without using cloned locus-specific probes to study the locus singly. Furthermore, the isolation of individual minisatellite loci not only allows the examination of segregation and mutation at minisatellites, but is a prerequisite for the application of PCR technology to the study of minisatellites (Jeffreys *et al.*,1988b,1990a).

The fact that DNA fingerprinting probes detected a large number of polymorphic loci led to isolation of these loci on a much larger scale than hitherto, and is described in more detail in section 3.1. However, the generation of highly informative markers by screening libraries of human DNA with DNA fingerprinting probes (Wong *et al.*,1986,1987) or G-rich oligonucleotides (Nakamura *et al.*,1987a,1988b) has provided a fund of genetic markers which have had a catalytic effect on the establishment of linkage maps (Nakamura *et al.*,1988a; O'Connell *et al.*,1988) and other analyses (Solomon *et al.*,1987).

Other multiallelic length variable loci exist in the human genome, which operate on a smaller scale. "Microsatellite", or poly-AC tracts consist of short (30-60bp) stretches of dinucleotide repeats (usually $[AC]_n$), at which alleles differ in the number of dinucleotide repeats (Weber and May,1989;Litt and Luty,1989). These may be typed by PCR amplification and resolution of alleles on polyacrylamide gels. A similar method may be used to resolve alleles at some polyadenylate tracts of Alu elements, some of which show length variation (Economou *et al.*,1990). While these short length-variable regions are sometimes of high informativeness and are undoubtedly very common in human DNA, their translation into useful data for genetic analysis will depend chiefly on the practical ease and reproducibility with which genotypes can be obtained.

1.3 *POINTS OF DEPARTURE*

1.3.1 *Isolation of minisatellite loci*

At the inception of this work, progress had been made in using DNA fingerprinting probes in the cloning of individual hypervariable loci from human DNA. In the studies of Wong *et al.* (1986,1987) six extremely variable minisatellites were isolated from a λ library of human DNA. Nakamura *et al.* (1987a,1988b) had been successful in cloning larger numbers of loci from cosmid libraries by hybridization screening with G-rich oligonucleotides, although the loci isolated were considerably less variable than those isolated by Wong *et al.* Thus while much headway had been made in the isolation of minisatellite loci, it was clear that many of the most variable loci detected by DNA fingerprinting probes had yet to be isolated. Chapter 3 describes the development of strategies for the isolation of minisatellite loci from human DNA.

1.3.2 *Genomic anatomy*

Minisatellites are of interest not only because of their uses in genetic analyses, but also for what they may tell us about the evolution of the human genome. Evidence had been presented that cloned minisatellite loci cluster near the ends of chromosomes, as detected by *in situ* hybridization and restriction mapping (Royle *et al.*,1988). Published linkage maps which include highly informative minisatellites confirm their tendency to appear near the ends of linkage maps (Nakamura *et al.*,1988a). Chapter 4 documents the structural analysis of cloned minisatellites to clarify their genomic disposition.

1.3.3 *Mutation and evolution*

The high population variability at minisatellite loci is maintained by a high neutral mutation rate to new length alleles, as may be demonstrated directly by pedigree analysis (Jeffreys *et al.*,1988a). Evidence for somatic mutation had also been presented from DNA fingerprinting analysis of human tumours (Thein *et al.*,1987). Chapter 5 extends these observations to the analysis of somatic change at individual minisatellite loci in human tumours, and develops initial approaches to the investigation of evolution at minisatellite loci.

CHAPTER 2

MATERIALS AND METHODS

Wovon man nicht sprechen kann, darüber muss man schweigen.

Whereof one cannot speak, one must pass over in silence.

Wittgenstein

2.1 *MATERIALS*

2.1.1 *Chemicals*

Chemicals for general work were supplied by Fisons (Loughborough) or BDH (Poole). Media for bacterial growth were from Oxoid, Basingstoke, except for tryptone and yeast extract which were from Difco, East Molesley. Antibiotics, bovine serum albumin, HEPES, IPTG, polyethylene glycol 6000, agarose, Ficoll 400, TEMED and DTT were from Sigma Chemical Company, Poole. BCIG was from Anglian Biotechnology, Colchester. Acrylamide was from Serva (Heidelberg) and N,N'-methylene-bisacrylamide was from Uniscience (Cambridge). Marvel dried skimmed milk was purchased from Sainsbury's.

Deoxyribonucleotides, dideoxyribonucleotides and synthetic hexadeoxyribonucleotides for oligo-labelling (section 2.2.3) were from Pharmacia (Milton Keynes). Synthetic oligonucleotide primers for PCR were prepared by John Keyte (Department of Biochemistry). Radiochemicals were supplied by Amersham International.

2.1.2 *Enzymes*

Restriction endonucleases were from Gibco-BRL (Paisley), New England Biolabs (via CP Laboratories, Bishop's Stortford) or Boehringer Corporation (Lewes). DNA polymerase I (Klenow fragment), T4 polynucleotide kinase and T7 DNA polymerase were from Pharmacia. DNA polymerase from *Thermus aquaticus* was supplied by Amersham. T4 DNA ligase and AMV reverse transcriptase were from Gibco-BRL. Calf intestinal alkaline phosphatase was from Boehringer. Proteinase K, Ribonuclease A and Pronase were supplied by Sigma.

2.1.3 *Biological materials*

2.1.3.1 E.coli *strains*

The strains of *E.coli* used in this work are listed in Table 2.1. NM554 was kindly provided by Prof.Noreen Murray, University of Edinburgh. Strain FBXL5 was made to provide a *rec*$^+$ host for growth of M13 phage, but bearing a positively selectable F′ marker. A rifampicin resistant mutant of JM101 (Table 2.1) was isolated, and used as the recipient in a conjugation experiment with XL1-Blue (Table 2.1) as donor. Transconjugants were selected on media supplemented with both tetracycline and rifampicin, one of which was FBXL5.

2.1.3.2 In vitro *packaging*

In vitro packaging of λ and Charomid ligations was carried out using Gigapack Plus (Stratagene), according to the manufacturers' instructions.

2.1.3.3 *Human tissues*

Human blood was collected in potassium EDTA or SSC and stored at -20°C, and DNA prepared as described in section 2.2.1.2. Solid human tissues (breast carcinomas and ileal mucosa) were transported in liquid nitrogen, and stored frozen at -70°C. DNA was prepared from human tissues as described in section 2.2.1.

2.1.4 *Cloning vectors and libraries*

Routine subcloning was carried out in the vectors pUC13, pUC18 and pUC19 (Vieira and Messing,1982), M13mp18 and mp19 (Yanisch-Perron *et al.*,1985), and pBluescriptII SK$^+$ and KS$^+$ (Stratagene).

A cosmid library was kindly provided by Dr.Brandon Wainwright, Department of Biochemistry, St.Mary's Hospital Medical School, Paddington. This was an amplified library of

Table 2.1 *E.coli* strains used in this work

Strain	Genotype	Reference
WL95	803,*supE,hsdR,tonA,trpR,metB*,P2 lysogen	(a)
ED8910	*supE,supF,recB,recC,hsdS,metB,lacY,galK,galT*	(a)
JC8679	AB1157 *recB,recC,sbcA*	(b)
DH5α	*endA,hsdR,supE,thi,recA,gyrA,relA* Δ(*argF-laczya*),*lacZ*ΔM15	(c)
NM554	MC1061(*hsdR,mcrA,mcrB*) *recA*	(d)
JM101	Δ(*lac-pro*),*supE,thi*[F'*traD,proAB,lacI*^{q}ZΔM15]	(e)
XL1-Blue	*endA,hsdR,supE,thi,recA,gyrA,relA* [F'*proAB,lacI*^{q}ZΔM15,*Tn*10(Tet^{r})]	(f)
FBXL5	JM101 Rif^{r} [F'*proAB,lacI*^{q}ZΔM15,*Tn*10(Tet^{r})]	(g)

References

(a) Loenen and Brammar (1980)

(b) Lloyd and Buckman (1985)

(c) BRL focus (1986)

(d) Raleigh *et al.* (1988)

(e) Messing *et al.* (1981)

(f) Stratagene Cloning Systems Inc.

(g) This work, section 2.1.3.1

human *Hin*dIII partial fragments cloned into the *Hin*dIII site of Lorist6, a derivative of LoristB (Cross and Little,1986). The ligation of large (5-15kb) human *Sau*3AI fragments into the *Bam*HI site of λL47 (Loenen and Brammar,1980) from the studies of Wong *et al.*(1987) was used in the experiments described in section 3.2.1.2.

Charomid vectors (Saito and Stark,1986) were supplied by the Japanese Cancer Research Resources Bank, Tokyo.

2.1.5 *Human and somatic cell hybrid DNA*

Human DNA from lymphoblastoid cell lines from members of the CEPH panel of families was kindly supplied by Profs. Howard Cann and Jean Dausset at the Centre d'Etude du Polymorphisme Humain (CEPH), Paris. Placental DNA samples from normal individuals were kindly supplied by Dr.Raymond Dalgleish, Department of Genetics.

DNA from human-rodent somatic cell hybrids were kindly supplied by Dr.Sue Povey, University College, London, and the hybrids used were largely those used by Wong *et al.* (1987)*;* The four hybrids used for the "mini-panel" used to map newly cloned minisatellites (section 3.3.5) were: 3W4C115 (Nabholz *et al.*,1969), FST9/10 (Kielty *et al.*,1982), HORL411B6 and HORP9.5 (Van Heynigen *et al.*,1975).

2.2 *STANDARD METHODS*

2.2.1 *General Methods; preparation of genomic DNA*

2.2.1.1 *General methods*

General methods for the preparation and manipulation of DNA were as described (Sambrook *et al.*,1989). Restriction endonucleases and other enzymes were used under the conditions recommended by the manufacturer.

2.2.1.2 *Preparation of genomic DNA*

DNA was prepared from blood as follows: frozen blood (20ml) was thawed and transferred to a 30ml centrifuge tube. After centrifugation (11,000 rpm, 10 minutes), the haemolysate was decanted and the nuclear and cellular pellet washed in 1 x SSC. The pellet was spun down as before and resuspended in 20ml 0.2M sodium acetate, pH 7.0, and SDS added to 1%. DNA was then prepared by phenol extraction followed by three ethanol precipitations.

Breast tumour samples (100-500mg) were disrupted by chilling in liquid nitrogen followed by pulverization in a Braun Mikro-Dismembrator II for 60 seconds. Ileal mucosa was disrupted using two fresh scalpel blades. Disrupted tissues were suspended in 1ml 150mM NaCl, 100mM sodium EDTA, pH 8.0. SDS was added to a final concentration of 1% and proteinase K to 0.5mg/ml, and the suspension incubated at 37°C overnight. After phenol extraction and ethanol precipitation, the crude preparation was purified further by treatment with ribonuclease A and pronase, followed by a second phenol extraction and three further ethanol precipitations. DNA concentration was assayed by measuring absorbance at 260nm, and undigested DNA (about 0.5μg) checked by agarose gel electrophoresis for degradation.

2.2.2 *Gel electrophoresis and Southern blotting*

Sigma type I agarose was used for gel electrophoresis, and gels were run in TAE (40mM Tris acetate, 20mM sodium acetate, 0.2mM EDTA, pH 8.3). Southern blot transfer was carried out using capillary transfer as described (Southern,1975). For DNA fingerprint analysis, nitrocellulose membranes (Schleicher and Schuell) were used; for all other applications, DNA was transferred onto nylon membranes (Hybond-N, Amersham). DNA was fixed onto membranes according to instructions supplied by the manufacturers.

2.2.3 *Hybridization conditions*

Hybridization probes were labelled by oligonucleotide priming using Klenow fragment (Feinberg and Vogelstein,1984), and labelled DNA recovered from unincorporated deoxynucleotides by ethanol precipitation using high molecular weight herring sperm DNA as a carrier. For DNA fingerprinting, single-stranded DNA templates were used for primer extension labelling as described (Jeffreys *et al.*,1985a).

Hybridizations with single locus minisatellite probes were carried out in phosphate/SDS buffer (Church and Gilbert,1984), with alkali-denatured human DNA as competitor to a final concentration of 10μg/ml (Wong *et al.*,1987) and washed in 0.1 x SSC, 0.01% SDS at 65°C. Hybridization conditions for DNA fingerprinting were as described (Jeffreys *et al.*,1985a). The hybridizations for MVR mapping by indirect end-labelling (section 4.4.3.3) and for genomic analysis of pJBT11 (section 5.2.2.4) were carried out in a milk-based buffer (1.5 x SSPE, pH7.7, 1% SDS, 0.5% Marvel, 6.12% PEG 6000) without any human DNA competitor.

Densitometric scanning of autoradiographs was carried out with an LKB Ultroscan XL laser densitometer, using film which had been pre-flashed and exposed without intensifying screens. Before re-use, nylon filters were stripped of probe using either boiling water or alkali treatment as recommended by the manufacturers.

2.2.4 *Purification of DNA fragments from agarose gels*

DNA fragments were isolated from horizontal agarose gels by electrophoresis at 3V/cm onto a piece of dialysis membrane placed vertically in a slot in the gel. In addition to this collecting membrane, a second membrane was placed behind the desired fragment, to prevent contamination with larger fragments. When all the DNA had collected on the membrane (as judged using a hand-held ultraviolet lamp), a block of gel was cut behind the collecting membrane and slid backwards, with the voltage still on, to expose the DNA on the membrane to the buffer. The membrane was then deftly removed into a microcentrifuge tube, and the small volume of buffer containing the DNA collected by centrifugation. DNA was recovered by ethanol precipitation. By this means it was possible to recover even relatively small amounts of DNA with good yield, and the recovered DNA could be used without further purification.

2.2.5 *Transformation of* E.coli

For routine cloning experiments, *E.coli* was transformed using cells prepared by the method of Hanahan (1983), or using competent cells purchased from BRL. In later experiments, *E.coli* was transformed by electroporation using a Bio-Rad Gene Pulser apparatus; cells were prepared as described (Dower et *al.*,1988), and 40μl samples subjected to an exponential pulse at 1.5kV (with a capacitance of 25μF and a parallel resistance of 940Ω) in a cuvette with a 2mm electrode gap. Under these conditions the time constant (τ) was about 20 milliseconds. Subcloning of minisatellites was performed using *E.coli* DH5α (for plasmid subclones), JM101 or FBXL5 (for M13 clones); in later experiments, in which pBluescript vectors were used, the host strain was XL1-Blue (Table 2.1).

2.2.6 *DNA sequencing*

The sequence of minisatellite-bearing clones was determined using the dideoxynucleotide chain termination method (Sanger *et al.*,1977) on single-stranded DNA templates. These were prepared by subcloning into M13 vectors, or by rescue of single-stranded DNA from pBluescript recombinants. pMS43, pMS228B and pMS607 were sequenced from random "shotgun" clones and directed clones proceeding from known restriction sites. pJBT10 and subclones of cMS608 were sequenced by the generation of nested subclones by exonuclease III/S1 nuclease treatment (Henikoff,1984). Sequencing reactions were carried out using Klenow fragment or T7 DNA polymerase (Tabor and Richardson,1987).

2.3 *SPECIFIC METHODS*

2.3.1 *Charomid cloning*

2.3.1.1 *Vector preparation*

Charomid 9-36 (Saito and Stark,1986) was obtained from the Japanese Cancer Research Resources Bank, Tokyo; the stab cultures sent were used directly to inoculate a 200ml culture in Nutrient broth supplemented with 50μg/ml ampicillin. After overnight growth, Charomid DNA was prepared by an alkaline lysis procedure (Birnboim and Doly,1979). Charomid DNA was digested with *Bam*HI, and the products analysed on a 0.4% agarose gel using intact λ phage DNA and λ DNA digested with *Xho*I as size markers. In addition to the expected linear molecules at about 36kb, shorter products were also seen, probably due to deletion of the "stuffer" repeats from the vector (Saito and Stark,1986). Full-sized (36kb) Charomid DNA was recovered from a preparative agarose gel (section 2.2.4) and ethanol precipitated before use.

2.3.1.2 *Ligation, packaging and infection*

Ligation was carried out at high DNA concentration, to promote concatemer formation, at a molar ratio of 2:1 (vector:insert). Specifically, 1.8μg of Charomid 9-36 linearised with *Bam*HI were ligated with 150ng 4-9kb human *Mbo*I fragments, in a total volume of 20μl. Ligation was carried out in the presence of 1mM ATP, and 0.1 Weiss units/μl of T4 DNA ligase (BRL), but using the ligation buffer recommended by New England Biolabs. The ligation reaction proceeded for 4 days at 15°C.

Ligated DNA was packaged into phage particles *in vitro* using Gigapack plus, according to the manufacturer's instructions.

4μl (corresponding to about 260ng total DNA) of the ligation was used in the packaging reaction. For infection, an overnight culture of host bacteria (*E.coli* NM554 for the ordered array library) was diluted 1:10 in Nutrient broth supplemented with magnesium sulphate to 10mM and maltose to 0.2%, and grown for a further 3 hours at 37°C. Dilutions of the packaged ligation were mixed with 200μl of these cells; after incubation at room temperature for 20 minutes, 700μl S.O.C. medium (Hanahan,1983) was added and the bacterial suspension shaken at 37°C for 60 minutes before spreading on Nutrient agar supplemented with ampicillin.

2.3.1.3 *Picking into ordered array*

After titration of the packaged DNA as described above, NM554 cells were infected and spread at a density of about 600 clones per 9cm Petri dish on nutrient agar plates supplemented with ampicillin to 50μg/ml. This density ensured that most colonies were well separated from their nearest neighbours while keeping the number of plates required reasonably low. Colonies were picked using pipette tips; colonies were picked by squashing into the end of the pipette tip, whence they were dispersed into the medium in the microtitre well by pipetting up and down. The wells of the microtitre plates were filled with 100μl of a medium consisting of 9 parts Nutrient broth to 1 part 10 x HMFM. 10 x HMFM contains 36mM K_2HPO_4, 13mM KH_2PO_4, 20mM trisodium citrate, 10mM $MgSO_4$ and 44% (w/v) glycerol. The bottom right hand well (H12) of each microtitre plate was left unfilled to aid unambiguous orientation of replica filters. Thus each plate contained 95 clones.

2.3.1.4 *Multi-locus probes*

Replica filters of the ordered array library were prepared by replica plating using a "hedgehog" as described (Coulson *et al.*,1986) onto nylon filters; DNA was fixed to the filters by microwave treatment (Buluwela *et al.*,1989).

The library was screened using the following multi-locus probes: 33.15 and 33.6 were as described (Jeffreys et *al.*,1985a,b) except that the double-stranded inserts from plasmid subclones, kindly supplied by Raymond Dalgleish, were used; the α-globin 3'HVR probe was the *Hin*fI insert from pSEA1 (Nicholls *et al.*,1985;Jarman *et al.*,1986), and was kindly supplied by Dr.Doug Higgs, Department of Molecular Medicine, University of Oxford; the M13 probe containing the tandemly repeated region of the protein III gene (Vassart *et al.*,1987) was a 1011bp *Alu*I-*Cla*I fragment isolated from M13mp8 RF DNA by Prof.Alec Jeffreys; the dispersed repeat sequences found in pMS1 (sections 4.3.2.2 and 4.3.2.4) were removed from the probe used by isolation of the largest fragment after digestion with AvaII, to produce the probe referred to here as "MS1.1A"; $(GGGCA)_n$ was a synthetic probe, made by tandem ligation of overlapping complementary 20-mer oligonucleotides, which corresponds in sequence to the repeat unit of a highly unstable mouse minisatellite (Kelly *et al.*,1989); a representative human Alu sequence probe was prepared by Mrs.Vicky Wilson by isolation of the 300bp S1 nuclease resistant fraction after partial renaturation of denatured human genomic DNA a described (Houck *et al.*,1979). All these probes were labelled by oligo-labelling (Feinberg and Vogelstein,1984).

2.3.2 *PCR methods*

2.3.2.1 *General methodology*

Contamination of samples for PCR work was minimised by adoption of a number of precautions. All chemicals, including water, were measured gravimetrically directly into new disposable plasticware. Pipette tips were used directly from sealed bags as supplied by the manufacturer. For work involving small numbers of target molecules (section 5.2.3) and in the preparation of stock solutions of buffers and oligonucleotide primers, work was carried out in a laminar flow hood.

2.3.2.2 *PCR primers and conditions*

Polymerase chain reactions were carried out in the buffer and nucleotide conditions described (Jeffreys *et al.*,1988b,1990a) using a Perkin Elmer Cetus DNA Thermal Cycler. The primers and cycle conditions used for amplification at the different loci studied are shown in Table 2.2. In all cases the amplification reactions were completed by following the denaturation/annealing/extension cycles by a single "chase", consisting of an annealing step followed by extension, without prior denaturation.

2.3.2.3 *Whole genome PCR; filter hybridization selection*

About 150ng of size-selected 1.8-2.5kb *Mbo*I fragments from JB tumour DNA were ligated with 500ng of *Sau*3AI linkers. The linkers were prepared by phosphorylation of primer SauLB (Table 2.2) with T4 polynucleotide kinase and ATP, followed by annealing with primer SauLA (Table 2.2). Ligated genomic DNA was separated from linker dimers by gel electrophoresis (section 2.2.4). Amplification of ligated molecules was carried out using primer SauLA alone, under the conditions described in Table 2.2.

Filter hybridization was carried out by preparation of about 20μg of the size fraction containing the novel fragment by PCR amplification; the DNA (in 10μl) was denatured by addition of

Table 2.2 Primers and cycle parameters for PCR amplification

Locus	Section	Denature	Anneal	Extend	Primers
607A(D22S163)	4.4.3.3	95°C 1′	67°C 1′	70°C 10′	607A,607B
608 (D12S40)	5.3.3	95°C 1′	50°C 1′	70°C 10′	608A,608B
228B(D17S134)	5.3.4	96°C 2′	60°C 1′	70°C 10′	228BA and 228BB or 228BC
MS31 (D7S21)	4.4.3.2	96°C 2′	65°C 1½′	70°C 10′	31AE,31B
MS32 (D1S8)	5.2.3	95°C 1′	67°C 1′	70°C 10′	32A,32B
"Whole genome"	5.2.2.4	95°C 1′	67°C 1′	70°C 10′	SauLA

Primer sequences

607A	5′CCTCTACAACCAGGTGCGACTGTG3′
607B	5′GCAGAGACAAGCCAGTAGGTATAC3′
608A	5′TTCAGATCTCCACTGAAAGGGTAC3′
608B	5′TAACTTATGTATATGCTTCCAGTC3′
228BA	5′AGCGCCACGAGCTCCTAGGGCCAG3′
228BB	5′CTTGGCTTTGACCCTGAGTCCCAA3′
228BC	5′TGGTGCAGACGCCCCGGAGCCCAC3′
31AE	5′CCTAGGATCCGAATTCTTTGCACGCTGGACGGTGGCG3′
31B	5′CCCACACGCCCATCCGGCCGGCAG3′
32A	5′TCACCGGTGAATTCCACAGACACT3′
32B	5′AAGCTCTCCATTTCCAGTTTCTGG3′
SauLA	5′GCGGTACCCGGGAAGCTTGG3′
SauLB	5′GATCCCAAGCTTCCCGGGTACCGC3′

1μl of 0.25M KOH, and left at room temperature for 5 minutes. The alkali was neutralized by the addition of 2μl 1M Tris-HCl, pH 7.5, followed by 1μl 0.25M HCl. This was added to 200μl phosphate/SDS buffer (Church and Gilbert,1984) containing 400μg/ml alkali-denatured human DNA as competitor, and a nylon filter bearing pMS1 DNA. This last was prepared as follows. 5μg pMS1 DNA was linearised with *Eco*RI, followed by phenol extraction and ethanol precipitation. The linear DNA was redissolved in 20μl water, and denatured by the addition of 2μl 1M KOH, followed by incubation at room temperature for 5 minutes. The solution was neutralised with 4μl Tris-HCl pH7.5 and 8μl 0.25M HCl. 2μl portions were spotted onto a 4 x 10mm rectangle of Hybond-N, allowing the spot to dry between applications. The DNA was fixed to the membrane by u.v. irradiation and the filter cut into small (about 2 x 2mm) pieces and immersed in the hybridization buffer.

The hybridization solution was covered with paraffin oil and incubated in a microcentrifuge tube overnight at 65°C. The filter fragments were washed in 100ml 0.1 x SSC, 0.01% SDS at 65°C. Hybridizing fragments were removed from the filter by washing in 100μl 10mM KOH, 0.01% SDS at room temperature for 5 minutes, followed by 100μl 0.5M Tris-HCl, pH 7.5, 0.01% SDS. These washings were pooled and the DNA was ethanol precipitated using 6μg high molecular weight herring sperm DNA as a carrier. 1/20th of this recovered DNA was used in each of the amplification reactions shown in Figure 5.9.

CHAPTER 3

ISOLATION OF HUMAN MINISATELLITES

πολλ'οι∂'αλωπηξ,'αλλ'εχινοs 'εν μεγα

Archilocus

The fox has many tricks, the hedgehog one big one.

Summary

The genetic mapping of the human genome requires the generation of large numbers of informative genetic markers. The loci detected in DNA fingerprints represent a large pool of loci, which if isolated could contribute enormously to the generation of a detailed genetic map. Previous approaches to the isolation of these loci are discussed. Cloning of minisatellites by screening λ phage libraries in *rec*$^+$ hosts with DNA fingerprinting probes, while initially very successful, appeared to lead to the isolation of only a small subset of the loci. Attempts to circumvent this selection by low density plating directly onto a *recBC* host were unsuccessful. Direct cloning in plasmid, although successful in leading to the cloning of a new hypervariable locus, was shown to be too inefficient to be used to isolate large numbers of minisatellites. An ordered array Charomid library was found to be a systematic and efficient method for cloning minisatellite loci. This method led to the isolation of a larger subset of human loci; 23 new minisatellite loci were cloned, characterized and assigned to chromosomal locations.

3.1: *HUMAN MINISATELLITES ALREADY ISOLATED*

3.1.1 *Fortuitous isolation*

3.1.1.1 *An anonymous cloned segment*

Wyman and White (1980) described the isolation of a random clone from a human genomic library which detected a highly polymorphic locus. This locus was localised to the telomeric region of the long arm of chromosome 14, and was highly variable with nearly everyone in the population heterozygous. Sequence analysis subsequently demonstrated that the locus was composed of short tandem repeats (Balazs *et al.*,1986).

3.1.1.2 *Minisatellites flanking gene sequences*

Many of the variable tandem repeat loci to be first described were discovered fortuitously in the analysis of DNA flanking gene sequences. Thus early among those discovered were those in the 3′ flanking DNA of the *c-H-ras* gene (Capon *et al.*1983) and the 5′ flanking DNA at the insulin gene (Bell *et al.*, 1982), both on the short arm of chromosome 11. Other examples of polymorphic tandemly repeated regions unexpectedly appearing in the non-coding DNA near genes include those near the type II collagen gene (Stoker *et al.*,1985), the apolipoprotein B gene (Knott *et al.*,1986) and no fewer than seven in the α-globin cluster on chromosome 16 (Higgs *et al.*,1981;Goodbourn *et al.*,1983;Jarman and Wells,1989).

3.1.1.3 *Large-scale screening of random cloned segments*

A labour-intensive yet successful method for isolating polymorphic regions was adopted by Braman *et al.*(1985). Their method was simply to screen more than 1500 λ clones either by isolating single-copy fragments or by using whole phage clones with a prehybridization step to suppress cross-hybridization from repeated sequences (Schumm *et al.*,1985). DNA from five unrelated individuals was tested using six different restriction enzymes. More than 500 single-copy RFLPs were thereby identified (Schumm *et al.*,1985), but 29 loci were also identified which were highly polymorphic, having PIC values between 0.7 and 0.9. Of these, 14 revealed the same pattern of polymorphism when tested with two or more restriction enzymes, suggesting that the polymorphism was due to a length variable region. The other clones, however, detected highly polymorphic loci only with one enzyme, the authors suggesting that their multiallelic variation was due to clusters of polymorphic sites for that enzyme (Braman *et al.*,1985).

This work provided the backbone of the first report including genetic maps of most of the human genome (Donis-Keller *et al.*,1987). However, having isolated enough loci to underpin such an initial effort, the method appears to be too inefficient to be applied to the "fleshing out" of the genetic map at higher resolution.

3.1.2 *Directed cloning*

3.1.2.1 *DNA fingerprints and "core" sequences*

The isolation of polymorphic tandemly repeated loci, which had hitherto been decidedly haphazard, became more systematic after the discovery of probes which hybridized to multiple polymorphic loci in human DNA (Jeffreys *et al.*,1985a,1985b). A tandem repeated sequence from within the first intron of the human myoglobin gene was shown to detect a set of polymorphic loci. Among the loci isolated by cloning and hybridization screening, two (probes 33.6 and 33.15) each themselves detected a large number of polymorphic loci in human DNA; the two probes recognise substantially independent sets of polymorphic loci (Jeffreys *et al.*,1986), and remain the most widely used of the DNA fingerprinting probes.

The discovery of these probes showed that there existed a large pool of hypervariable loci in the human genome from which individual loci could, in principle, be cloned by hybridization screening. The first examples cloned using the myoglobin repeat (Jeffreys *et al.*,1985a) seemed to share a common G/C-rich "core" sequence which, it was suggested, was important in the origin of variability at these loci. The question of the importance of "core"-like sequences is taken up again in the general discussion in chapter 6. However, the detection of these loci on DNA fingerprints suggested that the loci could be isolated by cloning and hybridization screening using probes 33.6 and 33.15, and any other probes giving multi-locus DNA fingerprints. Several such DNA fingerprinting probes have now been described, and include the α-globin 3'HVR (Fowler *et al.*,1988) and a tandemly repeated sequence from the protein III gene of M13mp18 (Vassart *et al.*,1987). The remainder of this introductory section (3.1.2.2 to 3.1.3) is given over to a discussion of the attempts to realise this promise of the isolation of large numbers of highly polymorphic markers.

3.1.2.2 *Cloning a locus defined by a DNA fingerprint band*

The first isolation of a locus defined by a DNA fingerprint band (Wong *et al.*,1986) was prompted by the possible segregation in a large pedigree of a band from a DNA fingerprint produced by probe 33.15 with a form of hereditary persistence of foetal haemoglobin (HPFH) not linked to the β-globin cluster. The method involved sequential size-selection of *Sau*3AI fragments from a given individual until the target band was separated from all other DNA fingerprint bands. This fraction was used to make a genomic library in λL47 (Loenen and Brammar,1980) and screened using the DNA fingerprint probe 33.15. The locus thus isolated (pλg3,Wong *et al.*,1986) was highly informative (population heterozygosity about 97%), but was subsequently shown not to be linked to the HPFH gene.

3.1.2.3 A *more general approach*

While the isolation of pλg3 (*v.s.*, section 3.1.2.2) used fractions highly enriched for a particular fragment, it remained in principle possible to clone a wider selection of loci by using a wider size fraction of fragments from DNA pooled from unrelated individuals. Allele sizes at minisatellite loci vary considerably, and so the use of DNA pooled from unrelated individuals maximises the likelihood of a restriction fragment from a particular locus appearing in the size fraction chosen. Wong *et al.*(1987) cloned 5-15kb *Sau*3AI fragments into the phage λ vector L47 (Loenen and Brammar,1980). The most variable loci detectable on DNA fingerprints have *Sau*3AI alleles larger than 4kb; such large *Sau*3AI fragments are very uncommon in human DNA, the vast majority being smaller than 2kb. Thus while the choice of a 5-15kb size fraction does not result in the extreme enrichment obtainable when cloning a single band, it nevertheless allows a considerable enrichment in hypervariable minisatellites over unfractionated DNA.

Five loci were cloned in the work reported by Wong *et al.*(1987). This work was extended on a larger scale by the application of the same methodology; however, the yield of new loci was not sustained, and suggested that only a limited subset of the hypervariable minisatellites present in the genome were amenable to isolation by this method (*v.i.*, section 3.2.1.1).

3.1.2.4 *Cosmid cloning*

Those loci detectable on a DNA fingerprint which show the highest variability tend to have alleles in the range 3-20kb, and below about 3kb there are many loci detectable, but many of them are monomorphic or minimally variable (Jeffreys *et al.*,1986; Uitterlinden *et al.*1989). Thus the studies of Wong *et al.*(1986;1987), which used large size selected DNA fragments, made the isolation of single-locus probes more efficient by excluding the large number of loci which cross-hybridize with the DNA fingerprinting probes 33.6 and 33.15 but yet have small, relatively invariant alleles. However, the isolation of minisatellite loci by Nakamura *et al.*(1987a;1988b) used libraries of human genomic DNA in cosmids without prior size-selection, thereby including the large majority of smaller, less variable loci. After hybridization screening with oligonucleotides based on G-rich repetitive sequences (Nakamura *et al.*,1987a) or a G-rich consensus derived from sequence analysis of cloned hypervariable loci (Nakamura *et al.*,1988b), positively hybridizing clones were tested for polymorphism by using them as hybridization probes against DNA from six unrelated people each cut with six different restriction enzymes. This relatively sensitive test not only allowed the identification of minisatellite loci but also gave rise to a "bonus" yield of loci which showed RFLPs with particular enzymes. In this way, despite some practical drawbacks (section 3.3.7.2), an impressive collection of moderately to highly polymorphic loci was isolated.

3.1.2.5 *Isolation of the α-globin 3'HVR*

Although recognized as a region of multiallelic variation by restriction mapping (Higgs *et al.*,1981), the tandem repeat region responsible for polymorphism 3' of the α-globin gene cluster proved extremely refractory to cloning in phage or cosmid vectors, even using recombination-deficient host strains. For this reason it was isolated by direct cloning into plasmid in a *RecA E.coli* host following sequential enrichment by rounds of restriction digestion and size-selection (Nicholls *et al.*,1985). While ultimately successful in the isolation of the locus, it is illustrative that even where a flanking restriction map and flanking DNA probes were to hand, extreme efforts were required to clone the locus.

3.1.3 *Overview of cloning methods and problems*

As outlined above, considerable progress has been made in the isolation of hypervariable regions of the human genome. The large scale isolation of informative genetic markers, particularly in the work of Nakamura et *al.*(1987a;1988b) and Donis-Keller *et al.*,(1987), has provided the basis of the primary genetic maps of the human autosomes. However, genetic mapping and analysis at higher resolution will require the continued generation of large numbers of highly informative genetic markers. Furthermore, the isolation of highly informative genetic markers not only provides a quantum leap in the efficiency of genetic linkage mapping, but also makes feasible many otherwise impracticable genetic analyses (section 1.2.3). This chapter of the thesis is devoted to the investigation of methods which might make the large-scale isolation of highly informative genetic markers an efficient process.

While the studies of Nakamura *et al.*(1987a,1988b) and Braman *et al.*(1985) used *cloning* techniques which were simple and efficient, a low proportion of the clones tested were hypervariable, and so a large investment of effort was required in the production of each new marker. The work of Wong *et al.* (1987, sections 3.1.2.2, 3.1.2.3) suggested that the screening of libraries of size selected DNA with DNA fingerprinting probes may have provided a simple and efficient method for cloning large numbers of hypervariable loci; however, the initial success was not continued (*v.i.*, section 3.2.1.1), and the sobering instance of the α-globin 3'HVR (section 3.1.2.5) implies that at least some such regions may, for reasons at present mysterious, simply not be amenable to cloning by conventional methods.

3.2: *CLONING HUMAN MINISATELLITES IN λ AND PLASMID VECTORS*

3.2.1 *Further studies with cloning in phage λ*

3.2.1.1 *Diminishing returns*

The cloning of human minisatellites by screening λ libraries of size-selected DNA fragments, as described by Wong *et al.*(1987), was extended on a larger scale in unpublished work of Dr.Nicola Royle, Richard Clarkson and Prof. Alec Jeffreys. As a result of this work, a further five variable minisatellites (λMS205, 207, 214, 228 and 301) were isolated. Like those isolated by Wong *et al.*(1987), they detected loci with high levels of population heterozygosity (mean value about 82%). However, the main feature of this larger-scale screening of λ libraries was the re-isolation of loci already characterized (most frequently that detected by λMS8 (Wong *et al.*,1987)), rather than the isolation of new loci. Given that there were estimated to be more than 50 loci detected by probes 33.6 and 33.15 on a DNA fingerprint (Jeffreys *et al.*,1986), it was clear that selection against the vast majority of loci was operating in λ libraries.

3.2.1.2 *Direct plating at low density*

The studies of Wong *et al.*(1987), and their continuation outlined above, involved the screening at high density of recombinant phage plated onto a *rec*$^+$ host and only subsequently onto a *recBC* host. A striking property of the minisatellite recombinants isolated in these studies was the very poor yield of DNA from recombinant phage (Wong *et al*,1986), suggesting a growth defect in these phage. If, then, selection against most minisatellite clones were simply a problem of poor phage viability leading to impaired competition for host bacteria, it should be alleviated by initial plating of recombinant phage at low density, such that many clones were clearly separated from their neighbours.

An additional feature of this experiment was to bypass the usual step of plating on the *rec*$^+$ *E.coli* strain L95 (Loenen and Brammar,1980). This strain is a P2 lysogen and was used to impose *Spi* selection against non-recombinants (Loenen and Brammar,1980); this step is not, however, necessary, since the size constraints of the packaging system should select adequately against non-recombinants.

The ligated DNA from the studies of Wong *et al.*(1987) was packaged *in vitro* and recombinant phage used to infect the *recBC E.coli* strain ED8910 (Loenen and Brammar,1980). About 4000 recombinants were plated onto five 9cm plates, such that many phage plaques were well separated from their nearest neighbours. These were screened by replication onto Nylon filters and hybridization using probe 33.15. 11 positive plaques were identified; of these, six underwent two further rounds of hybridization screening. DNA was prepared from purified phage (Blattner *et al.*,1977) and the *Sau*3AI inserts isolated. The *Sau*3AI inserts were labelled by random priming and used as hybridization probes against Southern blots of *Sau*3AI-digested DNA from three unrelated people. This showed that all six clones tested were isolates from loci already cloned, including two clones from the locus already identified by λMS8 (Wong *et al.*,1987)

3.2.2 *Direct cloning in plasmid vectors*

3.2.2.1 *Rationale and overview*

Many of the constraints imposed by cloning in λ vectors could be circumvented by the use of plasmid vectors; unlike λ vectors, their continued replication is not dependent on a size-limited packaging mechanism, and they can be propagated in *recA E.coli* hosts. Indeed, it was only by recourse to direct cloning into plasmid that the α-globin 3'HVR was isolated (Nicholls *et al.*,1985). However, the cloning of large restriction fragments into plasmid would be hampered by two inefficient steps in the process: first, the requirement to ligate the large insert plus vector into a circular product; second, the transformation into *E.coli*, a process which becomes less efficient as the size of the transforming DNA increases.

Probably for these reasons, the attempts at direct cloning of large restriction fragments into plasmid vectors were profoundly unsuccessful (*v.i.*), and its replacement by cloning of sonicated fragments from the same size fraction improved the prospects for the cloning of new loci. However, even though this latter approach led to the isolation of a new hypervariable locus, it was clear that direct cloning even of sonicated DNA fragments into plasmid vectors was too inefficient to be an acceptable method for the cloning of large numbers of minisatellites.

3.2.2.2 *Direct cloning of large restriction fragments*

Size-selected 4-6kb *Alu*I fragments were ligated into the *Sma*I site of pUC13. Test transformations of *E.coli* strains with the ligated DNA showed that the *recBCsbcA* strain JC8679 (Gillen *et al.*,1981) would yield more recombinants than DH5α (*recA*). However, even with large-scale transformation at an efficiency of 3 x 10^8 transformants/μg control plasmid (pUC13), only about 500 transformants were obtained. None of these gave a positive signal on hybridization screening with probe 33.15.

3.2.2.3 *Cloning sonicated fragments*

Although prior size-selection of large restriction fragments enriches genomic DNA for minisatellites, and thereby greatly reduces the number of clones needed to represent a genome equivalent, it is precisely the large size of the inserts which causes the extreme inefficiency of cloning into plasmid vectors (*v.s.*). The enrichment, however, can be uncoupled from the inefficiency by first size-selecting the large restriction fragments and then sonicating that fraction into DNA fragments small enough to be cloned with reasonable efficiency.

A fraction consisting of large (3-4kb) human *Alu*I fragments was sonicated to 200-600bp fragments, end-repaired and ligated into the *Sma*I site of pUC13. Transformation yielded many more recombinants than with the direct cloning of large *Alu*I fragments. The recombinants were screened by hybridization with probe 33.15, and after three rounds of screening, DNA was prepared and the insert used to probe Southern blots of DNA from unrelated people. Of six positively hybridizing clones tested in this way, four gave a monomorphic "ladder" of hybridizing fragments suggestive of satellite DNA (see section 3.3.2.3), one was another isolate from the locus detected by λMS31 (Wong *et al.*,1987), and one, termed pMS502, detected a newly defined locus.

pMS502 contained a 450bp human DNA insert which detected polymorphic hybridizing fragments in human DNA digested with *Alu*I, *Hin*fI or *Sau*3AI. A survey of DNA from 20 unrelated people digested with *Hin*fI allowed the heterozygosity level to be estimated at about 92%. pMS502 was used to probe a Southern blot of DNAs from somatic cell hybrids (Wong *et al.*,1987) and a provisional assignment of the locus was made to human chromosome 8 (data not shown).

Despite this success, it was clear that this was not a method efficient enough to be applied on a larger scale to the isolation of greater numbers of minisatellite loci. Conspiring to this end were two consequences of the small insert size: the low signal to noise ratio on hybridization screening, which resulted in the investigation of many falsely positive signals; and the low insert to vector size ratio, resulting in poor yields of insert DNA and low hybridization intensities when used to probe Southern blots of genomic DNA. Furthermore, this method would only usually result in the isolation of repeat units without flanking sequence, and so would not allow rapid conversion to PCR technology.

3.3: *CLONING IN ORDERED ARRAY CHAROMID LIBRARIES*

3.3.1 *Rationale*

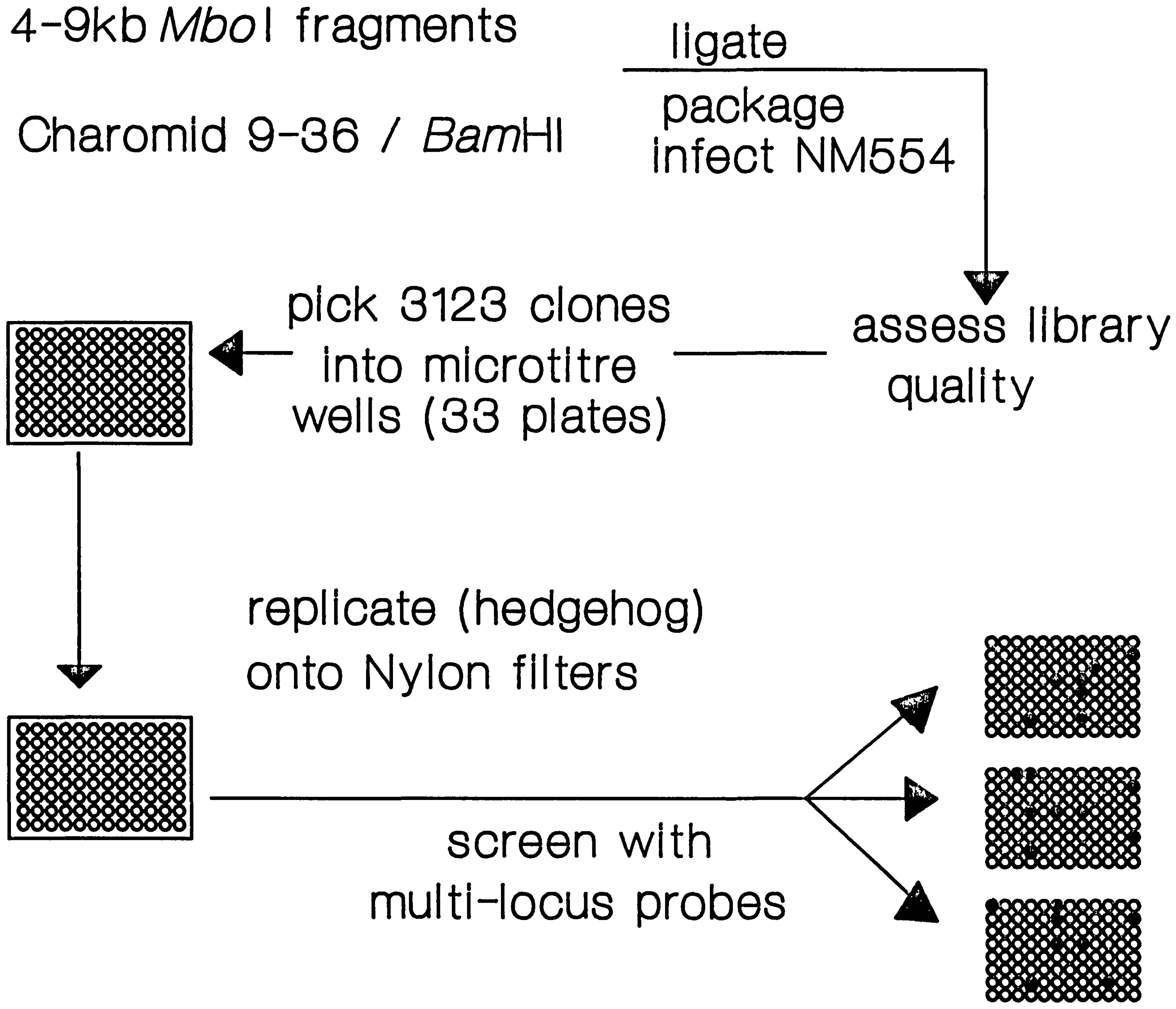
The cloning of large numbers of human minisatellite loci requires an efficient method which allows the isolation of many of the loci in the genome. While cloning methods based on λ vectors have high efficiency *in vitro* packaging, they appear to be strongly biased against the isolation of all but a small subset of loci (see above, section 3.2.1). The probable basis for at least some of this systematic selection against most tandemly repeated loci is shortening of cloned inserts by deletion of tandem repeats (Kelly *et al.*,1989), such that the shortened inserts make the recombinant phage DNA too short to repackage into viable particles.

One would expect *E.coli* hosts defective in *rec* functions, in particular *recA*, to propagate full-length tandemly-repeated DNA at higher frequency; while some bacteriophage vectors exist which can grow in *recA* hosts (Loenen and Blattner,1983), such systems still suffer from the requirement of recombinant phage to remain of packageable size at each generation of growth. Plasmid cloning (see section 3.2.2) appears to be too inefficient to isolate large numbers of loci. Cosmid cloning combines the efficiency of *in vitro* packaging with propagation in *recA* hosts, and has been the basis of some successful studies (Nakamura *et al.*,1987a,1988b, and see section 3.1.2.4). However, cosmid vectors require large inserts (typically 30-45kb) and hence cannot be used to construct libraries from those DNA fragments (3-20kb in size) richest in hypervariable human minisatellites.

Charomids (Saito and Stark,1986) are cosmid-based cloning

Figure 3.1

Schematic representation of the construction and screening of the ordered array Charomid library. The details are given in sections 3.3.2.1 to 3.3.3.3.



4-9kb *Mbo*I fragments

Charomid 9-36 / *Bam*HI

ligate

package
infect NM554

assess library
quality

pick 3123 clones
into microtitre
wells (33 plates)

replicate (hedgehog)
onto Nylon filters

screen with
multi-locus probes

vectors which allow the cloning of small inserts. The vectors are expanded by the inclusion of a tandem-repeated "spacer", so that constructs including small inserts have a large enough distance between adjacent *cos* sites to package efficiently. For example, Charomid 9-36 is a 36kb vector which can therefore accommodate inserts of 2-16kb into constructs of efficiently packageable size. Since, after the initial *in vitro* packaging, the recombinants can be propagated as plasmids in a *recA* host, such a system would allow the efficient cloning of size-selected large *Mbo*I fragments without the need for clones to remain of packageable size at each generation of growth.

3.3.2 *An ordered array Charomid library*

3.3.2.1 *Library construction*

Library construction and screening are summarised in Figure 3.1. Human DNA, pooled from 14 unrelated individuals, was digested with *Mbo*I and size fractions prepared by two rounds of gel electrophoresis and collection onto dialysis membrane (see section 2.2.4). 150ng of *Mbo*I DNA fragments from a 4-9kb size fraction were ligated with 1.8μg *Bam*HI-digested Charomid 9-36. This results in a vector:insert molar ratio of about 2:1. Of the resulting ligation, one fifth was packaged *in vitro* (Gigapack, Stratagene).

3.3.2.2 *Effect of host strain: ordering clones*

To investigate the effect of the *mcr* restriction system on the survival of recombinants, equal aliquots of the packaged library were used to infect *E.coli* strains DH5α (mcrA$^+$mcrB$^+$) and NM554 (mcrA$^-$mcrB$^-$). The titre of ampicillin-resistant colonies was about four times greater with NM554 ($\approx$2.1 x 10^6 cfu/μg) than with DH5α ($\approx$5.2 x 10^5 cfu/μg), suggesting that many recombinants in the library were destroyed by the *mcr* activity of DH5α; screening by hybridization with probe 33.15 under low stringency conditions showed that similar proportions of positively hybridizing clones were obtained with the two host strains, suggesting that the difference was not simply due to discrimination by DH5α against a frequent, non-hybridizing set of sequences. When used to infect NM554 cells, the titre suggested that the total library size was about 3 x 10^5 cfu.

By comparison of the DNA recovered from size-selection with the starting material, and assuming nearly full recovery of the required fraction, the 4-9kb *Mbo*I fraction appeared to account for about 0.19% of the human genome. If about 80% of clones in the library are recombinants (v.i., section 3.3.2.3) with mean initial insert size of 7kb, then about 1000 clones should contain one haploid genome equivalent for this subset of human DNA sequences.

The library was used to infect NM554 cells, and plated on ampicillin plates at low density (about 200 colonies per 9cm plate). A total of 3123 well-separated ampicillin-resistant colonies (about 3 haploid equivalents) were picked individually into the wells of microtitre plates. Each colony was dispersed into 100μl of Luria broth supplemented with HMFM (section 2.3.1.3), and the library microtitre plates were stored frozen at -20°C.

3.3.2.3 *Checks of library quality*

24 clones were picked at random from the ordered array, and DNA was prepared from overnight cultures. Digestion with *Sau*3AI showed that 18 had inserts (75%), commensurate with the rate of 80% found by Saito and Stark (1986). The average insert size was 4.9kb (range 2-6kb). While this includes inserts rather

smaller than the size-selected fraction, and none from the top end of the size range (7-9kb), it is likely that this reflects "collapse" of tandemly-repeated inserts (Kelly *et al.*,1989) during propagation rather than inefficient size-selection.

Inserts from 8 randomly selected clones were labelled by random priming and used to probe *Mbo*I-digested human DNA under high stringency conditions. Of the 8 inserts used, one hybridized to a single, apparently monomorphic fragment, while the remaining 7 hybridized to many-banded "ladders" of DNA fragments (for example, see Figure 3.3d); these "ladders" could be classified into two main groups. One would anticipate that arrays of satellite sequence would give such a pattern on Southern hybridization, with each rung of the "ladder" differing from the one below by a single satellite repeat unit. Indeed, one might predict that since any sequence appearing in the library needs to extend for at least 4kb without the appearance of a site for *Mbo*I, any such sequence would be expected to contain a substantial block of tandemly repeated sequences. Satellite DNA represents the most abundant class of tandemly repetitive DNA in the genome, and it is therefore not surprising to discover that it constitutes the bulk of the human DNA cloned in this library.

3.3.3 *Screening the ordered library array*

3.3.3.1 *Replication by hedgehog*

The ordered array was replicated by prodding with a metallic "hedgehog" at single or fourfold density as described (Coulson *et al.*,1986;Brownstein *et al.*,1989) onto Nylon filters resting on ampicillin agar. After overnight growth, DNA from each clone was fixed onto the membrane by microwave treatment (Buluwela *et* al.,1989).

3.3.3.2 *Probes used to screen the library*

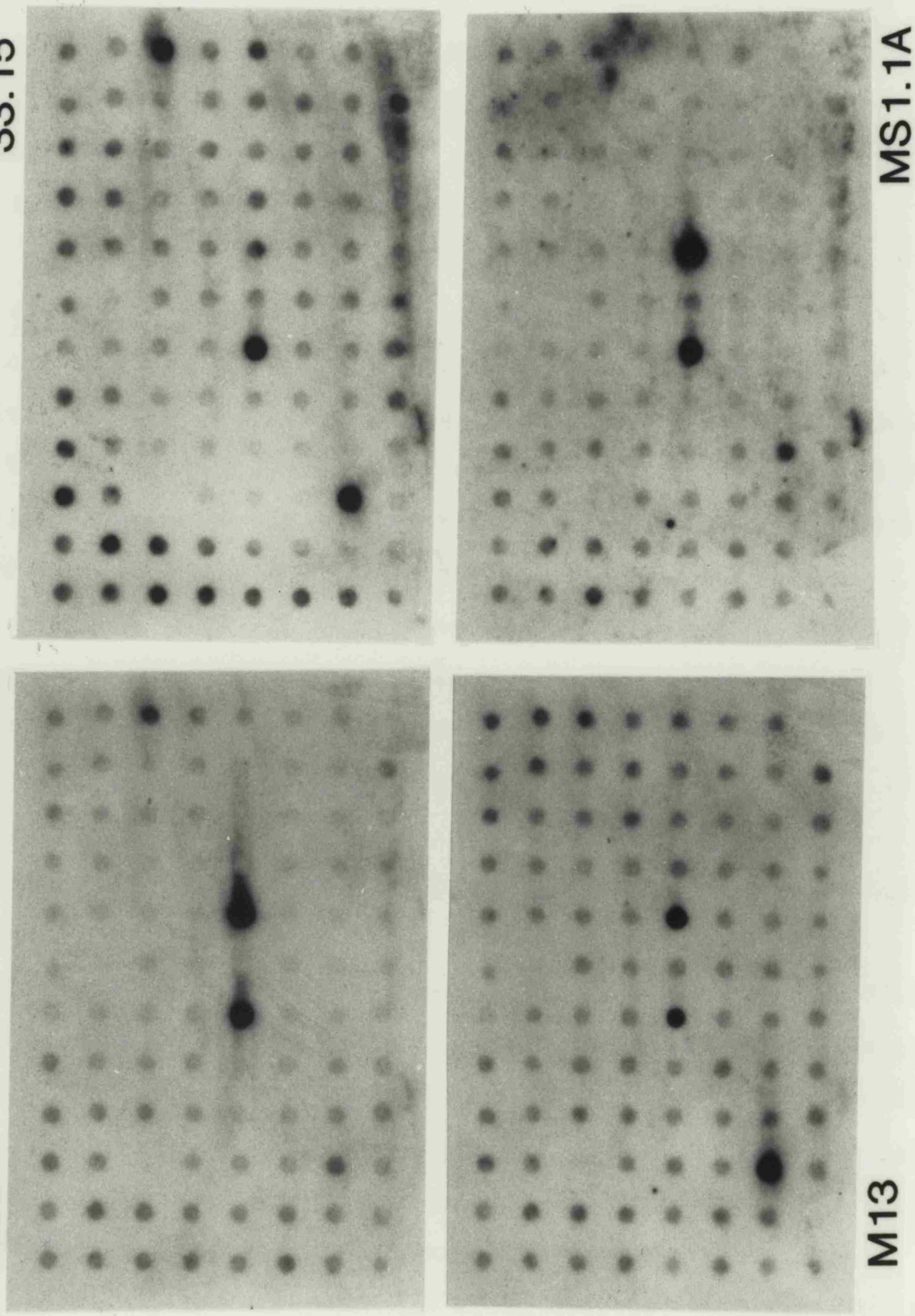
In order to identify clones potentially containing hypervariable minisatellites, the library was screened by hybridization with probes known to detect multiple variable loci in human DNA. The details of the probes used are given in section 2.3.1.4. They were: probes 33.6 and 33.15 (Jeffreys *et al.*,1985a,1985b); the α-globin 3'HVR (Jarman *et al.*,1986); the tandemly repeated region of the protein III gene of M13 phage (Vassart *et al.*,1987); a synthetic probe consisting of GGGCA repeats, which corresponds to the repeat unit of a highly variable mouse minisatellite (Kelly *et al.*,1989); and the highly variable human minisatellite pMS1, which consists of 9bp repeat units (Wong *et al.*1987). Since pMS1 contains two dispersed repeat elements (Armour *et al.*,1989b; section 4.3.2), the probe used to screen the library consisted of a shorter *Ava*II fragment which contained the tandemly repeated minisatellite, but neither dispersed repeat.

Screening of library replica filters with the M13 fingerprinting probe was carried out under the conditions described by Vassart *et al.*(1987). Screening with all other probes was carried out under the different conditions, namely overnight at 65°C in 1 x Denhardt's solution, 3 x SSC, 0.1% SDS and 6% PEG 6000; oligo-labelled probes were added at 0.5ng/ml, together with 1μg/ml Charomid 9-36 DNA and 2μg/ml high molecular weight *E.coli* DNA as competitors. Filters were washed at 65°C in 1 x SSC, 0.1% SDS.

In addition to the probes outlined above, which are all known to detect multiple hypervariable loci in human DNA, a consensus human Alu element probe was used to screen the

Figure 3.2

Example of screening a replica plated ordered array charomid library with multi-locus minisatellite probes (section 3.3.3.3). The results from a representative filter are shown after hybridization with the multi-locus probes 33.6 and 33.15, with the DNA fingerprinting region from phage M13 (Vassart *et al.*,1987), and using the single locus minisatellite probe MS1 at low stringency (see section 3.3.3.2). Note that while there is some overlap between the positive clones with each probe, the probes nevertheless each detect distinct subsets of clones. The full library was replicated onto 33 such filters.


33.15
MS1.1A
33.6
M13

library. There appears to be an association between dispersed repeat elements and hypervariable minisatellites in human DNA (Armour *et al.*,1989b; section 4.3.2); if so, then screening with an Alu probe might allow the detection of hypervariable minisatellite loci independently of the multi-locus probes mentioned above.

3.3.3.3 *Results of hybridization screening*

An example of one of the library replica filters after sequential screening with four of the multi-locus probes is shown in Figure 3.2; results for the whole library screened with the six multi-locus probes (*v.s.*) are summarised in Table 3.1. From a total of 3123 ordered clones, of which about 2500 were estimated to contain human inserts, 185 (about 7.4%) were positive with at least one of the multi-locus probes. The disposition of the library in ordered array not only allows the isolation of positively-hybridizing clones without further rounds of screening, but also allows comparison of the same clone screened with different multi-locus probes (Figure 3.2) and hence an indication of the degree of overlap or independence of the sets of clones detected by each multi-locus probe.

This approach showed, for example, that probes 33.6 and 33.15 appeared to detect substantially independent sets of clones (of 141 clones hybridizing positively with 33.6 and/or 33.15, only 26 clones were found to hybridize positively with both probes), a finding concordant with the genetic independence of the loci detected in DNA fingerprints using these probes (Jeffreys *et al.*,1986). By contrast, the M13 probe detected very few clones (Table 3.1), all of which were also detected by probe 33.15. The probe ("1.1A") derived from pMS1, which itself was

Table 3.1

Summary of the screening of an ordered array charomid library with six multi-locus probes. In the first two columns, the total number of positively hybridizing clones detected in a library of about 2500 recombinants is followed, in brackets, by the number hybridizing with that probe alone. The clones screened for polymorphism by Southern blot hybridization to human DNA are divided into "monomorphic", "satellite" and "polymorphic". The "satellite" category was inferred from the ladder-like multi-band patterns seen on hybridization (see Figure 3.3d). Polymorphic clones gave either multi-band patterns ("midisatellites", see section 3.3.6.4), simple patterns seen with clones already studied ("repeat isolates") or simple patterns from loci not previously seen. Only this last category was studied further. The multi-locus probes used to screen the library are shown in approximate order of productivity, and in the last column the number of new loci contributed by each multi-locus probe to the cumulative total of new VNTR loci is shown. "RFLP" refers to those loci at which variation is seen only with *Mbo*I, and so is probably due to a polymorphic site for *Mbo*I (section 3.3.4.2).

Table 3.1

					polymorphic					
Probe	positively hybridising (uniquely)	checked for polymorphism	mono-morphic	satellite	midi-satellite	single locus	repeat isolates	distinct, single loci	RFLP	new VNTR loci
33.15	84(38)	57(27)	28	6	1	22	6	16	2	14
MS1.1	48(32)	31(18)	4	13	1	13	2	11	0	6
33.6	58(24)	32(11)	12	6	3	11	5	6	1	2
α 3'HVR	31(14)	17(5)	6	2	2	7	2	5	0	1
$(GGGCA)_n$	23(17)	7(2)	4	0	2	1	0	1	0	0
M13	7(0)	4(0)	4	0	0	0	0	0	0	0
Total	185	100			5	33	8	25	2	23

originally detected by 33.15, detected a large set of clones of which two-thirds (32/48) were not detected by any other probe. Moreover, the clones MS440, 604, 608, 610, 619 and 623 which detect new variable loci (Table 3.2) all hybridized positively with pMS1 but not with 33.15. This suggests that pMS1 was detecting a set of clones more distantly related in sequence to 33.15. One might envisage sequential application of this technique, as in the "probe walking" method of Washio *et al.*(1989), to use clones from newly-isolated loci as hybridization probes to screen the library at low stringency, and thus to detect new sets of loci which had not been detected by the initial multi-locus probe.

3.3.4 *Clones detecting polymorphic loci*

3.3.4.1 *Checking clones for polymorphism*

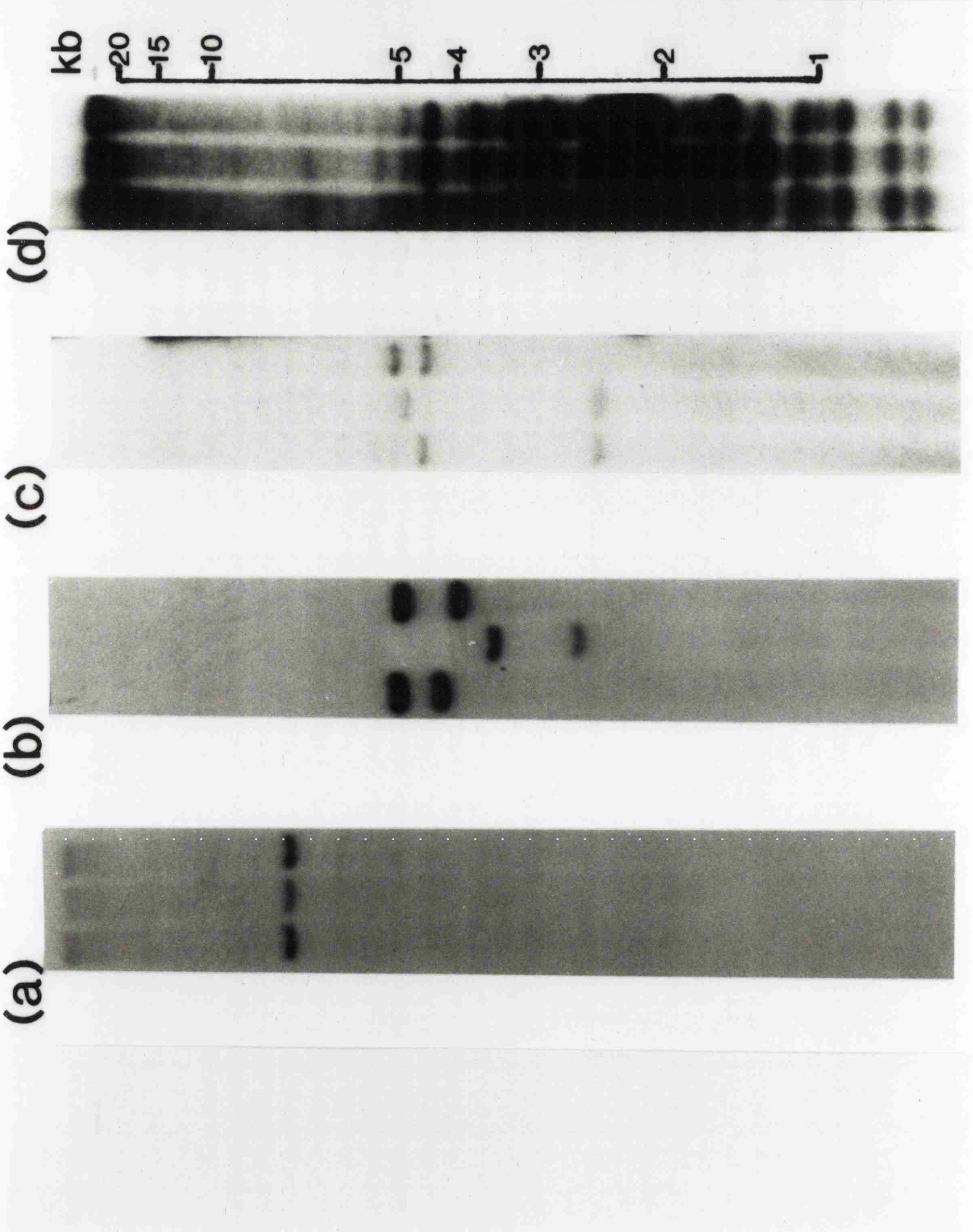
The *Sau*3AI inserts were isolated from 100 clones hybridizing positively with at least one of the multi-locus probes. When the clone had two inserts, these were isolated and tested separately. Labelled inserts were used as hybridization probes at high stringency against a standard Southern blot of *Mbo*I-digested DNA from three unrelated people (Figure 3.3). This provides a screen for reasonably polymorphic loci (a locus with a population heterozygosity of 60% or more would have a greater than 94% chance of detection), but it also allows each locus to be assigned a distinctive "signature" (Figure 3.3b,c; Wong *et al.*,1987) by which further isolates from the same locus may be recognized.

3.3.4.2 *Clones detected by multi-locus probe screening*

The results of screening 100 clones for polymorphism by the method outlined above are summarised in Table 3.1, from which

Figure 3.3

Screening of human DNA inserts from Charomid clones for probes which detect variable loci in human DNA (section 3.3.4.1). The inserts from clones which hybridized to at least one multi-locus probe were used as hybridization probes against Southern blots of *Mbo*I digested DNA from a standard panel of three unrelated people. This constitutes a screen for variability, as well as providing, for each locus, a characteristic 'signature' by which further isolates from the same locus may be recognised. Shown here are examples of clones which show (a) no variation; (b),(c) multiallelic variation (cMS605 and cMS610, see Table 3.2); (d) a largely monomorphic 'ladder' of hybridizing DNA fragments presumed to be derived from satellite DNA sequence.


(a)
(b)
(c)
(d)
kb
20
15
10
5
4
3
2
1

the relative performance of each multi-locus probe in detecting polymorphic loci can be assessed. Three common types of pattern were seen on Southern hybridization analysis of human DNA: (1) a single hybridizing fragment which showed no variation between the three individuals (Figure 3.3a), and was assumed to derive from a monomorphic locus, although some of these may derive from loci with low levels of variability; (2) no more than two hybridizing bands per individual, but polymorphic in size between them (Figure 3.3b,c); and (3) a largely monomorphic "ladder" of hybridizing fragments presumed to be due to satellite sequences (Figure 3.3d; section 3.3.2.3). Another less frequent Southern hybridization result was seen, in which multiple polymorphic fragments were detected. These were derived from two distinct loci, which both appear to be large polymorphic tandem arrays ("midisatellites"; see section 3.3.6.4).

33 clones detected simple polymorphic patterns on Southern blot hybridization. Of these, 8 showed patterns indistinguishable from loci previously detected in this library screen, and were assumed also to derive from those loci. Thus 25 distinct polymorphic loci were found in 100 clones screened. Of these, 8 showed only two alleles in the three people checked in the initial screen for polymorphism. All such loci were further checked with a variety of restriction enzymes (usually *Alu*I, *Hae*III, *Hin*fI and *Pst*I) to determine whether the polymorphism observed persisted with the other enzymes, and thus was due to variation in the length of a tandem repeat, or whether the polymorphism was only detectable with *Mbo*I, and thus was most likely due to a simple restriction site dimorphism for *Mbo*I. Two of the polymorphic loci were shown in

Table 3.2

Characteristics of the new minisatellite loci cloned in charomids. Heterozygosity was estimated from analysis of at least 40 unrelated Caucasian individuals. [1]The number of alleles shown is a conservative estimate of the number seen in *Mbo*I-digested DNA from 20 unrelated people. [2]After each chromosome assignment is shown the method of assignment (see section 3.3.5): (S) full somatic cell hybrid panel; (M) "mini-panel" of 4 somatic cell hybrids; (L) linkage. **Hinf*I alleles were used to estimate heterozygosity at this locus, as it gives a multi-band haplotype with *Mbo*I (see Figure 3.5a). +There is a frequent "null" allele at this locus (see section 3.3.6.2 and Figure 3.5b); the heterozygosity has been estimated assuming the "null" allele frequency $f_0 \approx 0.2$ deduced from pedigree analysis. ¶ Heterozygosity estimated from 22 unrelated females. § denotes loci at or near the ends of published linkage maps, or localized to subtelomeric regions by *in situ* hybridization (cf. Figure 3.6).

References

[a] Donis-Keller *et al.*(1987)
[b] Petit *et al.*(1988)
[c] Wong *et al.*(1987)
[d] Leppert *et al.*(1986)
[e] O'Connell *et al.*(1987)
[f] Nakamura *et al.*(1988a)
[g] Drayna *et al.*(1984)
[h] CEPH database,V3
[i] O'Connell *et al.*(1988)
[j] Clarke *et al.*(1984)
[k] Cooper *et al.*(1985)
[l] Colb *et al.*(1986)
[m] Jarman *et al.*(1986)

Table 3.2.

Probe	multi-locus probe(s)	HGM symbol	%hetero-zygosity	number of alleles^1	chromosome^2	nearest marker[ref.]
cMS440	MS1	D18S31	72	>10	18q (S,L)	CRI-L159 [a]
cMS600*	33.6	DXYS78	91*	>10*	X/Y (S,L)	§DXYS60 [b]
cMS601	33.15	D1S105	74	>10	1 (M,L)	CRI-L1226 [a]
cMS602	33.6,33.15, 3'HVR	D7S439	60	5	7p (S,L)	§MS31 [c]
cMS604	MS1(Alu)	D13S70	64	7	13q (S,L)	D13S6 [d]
cMS605	3'HVR(Alu)	D6S86	87	>10	6q (M,L)	§CRI-L1077 [a]
cMS607	33.15,MS1	D22S163	90	9	22q (S,L)	CRI-L1272 [a]
cMS608+	MS1(Alu)	D12S40	67+	9	12p (S,L)	§VWF [e]
cMS609	33.6	D21S155	66	5	21q (M,L)	CRI-L427 [a]
cMS610	MS1, 3'HVR	D19S77	80	>10	19p (S,L)	pJCZ3.1 [f]
cMS613¶	33.15	DXS438	35¶	4	Xq (M,L)	§DXS15 [g]
cMS614	33.15,MS1(Alu)	D10S92	77	6	10 (S,L)	CRI-JD12 [h]
cMS615	(Alu)	D18S32	51	4	18 (M,L)	D18S6 [i]
cMS616	33.6,33.15	D18S33	51	5	18q (M,L)	CRI-L159 [a]
cMS617	33.15	D20S26	79	8	20q (M,L)	§CRI-L355 [a]
cMS618	33.6,33.15, 3'HVR	D12S41	83	10	12q (M,L)	§MS43 [c]
cMS619	MS1.1,$(GGGCA)_n$	D22S164	79	5	22q (M,L)	SIS [j]
cMS620	33.15	D15S86	91	>10	15q (M,L)	§D15S3 [k]
cMS621	33.15	D5S110	92	>10	5p (M,L)	§CRI-L334 [a]
cMS622	33.15	D10S90	83	9	10q (M,L)	§VTR.4 [l]
cMS623	MS1	D12S42	79	9	12q (M,L)	§MS43 [c]
cMS624	33.15,MS1	D16S263	36	3	16q (M,L)	§CRI-O89 [a]
cMS625	33.15,MS1, 3'HVR	D16S264	47	2	16p (M,L)	§ 3'HVR [m]
		mean	71			
		median	77			

this way to be simple *Mbo*I RFLPs, and have not been studied further. Thus 23 distinct length polymorphic loci were found among the 100 clones detected by the multi-locus probes; of these one had been isolated by cloning in λ (MS31, Wong *et al.*,1987). The 22 new "VNTR" loci, together with cMS615 (below), were characterised further (*v.i.*, section 3.3.5).

3.3.4.3 *Clones detected by the Alu consensus probe*

In addition to the clones detected by the multi-locus probes, seven clones were tested which hybridized positively with the consensus Alu probe but with none of the multi-locus probes. Of the seven clones tested, one (cMS615 in Table 3.2) showed a multiallelic length polymorphism.

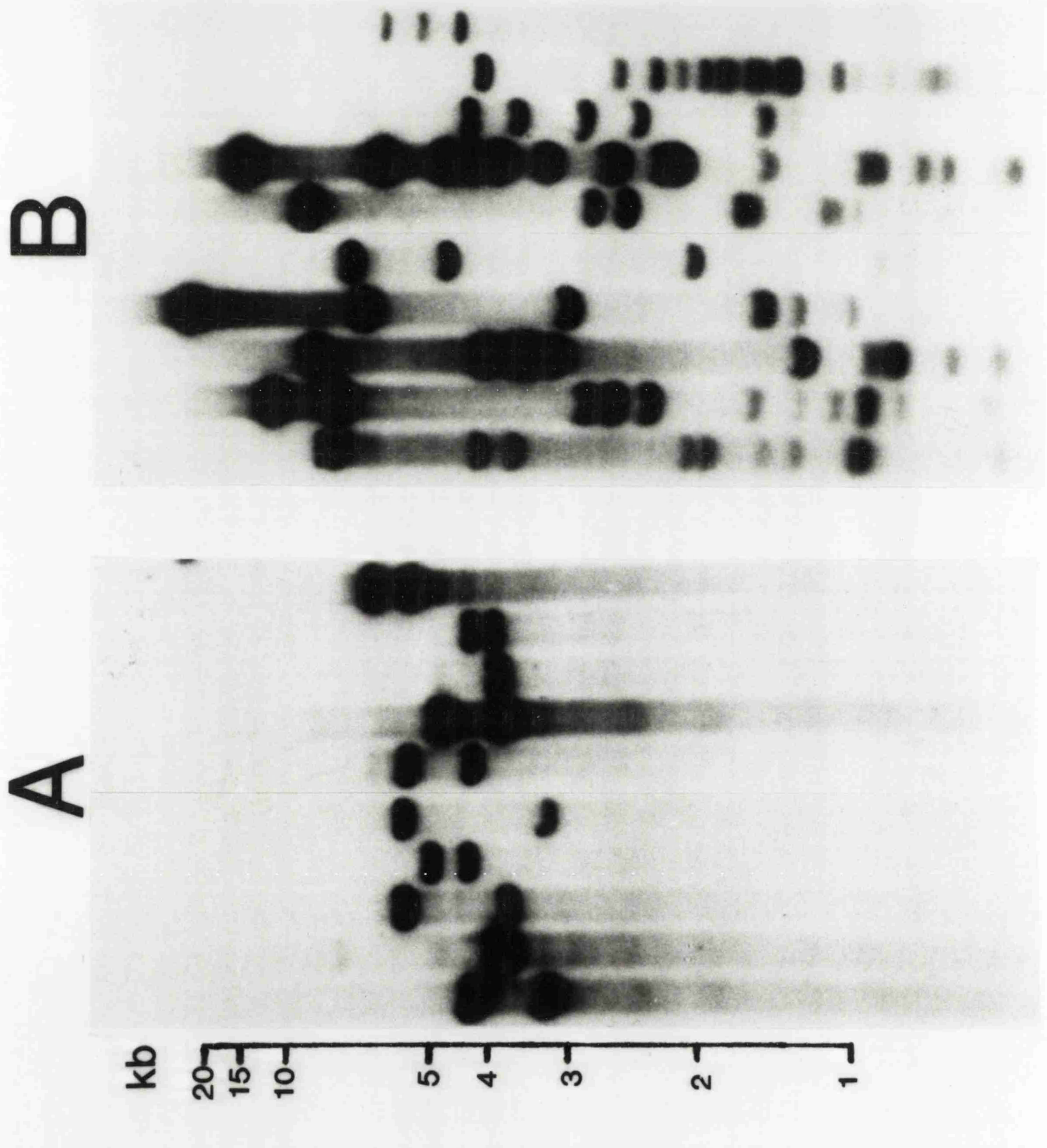
3.3.5 *Characterization of polymorphic loci*

Polymorphic loci were further characterized by assessing variability, inheritance, chromosomal localisation and mutation rates. This was done by probing DNA from 20 unrelated people, six three-generation pedigrees from the CEPH panel and DNA from somatic cell hybrids. Population variability was estimated from the 20 unrelated people and the 24 (unrelated) grandparents from the six CEPH families. These families served to establish the mode of inheritance, to determine segregation and to screen for length change mutation. In all cases the initial work was done using *Mbo*I-digested DNA, since this was the enzyme used in cloning the loci and so could be relied upon to give hybridizing fragments of scorable size (Figure 3.4a); however, in one case (cMS600, section 3.3.6.2) the complexity of the Southern blot phenotype with *Mbo*I (Figure 3.4b) led to the assessment of heterozygosity using *Hin*fI.

The chromosomal localisation was established by a

Figure 3.4

Simple and complex Southern blot phenotypes detected by hypervariable minisatellites (section 3.3.5). DNA from 10 unrelated individuals was digested with *Mbo*I and Southern blot hybridized with radiolabelled inserts DNA from (A) cMS605 or (B) cMS600. (A) shows a simple pattern, giving no more than two hybridizing allelic DNA fragments per individual, whereas in (B) many bands are seen in each individual, with all bands derived from the same locus as shown by family analysis (see Figure 3.5a).


A
B
kb
20
15
10
5
4
3
2
1

combination of somatic cell hybrid analysis and linkage mapping. The current CEPH database allows rapid assignment of loci by linkage, but is prohibitively tedious if more than about six chromosomes need to be searched. The most rapid mapping method used a "mini-panel" of just four somatic cell hybrids (section 2.1.5) which were sufficient to narrow the search to a subgroup of one to five chromosomes. Linkage mapping was then used to find the locus within one of the chromosomes identified. In eight cases (cMS440*, 600*, 602*, 604*, 607*, 608, 610 and 614) a full somatic cell hybrid panel was used, so that the loci could be assigned on these grounds alone; the analyses for the loci with asterisks were kindly performed by Drs. Sue Povey and Stephen Jeremiah at the MRC Human Biochemical Genetics Unit, Stephenson Way, London. Linkages were established for all the loci tabulated in Table 3.2.

Where no recombinants were found with a locus in the database, the newly-cloned locus was distinguished from the locus in the database by comparison of the genotypes of individuals in the CEPH panel. In the simplest case, an individual homozygous at one locus was heterozygous at the other. However, since the two studies used different restriction enzymes, it remained possible that the difference arises simply from a flanking RFLP for one of the enzymes. However, even allowing for this possibility, two or more individuals appeared *reciprocally* discrepant (for example individual X 1/1 at locus A, 1/2 at locus B, individual Y 1/2 at locus A, 1/1 at locus B). In these cases either the flanking polymorphism with one enzyme exactly compensates for the length variation shown by the other, or, more likely, the two loci are distinct. To such evidence could be added the general properties (number of alleles, heterozygosity level), which were often also discrepant between the newly-cloned locus and the database locus.

In two instances a locus mapped by linkage gave unexpected results with somatic cell hybrid DNA. The pseudoautosomal locus detected by cMS600 appeared to be present in the hybrid DUR4R3 (Solomon *et al.*,1976), although the X chromosome appeared to be absent from this hybrid. However, the absence of the X chromosome from this hybrid was ascertained from the absence of enzyme activity, and thus it remains possible that the signal is due to an inactive X chromosome in this hybrid (S.Povey, personal communication). The locus on chromosome 10 detected by cMS614 also appeared to be present in DUR4R3 despite the supposed absence of chromosome 10 from the hybrid. Furthermore, cMS622, which maps to the long arm of chromosome 10 (Table 3.2), failed to give a signal with this hybrid, which suggests that cMS614 may be identifying small fragments containing sequences from the short arm of chromosome 10 in this hybrid.

3.3.6 *Loci isolated from the library*

3.3.6.1 *Variability and mutation*

The 23 loci tested for population variability had a mean heterozygosity of 71% and a median of 77% (Table 3.2). Since the occasional isolation of a locus with a very low heterozygosity will have a large effect on the mean value, the median value, which shows that half the loci isolated had a heterozygosity greater than 77%, more fairly reflects the quality of the loci isolated. At only one locus were new mutant alleles detected; two new mutant alleles were observed in 100

Table 3.3

Comparison of relative success in minisatellite cloning by four approaches. Cloning of size-selected DNA fragments in λ vectors by Wong *et al.*(1987) is compared with cosmid cloning and screening with oligonucleotide probes (Nakamura *et al.*,1987a,1988b) and the ordered array Charomid library presented in this chapter.

Table 3.3

	Number of loci isolated	number of loci with heterozygosity						No. of loci with N alleles					
		0-50%	51-60%	61-70%	71-80%	81-90%	91-100%	N= 2	3	4	5	6-9	>9
Wong *et al.* (1987)	6	0	0	0	0	1	5	0	0	0	0	0	6
Nakamura *et al.* (1987a)	77	13	7	24	16	14	3	10	16	18	8	14	11
Nakamura *et al.* (1988b)	34	8	3	7	8	7	1	5	5	5	5	11	3
This work (Armour *et al.*,1990)	23	3	3	3	7	4	3	1	1	2	4	7	8

meioses at the D22S163 locus detected by cMS607 (see also section 4.4.3.3).

A comparison of four approaches to the isolation of variable minisatellites is made in Table 3.3. The salient feature is the small number but uniformly high quality of the loci cloned in λ by Wong *et al.*(1987). Both the work presented in this thesis and the studies of Nakamura (1987a;1988b) show that cosmid-based cloning seems to lead to the isolation of a wider variety of loci, but which includes a large proportion of much less variable loci. Comparing Nakamura's work with that presented here, the distribution of loci cloned in Charomids includes a higher proportion of the more variable multi-allelic loci compared with those cloned in cosmids, as expected, given that the Charomid library was constructed from a size-selected fraction enriched for the most variable loci.

3.3.6.2 *Inheritance*

Most of the loci showed codominant mendelian inheritance in the pedigrees studied, usually with a single hybridizing fragment per allele, although occasionally with alleles consisting of a pair of fragments segregating together. An extreme example of multi-band alleles is shown by the pseudoautosomal locus detected by cMS600 (Figure 3.5a); the *Mbo*I fragments inherited together as a linked haplotype are bracketed together in the figure. *Hin*fI cleaves to give only two or three larger hybridizing fragments, suggesting that this locus consists of large (5-35kb) arrays of tandem repeats which occasionally contain a recognition sequence for *Mbo*I, but much less frequently for *Hin*fI. This locus bears a superficial resemblance to another pseudoautosomal locus, DXYS20 (Page *et al.*,1987) but the pattern of fragments produced by different restriction enzymes with genomic DNA suggests that the two loci are distinct (data not shown).

Departure from simple mendelian inheritance was shown by two of the loci. cMS613 detected a locus at which males were always homozygotes and from which alleles segregated in a sex-linked fashion. This locus (DXS438) was mapped further using linkage to near the end of the long arm of the X chromosome. cMS608 detected a locus (D12S40) which mapped to the short arm of chromosome 12 by linkage. In some families, a parent who appeared to be a homozygote failed to transmit the allele to all offspring (Figure 3.5b), suggesting that this individual was not in fact a homozygote but rather a heterozygote for the visible allele and an allele either devoid of minisatellite repeats or too small to be detected on Southern hybridization. The nature of this "null" allele at D12S40 is discussed further in section 5.3.3.

3.3.6.3 *Linkage mapping*

Mapping by linkage allowed the 23 new minisatellite loci to be placed approximately on the genetic maps of human chromosomes (Table 3.2). In all cases linkage was detected at a LOD score of greater than 3.0. However, in some cases relatively few mutually informative meioses were used to establish linkage, and while the general area inhabited by the locus is firmly established, the precise placement is subject to wide errors, particularly where issues of marker order are concerned. Furthermore, although some of the loci can be placed fairly precisely on the physical map by close linkage to physically assigned markers, the majority of the genetic markers used had no physical localisation.

Figure 3.5

Pedigree analysis of the loci detected by cMS600 and cMS608 (see section 3.3.6.2). (a) shows the co-segregation of groups of *Mbo*I DNA fragments detected by cMS600 in CEPH kindred 1362. Hybridizing fragments which segregate as linked haplotypes have been bracketed together and alleles named A to D. The inferred genotype of each individual is shown above the corresponding lane of the autoradiograph. Analysis in other pedigrees has confirmed that the DNA fragments are all derived from a single locus. (b) shows the segregation of *Mbo*I alleles at the locus defined by cMS608 in CEPH family 1341. The mother passes allele A or D to her offspring. In contrast, their father transmits either allele B or an allele too small to be detected clearly in this analysis, here designated O. The arrow points to an extremely faintly hybridizing DNA fragment which in this family appears to segregate with allele O, and may correspond to this "null" allele. Note that the paternal grandfather may be either a homozygote for allele B (BB) or a heterozygote for this allele and the undetectable allele (BO).

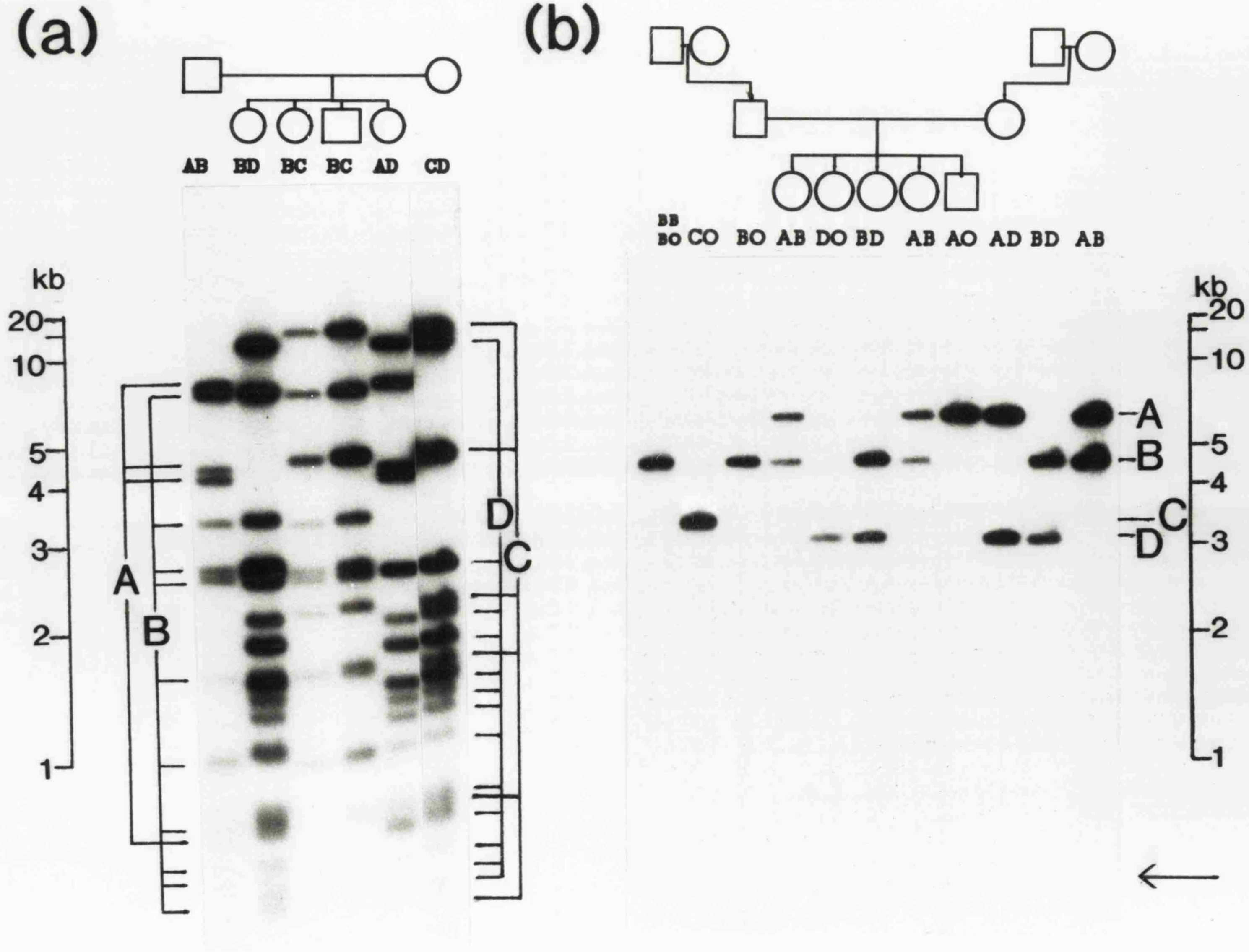
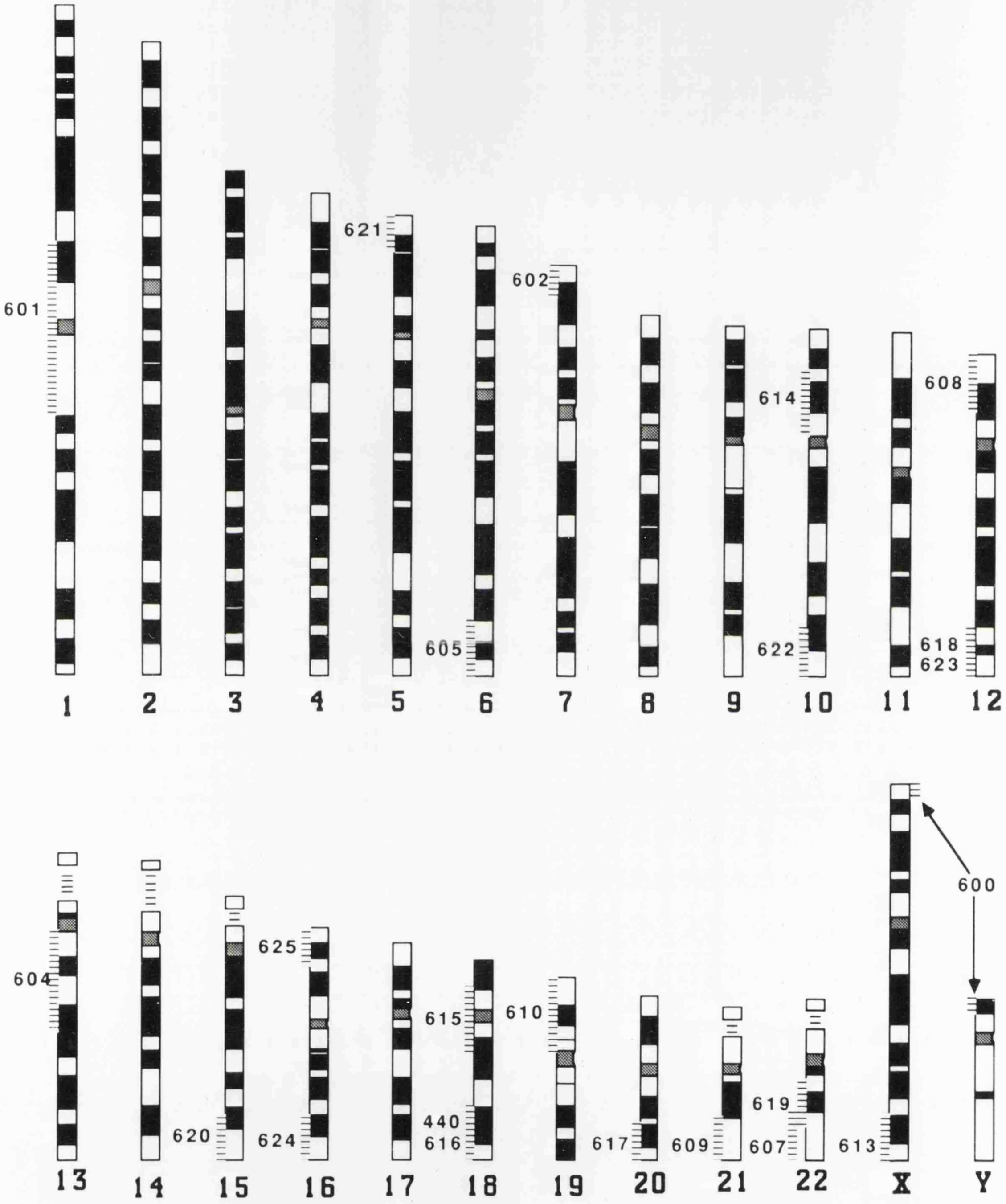

(a)
AB
BD
BC
BC
AD
CD
kb
20
10
5
4
3
2
1
A
B
D
C
(b)
BB
BO
CO
BO
AB
DO
BD
AB
AO
AD
BD
AB
A
B
C
D
kb
20
10
5
4
3
2
1

Figure 3.6

Schematic representation of the localisation of new minisatellite clones in the human genome (section 3.3.6.3). It should be noted that the imprecision in the physical locations of the loci is intentional; the assignments were made on the basis of linkage. In some cases, the placement is reasonably reliable, for example in the case of MS623, tightly linked to MS43, which has been localised by *in situ* hybridization to the telomeric region of 12q (Royle *et al.*,1988). In other instances, the placement is much more approximate, as in the example of MS601, which is tightly linked to FY, which in turn has been physiçally localised to 1p21-q23 (Donis-Keller *et al.*,1987).


601
621
602
605
614
622
608
618
623
1
2
3
4
5
6
7
8
9
10
11
12
600
604
625
620
624
615
440
616
610
617
609
619
607
613
13
14
15
16
17
18
19
20
21
22
X
Y

The loci mapped in this way were found to be on 14 autosomes and the X and Y chromosomes, and show a strong tendency to map to near the ends of linkage maps (Figure 3.6). Two examples of tight linkage groups ($\hat{\theta}$<0.05) were found. The first was on chromosome 18, and consisted of cMS440 and cMS616. The second consisted of cMS618 and cMS623, in turn both tightly linked to the minisatellite pMS43 (Wong *et al.*,1987;Royle *et al.*,1988) on chromosome 12. Looser linkage ($\hat{z}$=2.0 at $\hat{\theta}$=0.24) was also detected between two loci on chromosome 22, detected by cMS607 and cMS619.

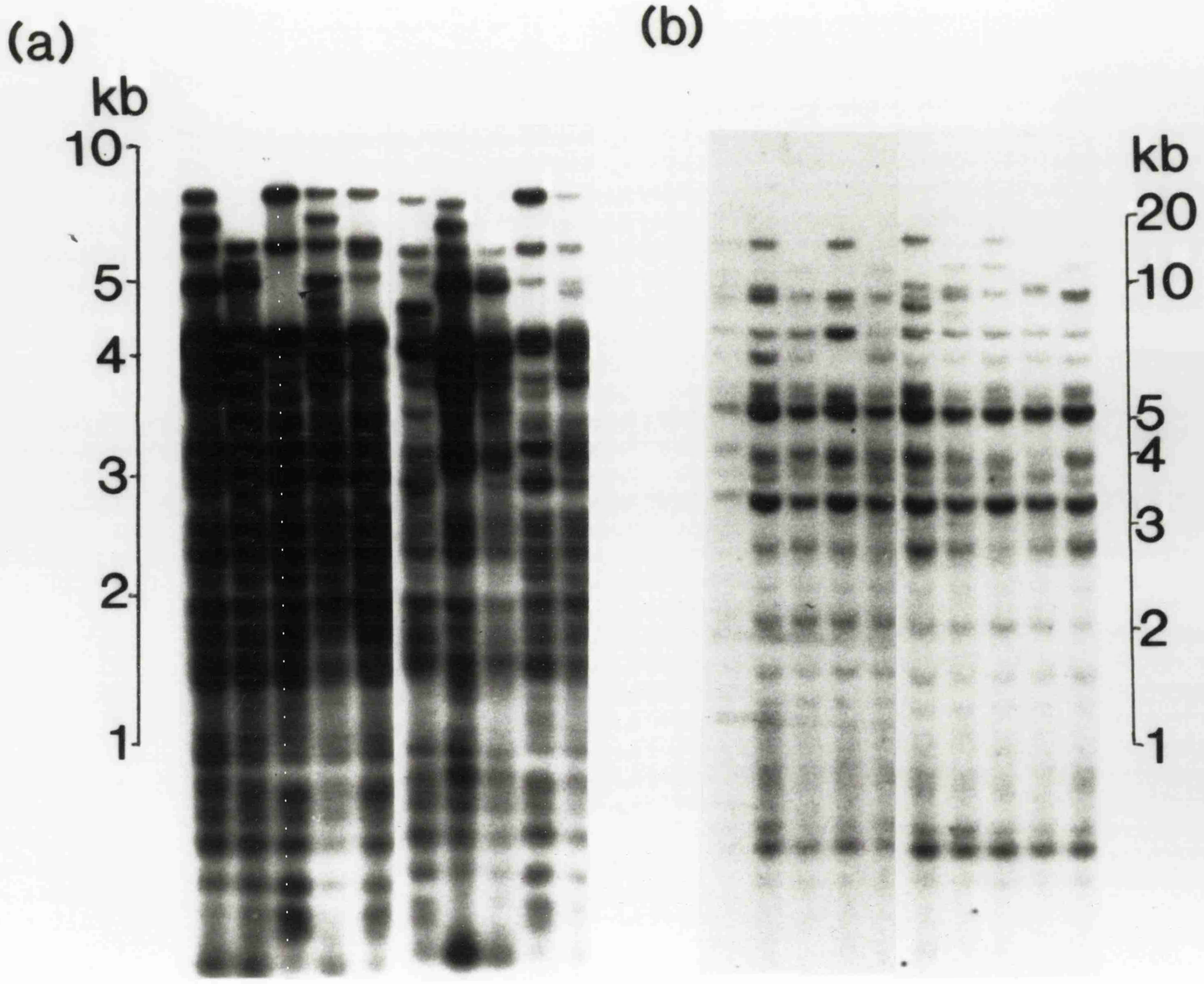
Thus the overall results of linkage mapping suggest a general dispersal of the cloned loci over the genome, in accordance with findings of genetic independence deduced from analysis of DNA fingerprints (Jeffreys *et al.*,1986), but also show the appearance of clusters of linked loci (predominantly at the ends of linkage maps) in agreement with predictions based on restriction mapping and sequencing (Royle *et al.*,1988;Armour *et al.*,1989b). The general discussion of the genomic placement of minisatellite loci is taken up again in section 4.2.2.

3.3.6.4 *Midisatellite loci*

Two clones detected complex Southern blot patterns in human DNA, consisting of many hybridizing fragments per individual, of which some were constant and some polymorphic (Figure 3.7). The clone VE2 (the name reflects the positioning in the ordered array; plate V, position E2) detected a locus at which most individuals had four or more polymorphic fragments (Figure 3.7a). Analysis in large pedigrees showed that all the polymorphic bands in an individual were very tightly linked to each other (no recombinants were seen) and so presumably were

Figure 3.7

Southern blot phenotypes at the "midisatellite" loci cloned in the Charomid library (section 3.3.6.4). (a) shows the profiles from *Mbo*I-digested DNA from 10 unrelated people probed with clone VE2; (b) shows *Mbo*I-digested DNA from 10 unrelated people probed with clone XVIID3A. Note that both probes recognize multiple monomorphic fragments, but also polymorphic fragments. The number of polymorphic fragments is clearly greater in the locus detected by VE2 (a) than by XVIID3A (b). The former appears to be the same locus as that described by Nakamura *et al.*(1987b), whereas the latter may be detecting fragments from more than one locus (see section 3.3.6.4).


(a)
kb
10
5
4
3
2
1
(b)
kb
20
10
5
4
3
2
1

derived from a single locus. Analysis of DNA from somatic cell hybrids showed that the probe detected no cross-hybridizing fragments at high stringency in mouse DNA, and that all the human fragments, not just the polymorphic ones, derived from a single human chromosome, chromosome 1. Linkage analysis showed that the polymorphic fragments were very tightly linked to the "midisatellite" locus detected by pYNI10 (Nakamura *et al.*,1987).

VE2 was used to probe human DNA digested with different restriction enzymes; comparison with the published account of the "midisatellite" detected by pYNI10 suggested that the two probes indeed recognised the same locus. However, apparent recombinants between the "two" loci have been detected; the genotypes detected by VE2 have been confirmed, and the discrepancy persists. Four of the "recombinants" can be attributed to two "double recombinants" in offspring from the same family, and would be nullified if the genotypes of the two children involved were exchanged. This strongly suggests that the discrepancy between our genotypes and those held in the CEPH database may be due to misreading of primary data, mistakes in input to the database, mislabelling of DNA samples or some other simple source of primary data error.

The other similar locus, detected by insert XVIID3A (the larger of two inserts in clone XVIID3) has not been characterized in the same detail. It, too, hybridizes to many fragments per individual in human DNA, but the average number of polymorphic fragments is lower than VE2, at about two (Figure 3.7b). Consequently it has not been possible to verify, by linkage between polymorphic bands, whether all polymorphic bands seen can be attributed to a single locus. Where an individual has two polymorphic fragments, however, they have been seen to segregate as alleles, suggesting a single locus at least in those individuals. Furthermore, somatic cell hybrid analysis suggests that hybridizing fragments may be contributed by two or more human chromosomes, and that the larger, more frequently polymorphic fragments may derive from a different chromosome from the smaller, constant fragments. However, the polymorphic fragments have not been successfully mapped by linkage on the most likely chromosomes (5 and 10) suggested by somatic cell hybrid analysis.

3.3.7 *Effectiveness of the cloning strategy*

3.3.7.1 *Genome coverage*

The frequency with which the same locus was isolated from the Charomid library can give a measure of the extent to which the analysis of 100 clones has exhausted the available loci: within the 100 clones tested, one locus (cMS612) appeared three times, and six (cMS440, 600, 601, 602, 616 and 625) appeared twice. This suggests that the mean rate of appearance of these loci among the 100 clones checked is about 0.7, and thus that if extended to all 185 positively hybridizing clones, another 12 new VNTR loci should be isolated.

This estimate, which is based on a simple Poisson model, makes the simple but incorrect assumption that all the loci have an equal chance of appearing in the library. However, at the most variable of these loci, *Mbo*I alleles will not be confined to the 4-9kb size fraction used to make the library. Since they will thus be underrepresented in that fraction, they will appear at a lower frequency relative to the monomorphic and minimally variable loci at which alleles always inhabit the

4-9kb size fraction.

In these analyses we are dealing purely with those loci which hybridize positively with a multi-locus "fingerprinting" probe, and other pools of variable loci may exist within the library but outside the subset defined by the multi-locus probe sequence. However, not all, and possibly very few of the loci cloned in the Charomid library actually account for hybridizing fragments in DNA fingerprints using the corresponding multi-locus probe (section 3.3.7.3). Thus some DNA fingerprints, for example those produced by probe 33.6, still have many loci which have yet to be isolated.

Thus while allowing the isolation of many more loci than seems possible in λ libraries of human DNA, this Charomid library may yet to be confined to the isolation of a subset of human minisatellites. Since it is likely that selection in *E.coli* is responsible for the failure to propagate some sequences, it may be that other host strains, for example those deficient in *RecBC* function (Wyman *et al.*,1985) will allow the isolation of a different subset of loci. Although the *recBC* gene product is usually associated with selection against inverted, rather than tandem, repeats, the use of strains defective in the functions of more recently characterized components of the recombination pathway, such as *sbcC* (Lloyd and Buckman,1985;Naom *et al.*,1989), may allow the isolation of a new subset of human minisatellite loci (see also section 3.3.8.1).

3.3.7.2 *Comparison with other cloning systems*

The variability of the loci isolated from the Charomid library has already been compared with those isolated in λ and cosmid systems (*v.s.*, section 3.3.6.1, Table 3.3). The general conclusions are that despite construction from very similar size fractions of human DNA, the λ and Charomid libraries contain different subsets of loci, with only one locus, that detected by λMS31, identified in both to date. Furthermore, the incidence of monomorphic loci in the Charomid library is much higher (about 50%) than in λ libraries; in the latter nearly all clones were from variable loci, but very often from loci already characterised.

A similarly paradoxical finding emerges from a comparison (Table 3.3) between the Charomid library and cosmid-based approaches (Nakamura *et al.*, 1987a, 1988b). Thus the frequency of appearance of polymorphic VNTR loci and their variability are approximately the same, despite the fact that the Charomid library was constructed from a size-selected fraction of human DNA estimated to enrich about 500-fold for highly polymorphic minisatellites.

The main advantage of the ordered array Charomid library is its convenience and high yield of information. After the initial drudgery of ordering the clones into microtitre plates, the rewards are many: only a single round of hybridization screening is required, since the clones are already isolated in their microtitre wells; clone DNA is easily prepared, by contrast with λ systems where minisatellite recombinants often yielded DNA at orders of magnitude below that expected (Wong *et al.*, 1986,1987); unlike in cosmid cloning, where the VNTR region within a cosmid insert needs to be identified and subcloned, the Charomid inserts correspond to the tandem repeat array with very little flanking DNA; and the ability to assign to each clone a grid position allows its profile of hybridization with many multi-locus probes to be assessed.

A useful illustration of this last point arose during the checking of clone inserts for polymorphism. It became apparent that MS1 cross-hybridized to a satellite sequence which was fairly abundant in the library. Insert DNA was prepared from one of these clones, and used to screen the library at high stringency. 16 clones hybridized positively, allowing them to be identified within the library array as clones containing that satellite sequence without the need to prepare DNA from each clone.

3.3.7.3 *How many loci are there left uncloned?*

The experience gained in cloning hypervariable loci by hybridization screening a library constructed from a 4-9kb size selected fraction is, unfortunately, not readily extensible to the genome as a whole, as the representation of a locus in a library depends upon its allele size distribution (*v.s.*, section 3.3.7.1).

Similarly, while our experience is chiefly restricted to minisatellites hybridizing positively with multi-locus probes, there is direct evidence that some of the loci cloned by hybridization screening do not appear in the corresponding DNA fingerprint. Thus whereas about 50% of clones from the Charomid library hybridizing positively with 33.15 detected monomorphic loci in human DNA, no monomorphic loci were seen in the 4-9kb size range in 33.15 DNA fingerprints (Jeffreys *et al.*,1985b). Moreover, while the clone cMS613, which detects a sex-linked locus (DXS438), was detected in the library by 33.15, no sex-linked loci were detected in genetic analysis of DNA fingerprints using 33.15. Thus the isolation of a locus by hybridization with a multi-locus probe does not guarantee that the locus is one of those that comprise the DNA fingerprint with that multi-locus probe.

The picture is complicated further by the fact that segregation data from many of the loci isolated by a similar hybridization strategy in cosmid libraries (Nakamura *et al.*,1987a,1988b) are not available in the CEPH database. We cannot therefore exclude the possibility that the loci cloned in Charomids overlap with loci cloned by Nakamura *et al.* However, the fact that no cases of identity were recorded between the loci cloned in Charomids and the loci cloned by Donis-Keller *et al.* (1987) suggests that we are still a long way off cloning most of the hypervariable minisatellites in the human genome.

3.3.8 *Extensions and prospects*

3.3.8.1 Cloning other subsets of loci

The loci cloned from the Charomid library to date have been mostly selected by hybridization screening with G/C-rich probes known to detect multiple hypervariable loci in human DNA. The use of other probes to select clones from the library may open up new subsets of loci to isolation.

For example, the success of the strategies of Nakamura *et al.* (1987a;1988b) in cloning many highly variable loci despite using libraries unenriched for large tandemly-repeated fragments may be due in part to their use of synthetic G/C-rich oligonucleotides as hybridization probes; a combination of oligonucleotide hybridization screening with the ordered array Charomid library may detect new variable loci. Similarly, the results of Vergnaud (1989) suggest that any probe consisting of tandem repeats of short random sequences may to some extent successfully detect new subsets of VNTR loci.

Isolation of clones from an ordered array library also allows direct comparison of the degree to which multi-locus probes overlap in detecting minisatellite clones. The fact that many of the positively hybridizing clones in the Charomid library were identified by more than one multi-locus probe suggests that there may in fact be considerable overlap between the loci detected by DNA fingerprinting probes currently thought to detect independent subsets of loci.

Movement away from the original "core" sequences may be made systematic and directional by the use of the ordered array library. Thus polymorphic probes detected by an initial round of screening with a "core"-like probe could be tested against human DNA at low stringency. Any which detected multiple variable loci could then be used to screen the library under similar conditions. The clones positive on the latter screen, but negative with the original "core" probe, would represent a movement away from the original subset defined by the "core" probe. This would make the approach described by Washio *et al.* (1989), that of "probe walking", efficient and directed. How far such a "walk" could be taken would be another useful indicator of the total size of the set of loci detectable by multi-locus probes.

As discussed above, other possible methods to broaden the range of loci cloned include the use of size fractions other than the 4-9kb *Mbo*I range used here, and the use of other *E.coli* hosts, particularly strains defective in other branches of the recombination pathway (section 3.3.7.1).

3.3.8.2 *Use of dispersed repeat probes*

As discussed below (section 4.3.2), DNA sequence analysis suggests that there is a significantly elevated frequency of dispersed repeat elements in the DNA immediately flanking minisatellites. This suggests in turn that dispersed repeat probes may alone serve as efficient indicators of clones containing DNA from hypervariable loci. This consideration prompted the use of an Alu consensus probe to screen the library.

There appeared in the library array to be an association between clones hybridizing positively with the Alu probe and those positive with minisatellite "core" probes. Of the approximately 2500 recombinants in the library, 140 (5.6%) were positive with the Alu probe, and 185 (7.4%, section 3.3.3.3) were positive with at least one of the multi-locus core probes. If independent of one another, one would expect about 10 clones to be positive with both the Alu probe and at least one multi-locus probe. In fact, 28 such clones were found ($p<0.01$). This association, however, may not necessarily reflect co-localization of these sequences in the genome. Since most of the recombinants in the library consist of satellite sequence, and since Alu elements will be underrepresented in blocks of satellite, the apparent association between the two types of clone may be explained simply by the fact that both would tend to inhabit the small compartment of the library consisting of non-satellite recombinants.

During the screening of Charomid clones for polymorphism, the finding of only one new hypervariable locus (cMS615) among seven Alu-positive clones tested did not represent a yield high enough, by comparison with clones identified by the multi-locus probes, to warrant large-scale screening. All the same, the Alu probe represents the only tested method for identifying clones likely to show VNTR polymorphism independently of DNA fingerprinting probes.

3.3.8.3 *"Negative" hybridization screening*

Another potential method which should enrich for minisatellites in a sequence-independent manner makes use of the fact that most recombinants in the library are derived from two or three classes of satellite sequence. If these were used as hybridization probes to screen the ordered array library at high stringency, the majority of clones which were *not* of interest could be identified, and the remaining clones would be enriched for minisatellites. However, the vast majority of remaining clones would be non-recombinants, although it should be possible to identify some of the remaining recombinants by screening with human dispersed repeat probes, or identify non-recombinants by hybridization with an oligonucleotide spanning the Charomid *Bam*HI cloning site.

3.3.8.4 *Other extensions*

In many laboratories, libraries have been constructed from specific human chromosomes or chromosome segments, often as part of studies on a genetic disease known to reside in that chromosomal location. In the initial stages of such work, where the locus in question needs to be localised to a precision of 1-2 megabases, the requirement is not so much for a physical map but for a series of informative genetic markers. The method applied here to the isolation of minisatellites from the whole human genome should be applicable to such chromosome-specific libraries of clones.

It is now well established that many non-human species show multiallelic variation at different loci with a number of DNA fingerprinting probes (Burke and Bruford,1987;Wetton *et* *al.*,1987;Georges *et al.*,1990), and ordered array Charomid libraries present a rapid and efficient method for the generation of useful numbers of locus-specific minisatellite probes from such species. Indeed, we have already been successful in applying the method to cloning minisatellites from Peafowl (Hanotte *et al.*,1990) and other bird species (T.Burke, personal communication).

CHAPTER 4

THE GENOMIC ANATOMY OF HUMAN MINISATELLITES

Das Eichhörnchen schließt nicht durch Induktion, daß es auch im nächsten Winter Vorräte brauchen wird.

The squirrel does not conclude by induction that it is going to need stores for the coming winter too.

Wittgenstein

Summary

The arrangement of human minisatellites in the genome is discussed. There appear to be no human chromosomes which are either significantly rich or lacking in the set of minisatellites studied. However, there appears to be a significant excess of hypervariable minisatellites near the ends of chromosomes. As suggested by this last observation, these subtelomeric regions have a high density of minisatellites, such that highly variable minisatellites in these regions are frequently clustered close together. Sequence analysis has confirmed the high density of tandem repeat regions in the DNA flanking other minisatellites, and has also shown that there is a high incidence of dispersed repeats in the DNA flanking minisatellites. Indeed, some minisatellites appear to have originated as tandem repetitions of sequences within dispersed repeats. The analysis of the internal structure of minisatellite alleles by minisatellite variant repeat (MVR) mapping is discussed, and examples presented.

4.1 *THE PROBLEM STATED*

4.1.1 *Levels of analysis*

The existence in the human genome of large numbers of highly repetitive structures, of which many show extreme levels of variability, is at first sight a little puzzling. As they appear to serve no obvious function, it is natural to inquire why human beings and many other organisms have acquired and tolerated them. Even if we are to accept, for example, some primary effect such as the facilitation of recombination (Jarman and Wells,1989), while this may have some bearing on the perpetuation of tandemly repeated minisatellites in the genome, it is probable that the explanation for the origin of these structures will be found elsewhere. It is difficult to accept that minisatellites have arisen *because* the cell needs a recombination signal; Wittgenstein's squirrel does not hoard nuts in autumn *because* it has solved the problem of induction. In trying to reconstruct an evolutionary history for these loci, the difficulty of distinguishing between a selected function (such as enhancement of recombination) and a selectively neutral "side-effect" of structures that have arisen for other reasons makes a purely structural discussion of the problem more rewarding.

In addition to accounting for the simple presence of tandemly repeated minisatellite loci in the human genome, some attention needs to be paid to the feature of these loci which has attracted most interest, namely their hypervariability. Thus a distinction is needed between those considerations which are sensibly applicable to any tandemly repeated block of DNA, and those which apply specifically to polymorphic loci, including the question of what it is that makes a minisatellite hypervariable.

These and other questions concerning the evolution of minisatellites are taken up in general terms in chapter 6; this chapter is concerned with what can be deduced about the origin of minisatellites from their arrangement in the genome.

4.1.2 *Ulterior motives*

Many of the analyses of the disposition of minisatellites in the human genome are concerned not only with the light they shed on the evolution of tandemly repeated DNA but also with more practical considerations. Thus, for example, the demonstration that the loci detected by DNA fingerprinting probes 33.6 and 33.15 are genetically dispersed in the genome (Jeffreys *et al.*,1986) is of great importance in assessing the power of such DNA fingerprints in paternity testing and linkage analysis. This general conclusion, indeed, has been challenged (Chimini *et al.*,1989) although it is fairly clear that their conclusion that the fingerprinting probes detect one or two "major" loci appears to be based on inaccurate interpretation of *in situ* hybridization and pulsed-field gel data (Jeffreys *et al.*,1990b). Similarly, the association between highly variable minisatellites and dispersed repeat elements (section 4.3.2) has been exploited in the cloning of human minisatellites (*v.s.*, section 3.3.8.2).

4.1.3 *Levels of description*

This chapter deals with the way in which evidence about the nature and origin of minisatellites can be gained from an understanding of their arrangement in the genome, and on the whole is entirely descriptive. The more speculative inferences which might be made from the data are reserved for chapter 6. The general disposition of minisatellites will be described in terms of their placement among and within chromosomes, their apparent clustering in subtelomeric regions, the high incidence of tandem and dispersed repeats in the DNA flanking minisatellites, and finally of the internal structure of individual minisatellite alleles.

4.2 THE CHROMOSOMAL DISPOSITION OF HUMAN MINISATELLITES

4.2.1 Placement among the human chromosomes

The placement of 32 human minisatellites cloned and mapped in this laboratory is summarised in Figure 4.1. The loci shown are those for which at least approximate subchromosomal localization has been made. The positions of loci placed by *in situ* hybridization (Royle *et al.*,1988;Armour *et al.*,1989b) are shown with filled bars. Those placed by linkage mapping (see section 3.3.6.3) have more approximate locations and are shown with broken bars.

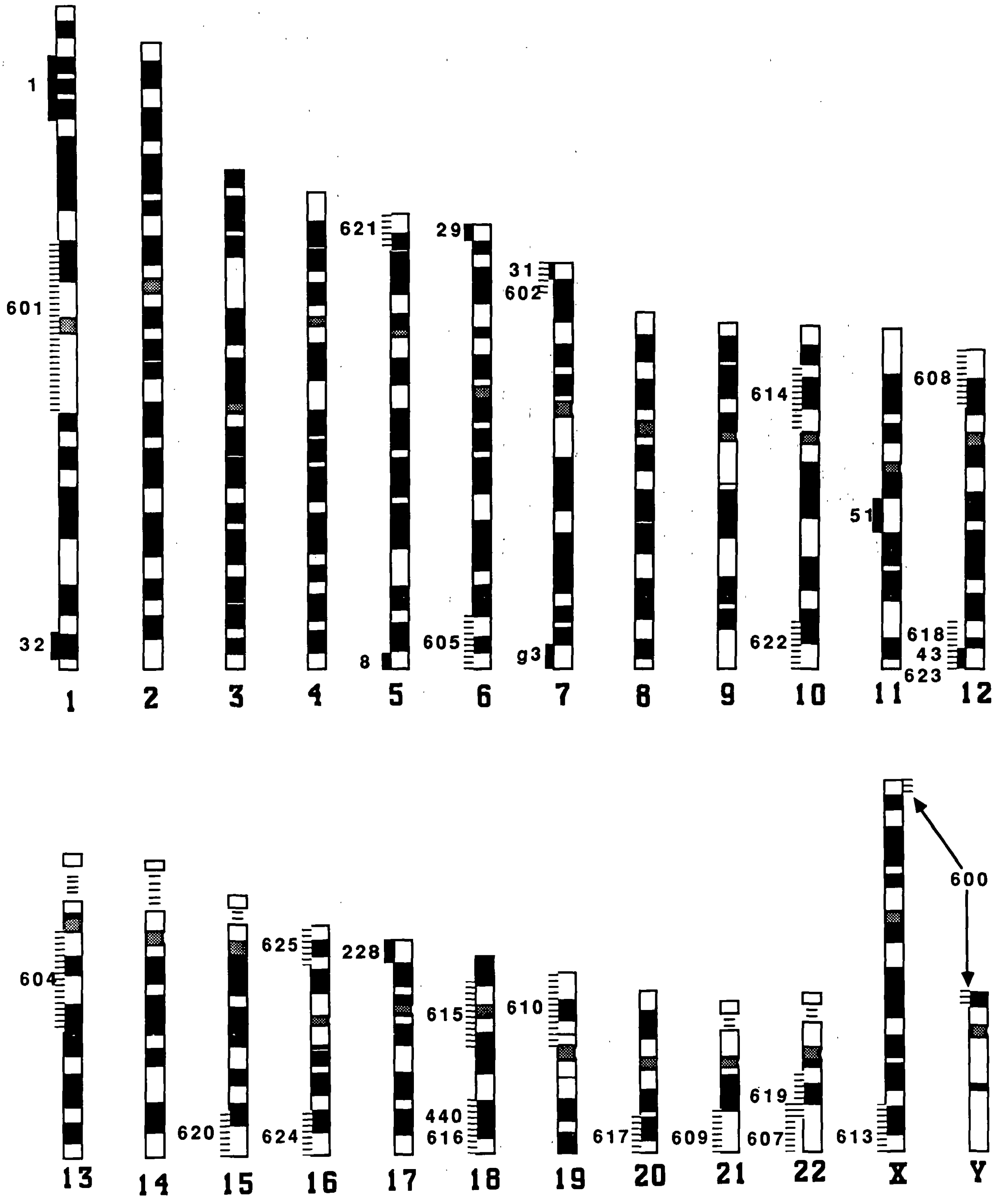
Despite the localisation of four minisatellites to chromosome 12 and three each to chromosomes 1, 7 and 18, the overall distribution does not depart significantly from a random scatter among the chromosomes. This conclusion holds whether one assumes that the probability of a chromosome having a minisatellite placed on it is uniform ($\chi^2_{[4\ d.f.]}=0.276$, $p>0.1$) or proportional to the length of the chromosome ($\chi^2_{[22\ d.f.]}=25.12$, $p>0.1$); details of the models tested are given in the legend to Figure 4.1. Similarly, while there are no minisatellites in the collection analysed which localize to chromosomes 2, 3, 4, 8, 9, or 14, the dearth here is not significant. There appears, however, to be a number of minisatellites localised to the smaller autosomes (15-22) which is disproportionate to their length ($\chi^2_{[1\ d.f.]}=5.53$ {with Yates' correction}, $0.05>p>0.01$), suggesting that the number of chromosome *ends* may be more important in determining a chromosome's minisatellite complement than its total length (*v.i.*, section 4.2.2.).

Figure 4.1

Distribution of cloned and mapped human minisatellites in the genome.

The location of minisatellites assigned to a genomic location by *in situ* hybridization (Royle et *al.*,1988;Armour *et al.*,1989) are shown with solid bars, for example MS228 on chromosome 17. The locations of loci isolated in this study and assigned by linkage (see section 3.3.6.3) are shown with interrupted bars, as with MS605 on chromosome 6.

The overall distribution of these loci does not depart significantly from a random distribution according to either of two models. In the first, it was assumed that the probability of a locus being found on any one chromosome was proportional to the length of that chromosome. An expected distribution was constructed by allotting to each chromosome a number of loci proportional to its length, to a total of 32. This was then compared with the observed distribution using the χ^2 test with 22 degrees of freedom. In this case, χ^2=25.12 and $p>0.1$. A second test compared the number of chromosomes which harboured 0, 1, 2, 3 and $\geqslant 4$ of the loci with the distribution expected from a Poisson process with a mean probability of a chromosome having a locus of 32/22. This model assumes that all chromosomes, regardless of length, have equal probabilities of harbouring a minisatellite locus. This comparison, made using the χ^2 test, here with 4 degrees of freedom, gave a χ^2 value of 0.276 ($p>0.1$). In both cases the Y chromosome was ignored, and the sex chromosomes assumed to consist of two copies of the X.


1
601
32
621
8
29
605
31
602
g3
614
622
51
608
618
43
623
1
2
3
4
5
6
7
8
9
10
11
12
604
625
620
624
228
615
440
616
610
617
609
619
607
613
600
13
14
15
16
17
18
19
20
21
22
X
Y

4.2.2 *Placement within human chromosomes*

The work of Royle *et al.*(1988) established that there is a strong tendency for human minisatellites to appear near telomeres. These conclusions were based on *in situ* hybridization studies using six cloned hypervariable minisatellites (Wong *et al.*,1987), of which four localised to the terminal Giemsa-staining band. Since then, two more human minisatellites have been localised by *in situ* hybridization (Armour *et al.*,1989b), of which one (MS51) was found to be interstitial, the other (MS228) near the telomere of the short arm of chromosome 17 (Figure 4.1).

These direct studies show the preferential, but not exclusive, localisation of human minisatellites to subtelomeric regions. Indirect evidence comes from the construction of genetic maps, in which the most informative markers are preferentially found near the ends of the linkage maps. For example, the genetic map of chromosome 19 produced by Nakamura *et al.*(1988a) contains two VNTR markers with more than 5 alleles each, which are both found to map, tightly linked to each other (*v.i.*), at one end of the linkage map. Similarly, another subtelomeric region, the human pseudoautosomal X-Y pairing region, is rich in hypervariable tandemly repeated sequences (Cooke *et al.*,1985;Page *et al.*,1987;Petit *et al.*,1988).

The localisation of the loci cloned in Charomid vectors in this study (section 3.3) similarly shows a pronounced tendency towards the ends of chromosomes (Figure 4.1). The physical locations of many of the markers used to establish linkage is in many cases imprecisely known or not at all, and so for many of the new loci it is not possible to make firm physical localisations on the basis of linkage mapping. Nevertheless, by the criterion of tight linkage ($\theta<0.1$) to markers either localised to subtelomeric regions by *in situ* hybridization or which are one of the two markers at the end of a published linkage map, 13 of the 23 new loci map to near the ends of chromosomes.

4.3 *LOCAL STRUCTURE*

4.3.1 *Clustering of human minisatellites*

4.3.1.1 *Evidence from linkage mapping*

If, as suggested in section 4.2.2., there is a strong tendency for minisatellites to appear near the ends of chromosomes, then one might expect minisatellites to be closely packed in these regions. That this is the case is supported by evidence from linkage analysis, restriction mapping and sequence analysis.

As remarked above (4.2.2), published linkage maps of human chromosomes show that many of the most informative markers are found at or near the ends of the genetic maps. Thus it is not surprising to find that in these regions the highly informative markers are also often tightly mutually linked. Some of these regions of high minisatellite density also appear to be areas of genetic map expansion in male meiosis. Thus, for example, the genetic maps of human chromosomes 7q and 16p of Donis-Keller *et al.*(1987) appear to show terminal map expansion preferentially in male meiosis, and a similar but more pronounced phenomenon has been documented for the human pseudoautosomal X-Y pairing region (Rouyer *et al.*,1986).

Similarly, among the new loci cloned in this work, two tight linkage groups are found near the ends of chromosome arms (Figure 4.1). One is on chromosome 18q, between MS440 and MS616 [$\hat{z}$=9.9 at $\hat{\theta}$=0]. The other is on chromosome 12q, between MS618 and MS623 [$\hat{z}$=10.6 at $\hat{\theta}$=0.04], in turn both linked to MS43 (Wong *et al.*,1987, Royle *et al.*,1988) [MS43/MS618:$\hat{z}$=16.6 at $\hat{\theta}$=0.04;MS43/MS623:$\hat{z}$=19.0 at $\hat{\theta}$=0]. Similarly, the newly-cloned locus MS602 is tightly linked to MS31 (Wong *et al.*,1987) on chromosome 7 [$\hat{z}$=7.65 at $\hat{\theta}$=0.03]. Looser linkage [$\hat{z}$=2.0 at $\hat{\theta}$=0.24] was also detected between MS607 and MS619 on chromosome 22.

4.3.1.2 *Evidence from restriction mapping*

Evidence was provided by Royle *et al.*(1988) that minisatellites often occurred in extremely close proximity to one another. In one example, MS43 on chromosome 12, a large, highly polymorphic locus (MS43A) had been cloned on the same *Sau*3AI fragment as a smaller, less variable tandem repeat block (MS43B). These could be studied either separately or together depending on the choice of restriction enzyme, and this demonstrated both tight genetic and physical linkage between the two tandemly repeated regions (Royle *et al.*,1988).

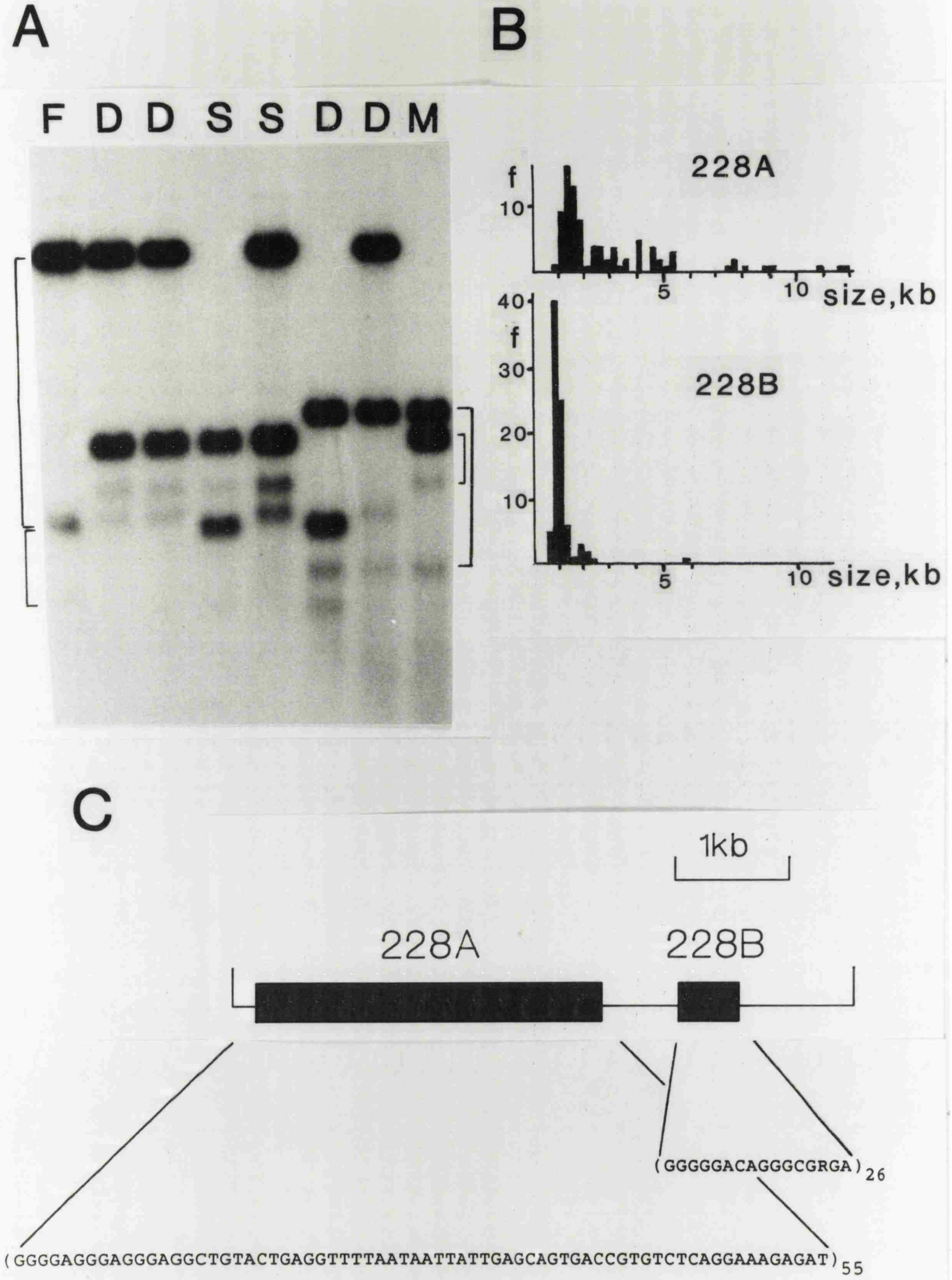
Another example of tight clustering between variable minisatellites was revealed by restriction mapping around the hypervariable locus detected by pλg3 (Wong *et al.*,1986). This demonstrated the presence of a highly polymorphic region of variable length in the DNA flanking the cloned locus (Royle *et al.*,1988).

An example similar to MS43 was found in the analysis of the locus detected by pMS228. This probe detected two distinct sets of hybridizing fragments on Southern blots of human DNA digested with *Alu*I (Figure 4.2a). It detected both a large, intensely hybridizing set of bands (228A) and a smaller, fainter set of fragments (228B). Family analysis (Figure 4.2a) showed that the two sets of fragments were derived from tightly linked loci (no recombinants seen in 25 fully informative offspring; $\theta<0.11, p>0.95$). Subcloning fragments from the *Sau*3AI insert of pMS228 and restriction mapping showed that the larger

Figure 4.2

Detection of two variable minisatellites by pMS228.

(A) shows the detection of two classes of hybridizing fragments on Southern blot hybridization of *Alu*I-digested DNA from a family group (F=father, M=mother, S=son, D=daughter) probed with pMS228. The strongly hybridizing fragments ("228A") and the weakly hybridizing fragments ("228B") both show polymorphism and are linked; fragments from each class in the parents which cosegregate in the offspring are bracketed together. (B) shows the size distributions of *Alu*I alleles determined from 89 (228A) and 48 (228B) unrelated individuals from the CEPH kindreds. Allele sizes at the two loci were grouped together in 0.25kb size classes. (C) is a schematic summary of the results of restriction mapping and subcloning of the *Sau*3AI insert of pMS228, which demonstrated that the strongly hybridizing bands (228A) are detected by a block of 69-70 base repeat units which occupies more than half the cloned insert, while the fainter minisatellite is detected by a smaller array of 15/16 base repeats.


A
F D D S S D D M
B
228A
f
10
5
10
size,kb
40
f
30
20
10
228B
5
10
size,kb
C
1kb
228A
228B
(GGGGGACAGGGCGRGA)26
(GGGGAGGGAGGGAGGCTGTACTGAGGTTTTAATAATTATTGAGCAGTGACCGTGTCTCAGGAAAGAGAT)55

of two tandemly repeated regions ("228A") was responsible for the detection of the larger, more intensely hybridizing bands, while the detection of the smaller, fainter hybridizing bands was due to a smaller repeat region, termed 228B (Figure 4.2c).

While 228A detected a typical hypervariable region, with a high heterozygosity (94%) and a wide range (0.8-12kb) of allele sizes (Figure 4.2b), the locus detected by 228B was unusual in combining a high level of variability (heterozygosity 85%) with a relatively restricted allele size range (0.6-5.9kb,Figure 4.2b). One of the limitations of amplification of minisatellites by the polymerase chain reaction is the large size of many alleles in the population at the most variable loci (Jeffreys *et al.*,1988), such that for many individuals only one or even neither of their alleles is small enough to be amplified efficiently. The combination of informativeness with limited allele size at the locus detected by 228B makes it a promising candidate for highly informative analysis, applicable to most individuals, by the polymerase chain reaction (see section 5.3.4).

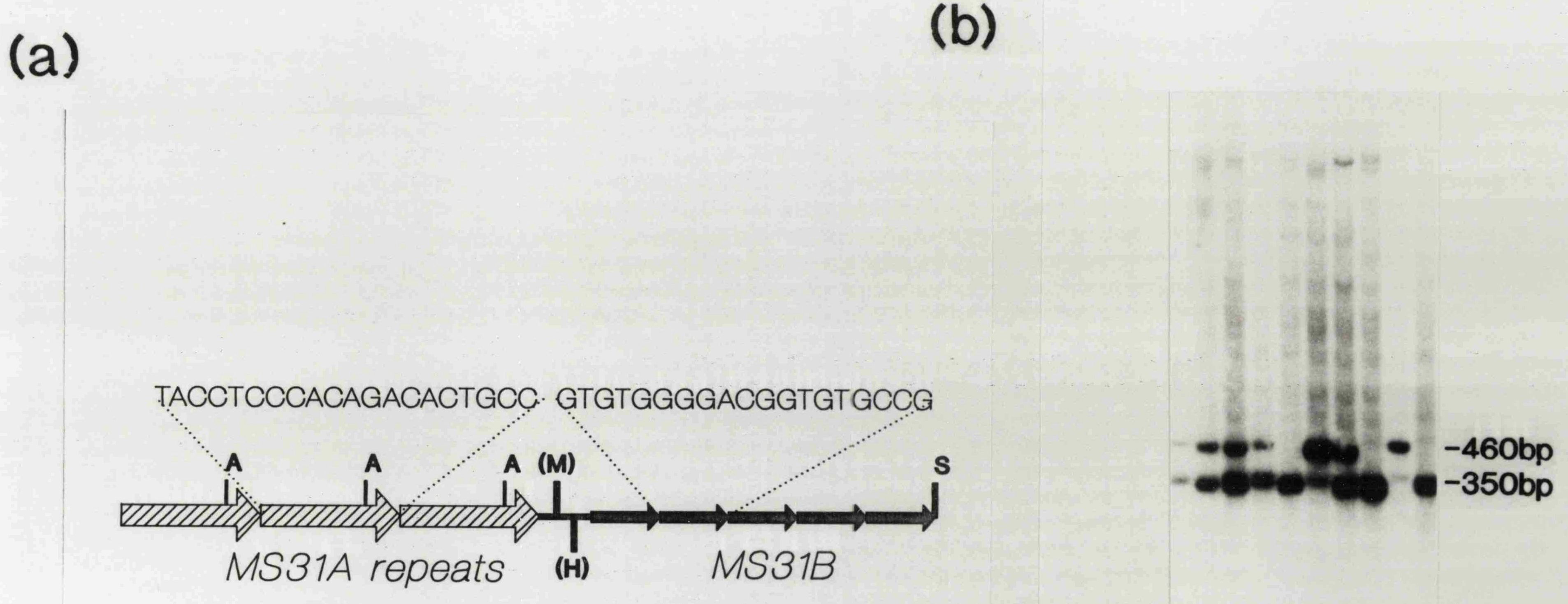
In both MS43 and MS228, it is the smaller, fainter locus which cross-hybridizes with the core probes (33.6 for MS43, 33.15 for MS228), and thus it is this smaller tandem repeat block which was responsible for the initial isolation of the locus. So the highly variable minisatellites MS43A and MS228A were isolated solely by virtue of their occupying the same *Sau*3AI fragments as these small minisatellites, and themselves do not cross-hybridize detectably with the "core" probes 33.6 and 33.15.

A further example of closely linked minisatellites comes from the sequence analysis of pMS31 (Wong *et al.*,1987;

Figure 4.3

(A) Schematic representation of the two minisatellites in pMS31. 31A has a 20-base repeat unit and accounts for most of the cloned insert, as well as for most of the variation at this locus. The smaller minisatellite, 31B, was only discovered on sequence analysis, and consists of a 19-base repeat unit which is G-rich on the opposite strand to 31A. A (*Alw*NI), S (*Sau*3AI), M (*Msp*I) and H (*Hae*III) indicate restriction sites; those in brackets are polymorphic in genomic DNA.

(B) Analysis of population variation at 31B. DNA from 10 unrelated people was digested with *Alw*NI and probed with a subclone containing 31B. The 31B region appears to be variable, but a contribution from repeat unit sequence variants at 31A cannot be excluded (see text). Some background hybridization to loci of similar sequence is also visible.


(a)
TACCTCCCACAGACACTGCC
GTGTGGGGACGGTGTGCCG
A
A
A
(M)
(H)
S
MS31A repeats
MS31B
(b)
-460bp
-350bp

v.i.,section 4.3.1.3), which revealed the presence of an unexpected second tandemly repeated region, thereafter termed MS31B. The population variability of this region was studied by subcloning the *Alw*NI-*Sau*3AI fragment containing it (Figure 4.3) and using it to probe Southern blots of DNA from unrelated people. These investigations were hampered by the fact that the *Hae*III and *Msp*I sites between MS31A and MS31B were themselves polymorphic (Figure 4.3a), such that in their absence a variable contribution from MS31A was included in the fragment detected. Digestion with *Alw*NI (Figure 4.3) suggested that MS31B was dimorphic in the population (heterozygosity about 50%), with two alleles of 350 and 460bp. However, since one of the bounding *Alw*NI sites is in an MS31A repeat unit, and since this site is not constant in MS31A repeats (see section 4.4.3), it remains possible that the observed variation is due to the presence or absence of *Alw*NI sites in the proximal repeats of MS31A, rather than reflecting length variation at MS31B.

4.3.1.3 *Evidence from sequence analysis*

The DNA sequence flanking the repeat arrays in the minisatellite clones pλg3, pMS1, pMS31, pMS32 and pMS51 were determined by Prof.Alec Jeffreys, Dr.Zilla Wong and Mrs.Vicky Wilson. These DNA sequences, together with the flanking sequences determined in this work (*v.i.*), are summarised in diagrammatic form in Figures 4.7 and 4.9. In these diagrams, regions of tandem repetition are shown by arrowheads, joined by a dotted line to indicate the omission of most repeats. As discussed above, the sequence of pMS31 demonstrated the presence of a second tandem repeat region close to one end of the clone. Like the main minisatellite MS31A, this smaller block was G/C-rich but with the G-rich strand in the opposite

Figure 4.4

DNA sequence data from the minisatellite clone pMS43 (two pages).

The entire *Sau*3AI insert was sequenced. Regions of tandem repetition are shown as aligned "paragraphs" of repeats; the two tandem repeat blocks (43A and 43B) responsible for the variability are shown in capitals. An incomplete Alu sequence element, which extends to the *Sau*3AI site defining the 3′ end of the clone, is shown in bold italics. Trailing dots (...) indicate that contiguous sequences have been split simply for the purpose of clear alignment. Gaps introduced to optimise alignment of repeats are shown with dashes (-). The tandem repeat arrays have not been sequenced in their entirety, and "nnn" indicates the boundaries of the sequenced region.

```
                   <-------------------- MS43A ------------------->
  1  gatctataca tgtTTACACACATGCCACACACCC-TTCCCAAGGCCGTCCCTATACCCA
                   TTACACACACACCACACCCCCCTTCCCAGGGCCGTCCCTATACCCA
                   TTACACACACACCACACACCC-TTCCCGGGnnn....
                                    ....nnnCCCGGGGGCCATCggc
     gcagtggcct cagcgt...
                                   ...ccag-cctccccgtcct
          ccggagctctctct----ggtgggggccccagtcctcctcgtctt
          ccggagctctctctctctggtgggggcccca...

                     ...cc tggccgacct gaaggtcagg gcatccactc

301  tttggggaaa cgaagttggg ttcgtagggg cagctctggt gttgttggtc

351  cttccctcca gacaccacgt ctcttcgtct ggctcccagg agaccccagt

401  tctatttcta tttctcccgc ctctgtcatc ccctcaaact gcttctgcat

451  agcccccact ctgcctctca ggctttaacc ccaactgtga gccccacctg

501  gctgtcctgg ggccactgaa ctccacgtgt ccatcgtggc tctccccaac

551  atgccaccac gcaccggctg gcattctgtc cctatgaagc cctgcctgcc

601  tggggactgg gaatcagccg ctgtcccagc...
                                  ...ccacccgtc
                                   aatcacccatc
                                   aatcaccggcc
                                   aatcacccgtc
                                   aatcaccgtc...
                                     ...tcttgcca gctttgcatc

701  ctaaatattt tttggaagag tcccgtgtcc gccacacccc acaagaaaga

751  aagttcaccc tactgggctc aacagtaaag atgttaatga agggaagctt

801  agggatgtgt ggttaggggc atagggacag acgggggcag tgcaacctca

851  ggggcggcaa cgtgcagttc tcgacaccgg cagagtgagg cctcgggcgc

901  cactcggacc acggtgcagc-ttccaac...
                  aacctgtggtgcctccacgggccaagac
                  agcctgc gtgcctccacgggccaagag...
                                     ...cacaaggt gggaggaggg
```

```
1001  aaggagctgg gggactggaa agcggtggtc agcatctgag ggcccggcac

1051  gTGGGGCTAGACGGG-GGATGTGGTGA  }
       TGGGGCTGGACGGGGGGGCGTGGTGA  }
       TGGGGCTGGACGGG--GGCGTGGTGA  }
       TGGGGCTGGACGGG-GGATGTGGTGA  }
       TGGGGCTGGTCGGGGGGGCGTGGTGA  }
       TGGGGCTGGACGGG--GGCGTGGTGA  } MS43B
       TGGGGCTGGATGGG--GGCGTGGTGA  }
       TGGGGCTGGACGGGGGGGCGTGGTGA  }
       TGGGGnnn                    }
              nnnCGGG--GGCCTGGTGA  }
       TGGGG acaggtgagt

1301  ctcctgtccc tcaccaggag aactgggttc tggttctcag cctgtctccc

1351  accgtagagc aacctctgac cgccagacag gacaccaagg tgaaaacttt

1401  caagcccca ggac...
                   ...gggagagaggtca
                   ...gggggagaggtcg
                    ...ggggagaggt...
                              ...ggcagggag
                                 ggcagggag
                                 gtcagggag...
                                ...tgg agtaaacgct gggcgcggcc

1501  acagctgtgc cttgtcaggg gatgaaggtg aaagacagaa cccgcggtga

1551  cacaggctcc cctgccccgc ggaggagtca ggccatcaga aaggccccta

1601  tgtcagccag gcgtggtggt tcacacctgg aatcccagca ctgtgggagg

1651  cagaggcggg cagatc
```

Figure 4.5

DNA sequence data from pMS228 surrounding the 228B minisatellite (two pages).

The data shown here come from a subclone which extends from the *Sau*3AI site at one end of the pMS228 insert to a *Kpn*I site between 228A and 228B. The 228B minisatellite repeat units are shown in capitals and aligned in a block. There is a complete Alu element, in inverted orientation, towards the 3′ end, which is shown in bold italics; the putative target sequence duplication around this element is underlined. Dots (...) indicate regions of contiguous sequence split for the sake of alignment.

```
   1  aggccacccc caccaacaac ataaccagag gggaaggaag tcaggcccct
  51  tctcactctg agaagctggt gcctgggatt ttaggctgtc acgacgattc
 101  cacccggcca gggcaggccc gaaccggccg gaggccacag gagaaccaat
 151  gagcctggct ggactcctgc aaaccttgag ggacgccgag attcacattc
 201  actgaatttt catgccacgt gacactcttc ttcttttgtt ccctcccgcc
 251  ctgccccacg gagccattta caaacataaa accattttta aaccattttt
 301  aaatggttta aaacctgttt ctgtaccaag tggtacagaa ataggggcca
 351  gcccccggtc cagcgccacg agctcctagg gccagagtgc aagagaggc
                ...ACAGGGCGAGAGGGG
                   ACAGGGCGAGAGGGG
                   ACAGGGCGAGAGGGG
                   ACAGGGCGAGAGGGG
                   ACAGGGCGGGAGGGGG
                   ACAGGGCGGGAGGGGG
                   ACAGGGCGAGAGGGG
                   ACAGGGCGGGAGGGGG
                   ACAGGGCGGGAGGGGG
                   ACAGGGCGGGAGGGGG
                   ACAGGGCGGGAGGGGG
                   ACAGGGCGGGAGGGGC
                   ACAGGGCGGGAGGnnn...
                        ....nnnGGG
                   ACAGGGCGAGAGGGG
                   ACAGGGCGAGAGGGGG
                   ACAGGGCGAGAGGGGG
                   ACAGGG...
         ...tcat cagggtgctt agggtgggct ccggggcgtc tgcaccacca
 701  ggcgcacagc ccaggaggtg gcaggagtca tccgttctga aacagccaga
 751  ggtacaacct cgtcgtccag gcaccggccg agttgggact cagggtcaaa
 801  gccaagctga ggcaacgtcg agatggaggg taacagccct cagcctgcac
 851  ctgccacact gcggaggccc cacaggaaca atccgggagg gtggggtggt
 901  gcctgcctgg cagctgcggg ggctgggtag ggaagggcta ctccaccctg
 951  gaggcccagc tcacaccaac ctcctagccc ctgacgtccc accaggcagc
1001  ttcacaaggt tacaggtcgg ttccttctcc actggattcc tcccacatcg
1051  ggtgacctga ccacacacac ggcaggtgcc cagcggtggt cccagcccca
1101  acatctcaag agcaggacac ccgagtggag atactaggtc acaggaatgt
1151  ctccacacag acattcaggc aggttcgagg gaagaagaca gcttcccggc
1201  cactctccca ccacgccaca cccggtgggc tcctctcctc aacctgggcc
1251  cacattctct cccaggttac tcacatcact cagtcatccc tcacatcact
```

1301 cacatcgccc catgacatcc gcctctgagc tgccagcctc cctccccagc

1351 ccctttcttc cttccctccc tccctccctc ***ctttcttttt tttgaaacag***

1401 ***gtcacccagg ctggagttca gtggcacaat cttgactcac tgcagcctcc***

1451 ***gcctctgggg ctcaagcctt ccaggctcaa gcaatcctca gcctcagcct***

1501 ***cccaagtagt tggaaacgta actgggcacc accatgccca gctatttttt***

1551 ***ttttttttca gtagagatga ggtctcacta cattacccag gctggtctca***

1601 ***gactcctggt ctcaagcaat cttctcacct tggcctccca aagtgctagg***

1651 ***attacaggtg tgagccactg cgcccgactt*** ccccagcccc tttctgaccc

1701 acagcctggg atc

direction to MS31A. MS31B was separated from MS31A by 12 bases apparently unrelated to either repeat unit. As detailed above (section 4.3.1.2), genomic restriction mapping suggested that MS31B was dimorphic in the population.

The DNA sequence of the minisatellite clone pMS43 is presented in Figure 4.4. The repeat unit sequence of each of the two minisatellite blocks had already been determined (Wong *et al.*,1987;Royle *et al.*,1988) and were confirmed in these studies. The presence and positions of the two minisatellites were thereby confirmed, and in addition to an Alu dispersed repetitive element (*v.i.*, section 4.3.2.1), the detailed sequence information showed the presence of numerous shorter tandem repeat regions, including two copies of a sequence immediately adjacent to MS43A (Figure 4.4).

The DNA sequence determined from pMS228 is from a 2kb *Kpn*I-*Sau*3AI fragment including the 228B minisatellite, which occupied about 600bp in this clone. The sequence data are presented in Figure 4.5, and in addition to confirming the presence of a tandemly repeated region, show that there is an Alu dispersed repeat element in the flanking DNA. The repeat unit sequence of the larger minisatellite 228A (presented in Figure 4.2c) was determined separately, from clones of random sonicated fragments, and no flanking sequence has been determined around this minisatellite.

A fourth example of the presence of two substantial tandemly repeated regions in a single cloned *Sau*3AI fragment was unexpectedly discovered during the sequencing of cMS607. DNA sequence data from this clone are presented here in Figure 4.6, and results of genomic mapping of the two minisatellites are presented in section 4.4.3.3.

Figure 4.6

DNA sequence determined at MS607 (two pages).

The sequence shown extends from the *Sau*3AI site defining one end of the cloned insert to a *Pst*I site about 150 bases from the other end. The remaining *Pst*I-*Sau*3AI fragment bearing the minisatellite MS607B proved extremely refractory to sequencing by dideoxy methods, but a clearly repetitive structure was nevertheless discernible. The MS607A repeat units are shown in capitals and aligned to show the repetitive structure. The sites of the PCR primers used in MVR mapping (see section 4.4.3.3) are shown in bold, and the *Pst*I and *Acc*I restriction sites defining the probe ("60714") used in indirect end-labelling are underlined (see also Figure 4.14b). Dots indicate where the sequence has been split in this figure to allow clear alignment, and dashes indicate where gaps have been introduced to optimise repeat alignment.

```
   1  gatccaccag gtgtgcaggg aggcgaggtg gggtcccggc ctctgtgtgc

  51  tgggttgggg gtcctggctc tgtctgtagg ggtgggggcc ctagctctgc

 101  ccagggaccc tacagcacct tgctcttccc ccaggcctgc cgctttgggc

 151  acgtgcagca tctggagcac ctgctgttct atggggcaga catgggggcc

 201  cagaacgcct cggggaacac agccctgcac atctgtcgcc tctacaacca

 251  ggtgcgactg tgtgtcctgc acatgcc...
                      ...TGCACCAGCGAGTGTGCATATACTTGCCTCTTCTGGGGGTGTATGTGTGT--GTGGGCAC-CAAGTG
GACCCTGTACGGTGATTGCATGTGTGCACC---GAGTGTGGATATACTTGCCTGTTCTGGGGGTGTACGTGTGTTTGTGTGCACACAGGT
GACCCTGTACAGTGATTGCATGCGTGCACCAGGGAGTGTGGATATACTTGCCTGTTCTGGGGTTGTACATGTnnn...
         ...nnnGCATGCGTGCACCAGGGAGTGTGGAAATACTTGCCTGTTCTGGGGGTGTACACGTGT--GTGTGCACACATAT
GACCCTGTACAGTGATTG...
         ...tgtagt gtcatccctg cctctgtgcc atggtataga tatgtgctct

 651  gtgtcctgca gcacagcctc ataggcatat gtgtgcacat ttgttctctg

 701  aacacacagg ggcttcacat gtgtgcacgt gtgttctgaa taaccaggta

 751  tgaattgggt acatctaggc cctctgcgag gtgagacctg agcgtgtata

 801  cctactggct tgtctctgca actcaggtgt acatggaaca aataggtgtg

 851  agtccgtgtg tgtgagcctg tgccctgcgc acgccatgtg tgcattcctg

 901  tgtgcgcatg tgctgttgtg ctcggatggt ctctccagcc acccagctgt

 951  gattccctct tccccgcaac aggagagctg tgctcgtgtc ctgctcttcc

1001  gtggagctaa cagggatgtc cgcaactaca acagccagac agccttccag

1051  gtacaccggt ggtttacagg agctcaaggc tgccccagag gtgtctgtct

1101  ctgtgtccat gtgacttgac ttctctgaac cttggttctt ccctggaagg

1151  ccctaaggga gcacctcccc caggactgcc cacaggaggt gttgggggac
```

```
1201  gagcccagca cgcgaggggt atttggtgtt gatgttccct tcgtcccctc
1251  gccagggaga gaggagggtc agcagggctc tggggcaggg gtatggggaa
1301  aatgagaaga ctggggtgac aggtgtgggt ctgacccccc aaccccgaga
1351  gaccgctagg ggtgcagaag ccaaactgca g
```

In the last three sections, different lines of evidence have been presented for a total of seven examples of clustered minisatellites. Furthermore, all seven of the examples are from loci which map to subtelomeric locations, suggesting that where minisatellite density is high, they may be very tightly crowded indeed. While in some cases the evidence, from linkage mapping, suggests grouping together on a fairly large scale, no fewer than four of the cloned minisatellites which have been studied in detail have two minisatellites so close together that they appear on the same *Sau*3AI fragment. In fact, nine cloned minisatellites have been fully characterized by sequence analysis, and of the four which contain more than one minisatellite in the cloned insert, all four map to subterminal locations. Three of the remaining five (MS1, MS32 and MS51) have been localised by *in situ* hybridization to interstitial locations.

4.3.2 *Association with dispersed repeat elements*

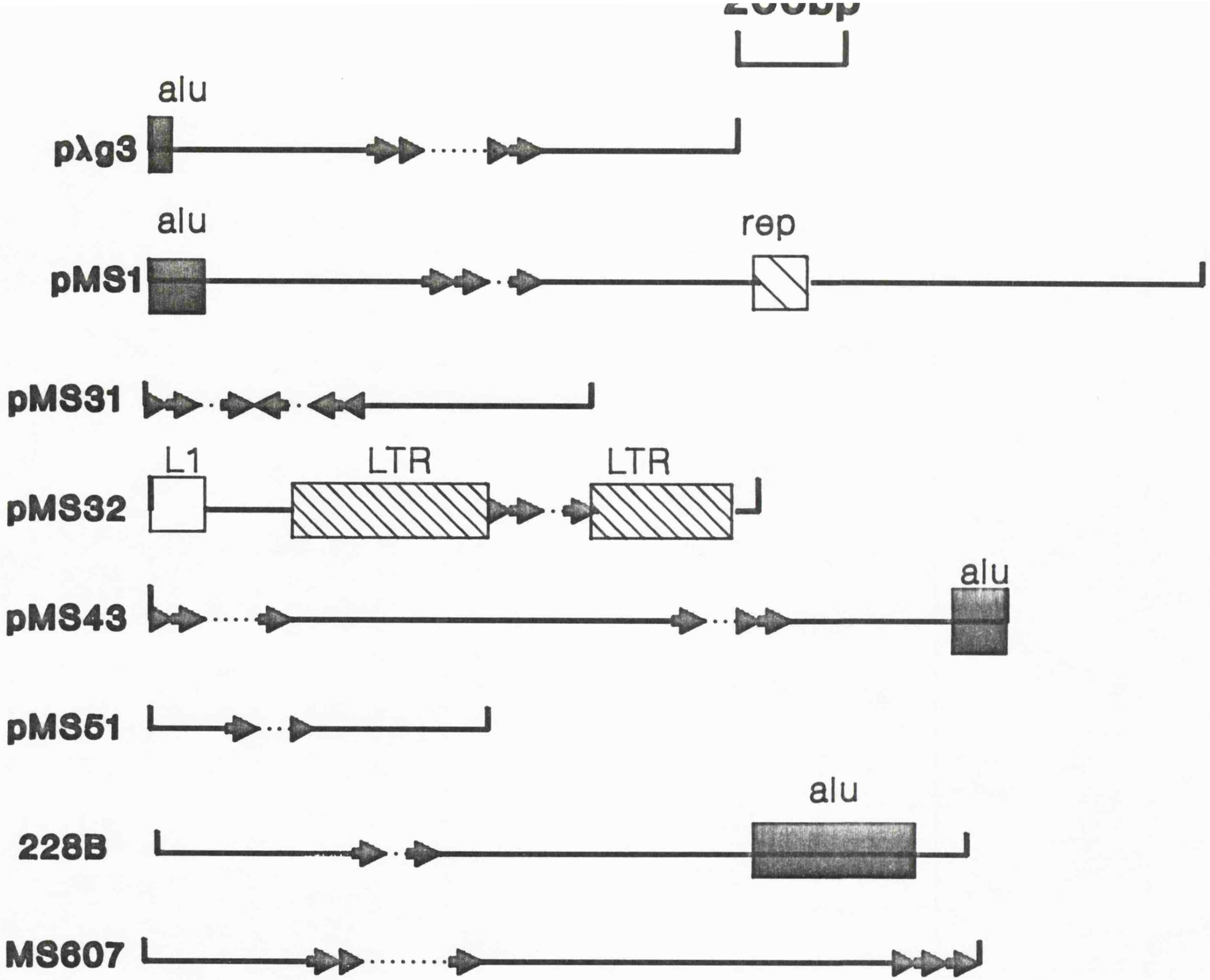
4.3.2.1 *Sequences flanking cloned minisatellites*

The determination of the sequences flanking the repeat arrays in pλg3 (Wong *et al.*,1986), pMS1, pMS31, pMS32 (Wong *et al.*1987) and pMS51 (Armour *et al.*,1989b) were determined by Prof.Alec Jeffreys, Dr.Zilla Wong and Mrs.Vicky Wilson. Computer analysis of these sequences was performed by Prof.Alec Jeffreys. These and four other flanking sequences (Figures 4.4, 4.5, 4.6 and 4.8) are summarised in diagrammatic form in Figures 4.7 and 4.9. In these figures the tandemly repeated regions are abbreviated to arrowheads separated by a dotted line, while dispersed repeat elements are shown as boxes. It can be seen that among the minisatellite flanking sequences

Figure 4.7

Summary of the features of DNA sequences flanking cloned minisatellites (see also Figure 4.9).

In order to show features of flanking DNA more clearly, the tandemly repeated arrays are shown in abbreviated form as arrowheads separated by a dotted line. If shown to scale most of each clone would be occupied by tandem repeats. Two distinct tandem arrays are present in pMS31, pMS43 and pMS607. Alu elements are shown as filled boxes; other dispersed repeat elements are (i) in pMS1 a novel 70-100 base dispersed repeat element ("rep", diagonally striped box) and in pMS32 (ii) an L1 dispersed repeat (open box) and (iii) a retroviral LTR-like element (striped box).


200bp
alu
pλg3
alu
rep
pMS1
pMS31
L1
LTR
LTR
pMS32
alu
pMS43
pMS51
alu
228B
MS607

there is a high incidence of dispersed repeat elements; in a total of about 12.3kb of flanking sequence there are eleven dispersed repetitive elements.

4.3.2.2 *Alu elements*

Eight of the eleven dispersed repeats found flanking minisatellites were Alu dispersed repeats (Schmid and Jelinek, 1982). In pλg3, pMS1 and pMS43 the *Sau*3AI site defining one end of the clone is within an Alu element, and so these clones contain only part of an Alu element, whereas full-length elements are found in the DNA surrounding 228B (*v.s.*, section 4.3.1.2) and MS608 (*v.i.*, section 4.3.2.5). In the DNA flanking MS608 no fewer than four Alu elements are found within 2kb of the tandem repeat array (Figures 4.8,4.9).

While substantial local variations are observed, one would expect an Alu element to occur about every 5-6kb in human DNA if randomly dispersed (Schmid and Jelinek,1982); thus the appearance of eight complete or partial Alu elements in just over 12kb of flanking sequence represents a significant excess of Alu elements in the DNA flanking minisatellite tandem arrays ($p<0.01$). However, there are other regions in which significant clustering of Alu elements is seen, for example in the non-coding DNA at the human tissue plasminogen activator gene, in which 28 complete or partial Alu elements, together with a partial L1 element and a substantial block (570bp) of 7 base tandem repeats, are found in 36.6kb (Friezner Diegen *et al.*,1986).

4.3.2.3 *Dispersed repeat elements in pMS32*

The first 90 bases in pMS32 bear a 64% similarity to bases 4572 to 4663 of the human L1 element consensus of Singer (Demers *et al.*,1986), suggesting that this is the start of a 5'

Figure 4.8

Sequence data around MS608 (four pages).

Sequence determination of the Charomid clone cMS608 showed that the *Sau*3AI insert contained repeat units and flanking Alu elements only; therefore, in order to determine flanking single copy sequence, a cosmid clone containing this region was isolated from a cosmid library [in Lorist 6, (Cross and Little,1986)] kindly provided by Dr.Brandon Wainwright, St.Mary's Hospital Medical School, Paddington. The sequence data shown in this figure are from a 6.1kb *Eco*RI-*Bam*HI subclone from this cosmid. The sequence shown only extends for 4.85kb since only the outermost repeat units from the 2kb minisatellite tandem array have been sequenced. The tandem repeats are shown in capitals and aligned, with dots (...) indicating that the two sequences so separated are contiguous and have been divided purely for clear alignment, and dashes (-) indicating where gaps have been introduced to preserve repeat unit alignment; Alu dispersed repeat elements are shown in bold italics; putative target site duplications around these Alu elements are shown underlined. The two *Sau*3AI sites (GATC) defining the ends of the original Charomid clone are also underlined.

1 gaattcccta gcctagaaca tattcaaata ttagcacttt taacctcaaa

51 gatttttgca aacctaacac gatagcagtt gtactaatac ggtttgttga

101 ttaaatacag acaaaaaact agtaggctgc aataatgttt taaaagaaag

151 ttgtatttta ttaatcacaa tagcatacca gcatttaagt aagtaggagc

201 acatactgaa atatattatt tagttggcct atagacaatg tttggtatgc

251 atatggattg aagacccaaa aattccacag ccaacttagg tatcctctgc

301 tagcggatcg aaaaccaaag ttggggcctc aaatgttggt gtataacaga

351 atcacctgtc aagcccccag cactaaccaa gtcttgcctg taattcagca

401 tttgtccaag agatgtcccc actcttctaa cagtcaaatg aaaggagcta

451 cagcgaaata atacaaaagc taagaataag cagtcaaaca acaatgtgat

501 ttgagtagaa acagaaaata caatgcttca ccgaaactcc acttttttacc

551 tctcggcttt tgggggggag aattgtttac ctgccagcta caaaatcatg

601 ttaataaaga aaacgtaaaa ctaattacaa atgttactac taaatagcta

651 caaatgccat tttaatgcta tcttatttca tcaaactaag atgccactga

701 ttataagaca caccactgtt gtacatgctg aaaagaaata gtggaaatgc

751 ctccaaataa atcatgacac aaaactttat cataaa***ttta gggttttgtt***

801 ***tgtttgtttt tgaggcagag tgtcactctg cacgctggag tgcagtggtg***

851 ***caatcctggc tcactgcaac ctccgcctcc tgggttgaag caattaccat***

901 ***gtgccagcct cacaaagagc tggattacag gcgtgcacca ccacactcag***

951 ***ctaatttttg tatttttagt agagacaggg ttttgccata ttgcccaggc***

1001 ***tggtctcgaa ctcctggctc aggcgatccg cccgcctggc ctcccaaagt***

1051 ***gctgggaata cagacgtgag ccattgcacc cgccct***agaa tttttcatta

1101 atactgattg caagagttag tttttttaaa cttatttaaa cagagtttaa

1151 tcacatacca tcataaactg ttttcataac tttctccata ccgggtaatt

1201 ttgtttcaat gctgattcac agaatattta tgcagaaaca ctgaacaaca

1251 aagttaacat gggaaatgcc aacaatgctc atacctcagc tgaaatgatg

1301 acaaaatgaa tggacatgat tattatttgt gaatacacgg gcaaacaaat

1351 ataccaggac tgctgctttg ctgatatgac tgtcaacatg acaccaagtg

1401 ttaacactta actcaatttc attgttaaaa cagggaagtt gtgactttta

1451 gaattgatta aatgt***gggcc gggcgcggtg gctcacgcct ataatcccag***

1501 ***catattagga ggccgaggcg ggcggatcac gaggtcagga gttcaagacc***

1551 ***agcctgacca acatggtgaa acctcgtctc tactaaaac agaaaatta***

1601 ***gctgggcctg gtgatgggtg cctgtaatcc cagctactcg ggaggctgaa***

1651 ***gcaggagaat cgcttgaacc caggggcgc aggttgcagt gagccgagat***

1701 ***cgcaccattg cactccagcc tgggcaacag agcgagactc cgtctcaaaa***

1751 ***aaaaaaaaaa gaa***ttgatta aatgccactt tttttaaaa gatggacaat

1801 aattcaataa ctaatgaaat ttctggatac cctagcattg ccagcccata

1851 attctttgtg ctgcttcctt gatacgctag tgatattctg tacatctcca

1901 gcaatcattt caagtcccca ttcccctaat ctagattaca gcagttaaat

1951 tccttactat actcccttgt tcactcttct ctccttcctc catttattct

2001 cagcacaacc agagtaaatt ttaaaagcaa atcagattgt gtcactccct

2051 tgattcaacg gtggaatggt tttctctccc attttgaata aaatggttgc

2101 attatctgca agggatcata aagggcaggg tactttttct caaagacagc

2151 atggatttgt ttgctgctat atccttagtg ctgtgcgcat ggttggtgtc

2201 cagtaaatac gtgataagtg aatctgactc tggccacctc tccaatttcg

2251 catctcatta agtcttcccc ttatacagta tattcttgcc ctccagcctc

2301 cctcctgttt cctcaaaact ccttcctatg aggtctttaa aactgttaat

2351 gcttcttcat gaaattattt tacttggcaa agctggcttc ttctcttcag

2401 atctccactg aaagggtacc tactaaagga cacttaccct gttataaaat

2451 ccccttccca acttaattag cactccaccc agaaagtctt tctgaacagt

2501 gctgatgcac agcattccat ccttgtaaga taaaaaaaga aacaaataat

2551 gaaggaatct ataaataata aaaat***gggc tgggtgcagt agcttacgcc***

2601 ***tgtaatccca gcactctggg aggccgagtc aggcagatca cctgaggtca***

2651 ***ggagtttgag accagcctgg ccaacatggc gaaaccccat ctctactaaa***

2701 ***aatacaaaaa ttagcttgtg cacgcctgtt gtcccagcta ctcaggaagc***

2751 ***tgaggcagga gaatagcttg aacccgggag gtggaggtta caatgagcca***

2801 ***agattgcgcc agtgcactcc agcctggg***ac aTAATAATAATAATAATAATAA
CAACAACAACAA...
CAACAACAACAACAATAATAATAATAACGGGCCAGGCATAGGCATATTGCCTGTAATCTCAGCACTTT
CAACAACAACAACAATAATAATAATAACGGGCCAGGCAT-GGCATATTGCCTGTAATCTCAGCACTTT
CAACAACAACAACAATAATAATAATAACGGGCCAGGCATAGGCATATTGCCTGT...
...GCCTGTAATCTCAGCACTTT
CAACAACAACAACAATAATAATAATAATGGGCCAGGCATAG-CATATTGCCTG
TAATCTCAGCACTTT
cagaagccaaggaggg aggattgcttg
aggccaggag ttcaagacca gcctaggcaa

3201 ***catagggaga ctctgtctct acaaaatttt tttttaatt taaaaattaa***

3251 ***caatgcatgg tggcatgcac ctgtagacct acctactagg gaggctaagg***

3301 ***cagaaggctc acctaagccc aggatttcaa gctgcagtg agccatgatc***

3350 ***atgccactgc actccagcct gagtgacaga gtgagaccct atctctaaaa***

3400 ***ataaaaacaa taaaaattta aaataaaaaa aaaaa***tgact ggaagcatat

3450 acataagtta ataattatta ctaggtaaca atactttgag ttataagttt

3500 ttgttaatct atttcccaat tttactacaa tggatttcca gtgtaataac

3550 atcatgtttt taaaaaacaa tgatgttcta cttcaaaaat aaaaatgccc

3600 tcactgttaa aaaagatttc agacttttta caaaaggaaa aattatagaa

3650 caacctacaa atttagaagt atttttaaaa attattattt tgctgctttg

3700 ggaatacgcc atttttttaaa cctccttcct cacattttat ttaaaaagga

3750 attattcaga ctcatactaa cacagattag ggatttattt tcccctgccc

3800 aaaagtgaaa attacctgaa gtccaactat atatagaaaa gttattatgg

3850 ctattaattt tttaattcaa tataagatgg tttcttcctt tacgccttct

3900 taaaatgata cattatttag aaaagtaaaa aattatcagt agtcacctag

3950 gtcagctaaa aataagtaac actgtgtctc tacatcactc tctccagcta

4000 tgcctccatg gcagccaatg atttttatca cgtatctttt acagcaaggc

4050 aacagtttcc atggaaacca cagggcaaat atccattctt gctgccaagg

4100 acttttataa gcacacataa gctgagataa ccgcttactc tcttttacta

4150 atatttttaa attaagcaat ggttagctat tctgattctg tttgttgtgt

4200 ttgttgtctc tatgaacaga tctagagatg agtgaaaaat caaggaattt

4250 tttacccctt tagcaacaaa tgaactctca aaatcagggg agcaggagag

4300 aaaagaccca gaaagtcact gagccagcat acataccaga caatgtcatc

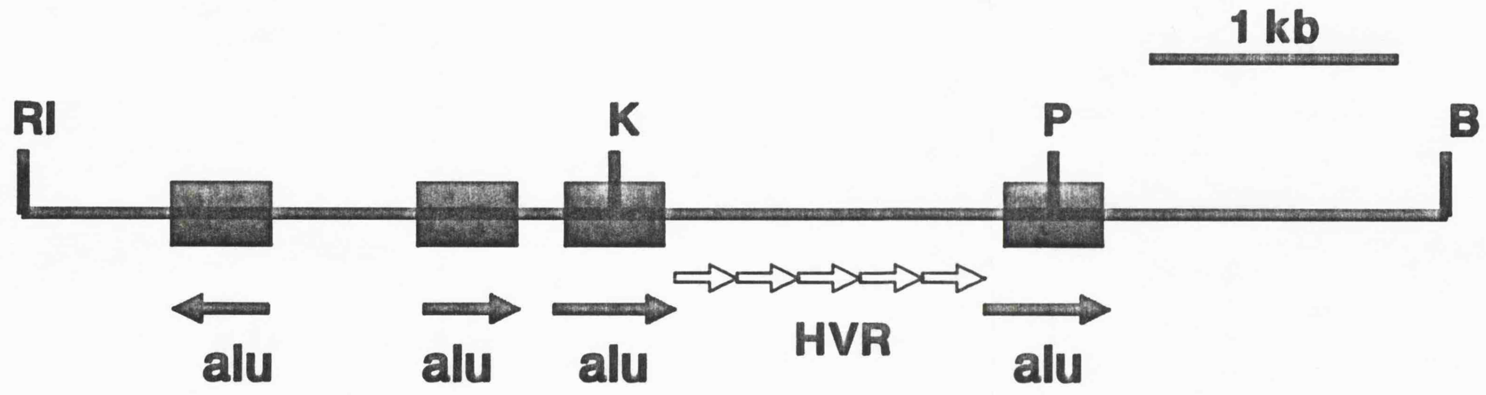
4350 ccatacaaca gggcaaatca gaacagaaat cagtgacgca cttaaataca

```
l00   aaagatgaaa ttctcttcca ccagcacggc agagcacgta gattcactgg
l50   ggccaagatg ctgattaatc agtgaaaaaa aagataacca acaaatcata
ı00   aaactaaaaa agtagtcaca tacctctctc aacatactgt tctctttttc
ı50   cttctgctcc aaaaggcttt ctaggtgatg aacgtgcatc tctgcctctg
ı00   ccagtcgtct tgttctctca tggtcttcct cggtagcctt ggcagaaagt
ı50   cctttgctct gcaacatttc cagaagcttc ttatggattc atcccgagca
'00   tttagggtct gcttttgagt ctcaatacgc agctccattt cctccaatgt
'50   ctttcgaaga agaaacagct ctttggcctg ccgctcatgc tcagcatgaa
ı00   gcctctgaaa gttctcctct gtcagctctg ctacacaagg ttcgccagtc
ı50   ctgctgctac tatcctgctg aaacagctga ttcaggtccc tctggatcc   4898
```

Figure 4.9

Summary of DNA sequence features at MS608.

The diagram summarises the sequence data presented in Figure 4.8. The hypervariable region (HVR) is shown as a series of open arrows (not to scale). Alu elements are shown as filled boxes, with their orientation indicated by the arrow below. RI = *Eco*RI, K = *Kpn*I, P = *Pst*I, B = *Bam*HI.


1 kb
RI
K
P
B
alu
alu
alu
HVR
alu

truncated member of the L1 family of dispersed repeats (Singer and Skowronski,1985). Much of the DNA flanking pMS32, however, is occupied by an element with a 70% similarity to the putative LTR region of a retrovirus-like sequence element (RTVL-I) discovered in the human haptoglobin-related gene by Maeda (1985). The sequence similarity did not extend beyond the bounds of the putative LTR, suggesting that the element in pMS32 represented an isolated LTR from the same family of retroviral elements. The sequence similarity is interrupted by the tandem repeat array of pMS32, and resumes on the other side, suggesting that the tandem array may have arisen from within this element (*v.i.*, section 4.3.2.5).

4.3.2.4 A *novel dispersed repeat element in pMS1*

In addition to the partial Alu element in pMS1 (*v.s.*), there is a short region which shows significant similarity to many regions of non-coding DNA from mammalian DNA sequences in the EMBL database (analysis of Prof. Alec Jeffreys). It showed no relation to any of the major classes of dispersed repeat elements, but a similar element was described in a report by Donehower *et al.*(1989). This region is about 70-100 bases long; the apparent length varies according to the criteria upon which similarity is defined. A central 55 bases is almost invariant among the examples studied, whereas a lower but still significant sequence similarity is found extending away from this central region, suggesting that the full-sized element may be as large as 110 bases or more.

The association of these elements with coding sequence, and examples of elements which appeared to have been conserved between humans and rodents prompted the suggestion that they subserve a *cis*-acting function in gene expression (Donehower *et* *al.*,1989).

However, it is difficult to evaluate the contention that there is an association with coding sequence, as the element was defined as a sequence of frequent occurrence in the database, and gene sequences are, for the most part, what the databases consist of. Furthermore, the rodent examples studied by Donehower *et al.* were found in the database by virtue of similarity to their human consensus sequence, and selection might thereby have been introduced in favour of those rodent elements which had been better preserved than others in mammalian evolution.

Further analysis made use of *unselected* comparisons. This was done by choosing at random twenty human elements from the EMBL sequence database which matched the derived consensus sequence. The cognate position in a homologous rodent sequence, if available from the database, was then analysed to see if there was such an element and if so, how well preserved it was with respect to the human element. The database allowed six elements to be analysed in this way, in the human and rat enkephalin, fibrinogen γ-chain and cytochrome P450 genes, and three elements in the human and mouse myoglobin genes. Of these, four were no better preserved than surrounding non-coding DNA, and the remaining two elements, if present, were too poorly preserved to be identified within the rodent sequence.

These elements show no obvious features typical of retroposons; there is no polyadenylate tract associated with them, and while their exact extent is ill-defined, no examples of target site duplication have been documented. Thus while the origin and possible function of these elements remains unclear, the fact that an element appears in the DNA surrounding a minisatellite (*v.s.*) and also between an Alu element and the poly-A tail in the 3′ untranslated region of a processed GAPDH pseudogene (Hanauer and Mandel,1984), suggests that not all of these elements are associated with functional coding sequence, and that similarity of sequence between human and rodent versions may reflect haphazard preservation, rather than evolutionary conservation, of DNA sequence.

4.3.2.5 *Tandem repeats arising from dispersed repeats*

The tandem repeat sequence of pMS32 is found to disrupt the sequence similarity to the RTVL-I LTR (Maeda,1985). The similarity resumes on the other side of the tandem repeat block, and sequence alignment suggests that the tandem repeat array at pMS32 may have arisen by tandem repetition of a sequence in this LTR-like element (analysis of Prof.Alec Jeffreys). Similarly, comparison of the repeat sequence of MS608 with the consensus Alu element sequence (Bains,1986) indicates a close relation between the two (Figure 4.10). Most differences between the Alu consensus and the MS608 repeat unit can be accounted for by short reiterations of the trinucleotide AAY (RTT). The first repeat unit in the MS608 array is an atypical repeat, having an extended $(AAY)_n$ tract (Figure 4.8).

The demonstration of such a close sequence relationship with an Alu element begs the question of how such a locus could give a locus-specific hybridization pattern; presumably the high concentrations of human DNA competitor used in hybridizations (section 2.2.3;Wong *et al.*,1987), combined with the fact that tandemly-repeated probes may give good hybridization signals with a relatively poorly-matched tandemly-repeated targets, prevents the MS608 probe from hybridizing to dispersed Alu

Figure 4.10

Sequence similarity between the MS608 repeat unit and a consensus human Alu repeat element.

The figure shows a comparison between the repeat unit sequence of MS608 (numbering as in Figure 4.8) and a consensus human Alu element (Bains,1986). Gaps (-) have been introduced to optimise the match. Note that the main difference between the two sequences consists of tandem reiterations of AAC and AAT trinucleotides.

```
                         MS608 repeat sequence
 2914                                                               2941
..GCCTGTAATCTCAGCACTTTCAACAACAACAACAATAATAATAATAACGGGCC-AGGCATGGCATATT
  |||||||||| |||||||||                             |||| ||||  |||  ||
acgcctgtaatcccagcactttt-------------------------gggaggccgaggcg-ggcggatc
1                                                                   60
                         Alu consensus sequence
```

elements at significant levels.

Similar examples, in which the tandemly repeated unit of a minisatellite appears to have derived from a sequence within a dispersed repeat element, have been described elsewhere. Thus Kelly *et al.* (1990) describe a highly unstable mouse minisatellite locus composed of GGGCA repeats, which appears to have arisen from within a mouse MT dispersed repeat element. A human minisatellite has also been shown to have arisen from within a copy of a dispersed repeat ("MstII") sequence with some similarity to "O" and "THE" elements (Mermer *et al.*,1987).

4.4 *THE INTERNAL STRUCTURE OF MINISATELLITE ALLELES*

4.4.1 *Minisatellite variant repeats*

The tandemly repeated arrays of minisatellite alleles, while maintaining a clear pattern of repetition throughout, almost always show some sequence variation between repeat units. Thus the repeat array of λMS8 is composed of two mutually interspersed repeat units 29 and 30 bases long (Wong *et al.*,1987). Similarly, nearly all minisatellite blocks which have been sequenced show some variation in sequence between repeat units (Jeffreys *et al.*,1985a; Wong *et al.*,1986,1987;Nakamura *et al.*,1987a;Royle *et al.*,1988).

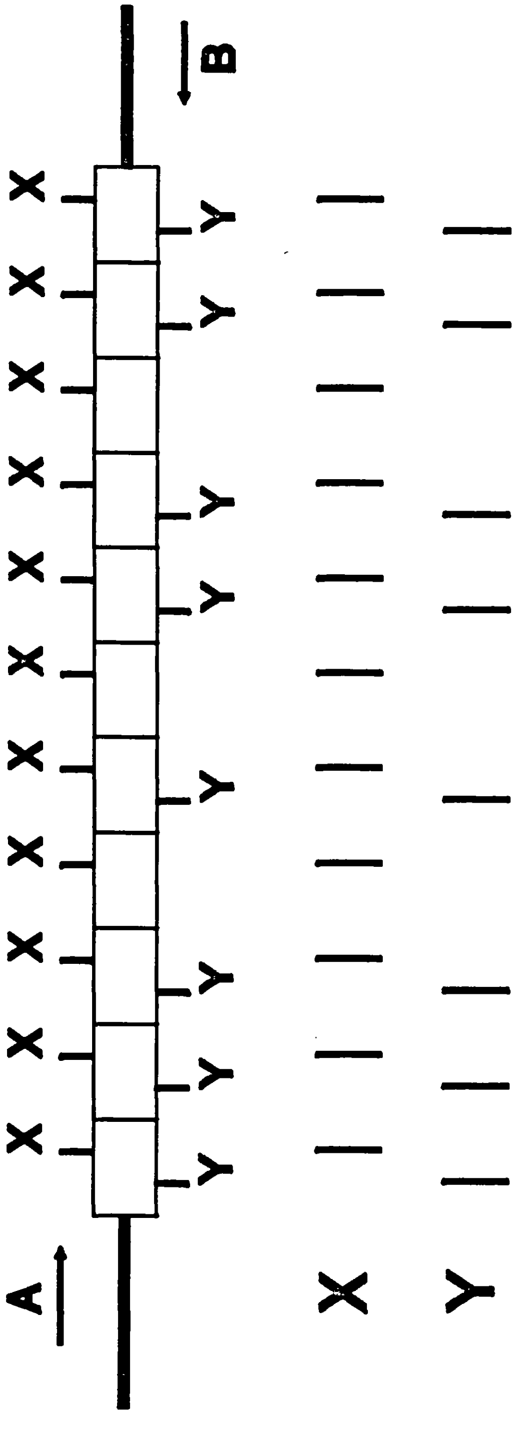
If the sequence variation between repeat units creates or destroys a recognition sequence for a restriction enzyme, then this variation can be exploited in a direct assay of the internal structure of minisatellite alleles (Figure 4.11). Minisatellite alleles are amplified by the polymerase chain reaction, end-labelled and subjected to partial restriction digestion. Thus an enzyme (X in Figure 4.11) for which a site always appears in the repeat unit will cut to give an uninterrupted ladder of DNA fragments corresponding to each repeat unit in the array. Enzymes (such as Y in Figure 4.11) which cut only some repeat units will cleave to give an incomplete ladder of partial digestion products, showing the positions within the allele of variant repeats which contain or lack sites for the enzyme Y.

This level of variation, in the internal structure of minisatellite alleles, adds a new dimension to the genetic analysis possible at minisatellite loci. In conventional analysis by Southern blot hybridization, discrimination between

Figure 4.11

Schematic summary of the general principles of minisatellite variant repeat (MVR) mapping.

The example shows a minisatellite allele in which all repeat units are cleaved by the restriction enzyme X, but only some by enzyme Y. The positions of the variant repeat units can be mapped by amplifying the allele using primers A and B. After end-labelling, the PCR product is partially digested with X or Y, the partial digestion products separated by gel electrophoresis, and those bearing a labelled end identified by autoradiography. In this case enzyme X would show an uninterrupted "ladder" of fragments, with one rung per repeat unit, whereas Y would give rise to an interrupted pattern showing the position of variant repeat units in the array.



A
B
X
Y

alleles is in practice limited by gel electrophoretic resolution, together with the fact that alleles of distinct ancestry may happen to contain identical numbers of repeat units. Mapping of minisatellite variant repeats (MVRs, Jeffreys *et al.*,1990a) not only allows the distinction to be made between alleles of identical length which are not truly isoallelic; it also makes available a new fund of variation at these loci, and allows the absolute measurement of allele repeat unit copy number (Jeffreys *et al.*,1990a).

4.4.2 *Studies at the locus D1S8*

Mapping of minisatellite variant repeats was first applied to the D1S8 locus detected by pMS32 (Jeffreys *et al.*,1990a). This locus was known to be extremely variable (Wong *et al.*,1987), to have a high mutation rate to new length alleles in both germline and soma (Armour *et al.*1989a, see also section 5.2.1.2), and to be amenable to amplification by the polymerase chain reaction (Jeffreys *et al.*,1988b). Furthermore, sequence analysis (Wong *et al.*,1987) demonstrated the presence of a variant base in the repeat unit sequence; an A to G transition created a site for *Hae*III, while a site for *Hin*fI appeared in all repeat units sequenced. Internal mapping of alleles at D1S8, using *Hin*fI to show the position of each repeat and *Hae*III to show the variant positions, demonstrated that the variant repeats accounted for about one third of all repeat units in alleles at D1S8.

The two types of repeat unit were often intermingled along an allele, although uninformative alleles, which contained long stretches of repeat units of similar type, were a frequent occurrence. Some common motifs within the internal maps emerged: firstly, while no common patterns were seen at the 3′ end, the patterns at the 5′ ends fell into three main haplotypic groups, suggesting that the variation and mutation at this locus was mainly due to instability at the 3′ end of the array.

The use of the polymerase chain reaction to amplify minisatellite alleles in the MVR mapping procedure allows the possibility of the analysis of single target molecules (Jeffreys *et al.*,1988b;1990a). This in turn makes possible the isolation and mapping of single spontaneous mutation events in bulk DNA. DNA from the germline (sperm) or soma (blood) was restriction digested and size-fractionated, and a fraction much smaller than the unmutated progenitor alleles recovered. Single deletion mutant molecules were then amplified and internally mapped; comparison with the known maps of the progenitor alleles demonstrated that there were no events which had to be ascribed to recombination between alleles, suggesting that for this class of mutants at least, the alleles were evolving primarily along haploid chromosome lineages, disregarding the presence of another allele on the other homologous chromosome. Furthermore, the breakpoints of the deletion events could be mapped, and in accordance with the conclusions of the population survey, showed a significant excess of deletion breakpoints towards the 3′ ends of alleles at D1S8.

This detailed analysis of mutations at D1S8, however, is confined to the analysis of deletion events which remove more than 20% of the repeat units from a given allele. It is known that the commonest mutations at minisatellite loci in both germline and soma are small length changes (Jeffreys *et al.*,1988a;Armour *et al.*,1989a;section 5.2.1.2), and that increases and decreases in allele length are approximately equally common. Thus while large deletions at D1S8 appear to evolve in haploid lineages, with a bias towards mutational change at one end of the array, it may be that other mechanisms of mutation predominate among the mutations that most frequently occur.

4.4.3 *MVR analysis at other loci*

4.4.3.1 *Disadvantages of D1S8*

While both providing a highly informative genetic system and a powerful tool for mutational analysis, MVR mapping at D1S8 has three significant disadvantages: firstly, in common with many hypervariable minisatellites, it has alleles of a wide range of sizes, and many alleles in the population are too large to be efficiently amplified and mapped; secondly, only two kinds of repeat units are found, limiting mapping to binary coding of alleles; thirdly, and related to the second point, "bland" alleles, apparently containing long stretches of a single type of repeat, are frequently seen. This section details the investigation of other loci at which MVR mapping may not be subject to these constraints.

4.4.3.2 *MVR mapping at D7S21*

The locus D7S21 detected by λMS31 (Wong *et al.*,1987) is a hypervariable minisatellite consisting of variable numbers of a 20 base repeat unit. The locus has a high population heterozygosity level (99%), but its relatively restricted allele size range (3.5-13kb *Hin*fI alleles were found in an initial survey; Wong *et al.*,1987) suggested that a larger fraction than at D1S8 (about 50%) of the alleles present in the population might be of a size (<5kb) amenable to MVR mapping.

Sequence variants suggest that while a site for *Mnl*I should be present in most or all repeats, a site for *Alw*NI is present in only some repeats (Figure 4.12a). It had already been established that alleles at this locus could be amplified by the polymerase chain reaction, at least to the point at which amplified products could be detected by Southern blot hybridization (Mrs.Rita Neumann, Mrs.Vicky Wilson and Prof.Alec Jeffreys, unpublished results; see also Jeffreys *et al*.1988b). Modified PCR primers were designed which included a "tail", not matching the target sequence, but which would result in the formation of a site for *Eco*RI at one end of the PCR product (see Figure 4.12 legend).

The smallest alleles at D7S21 found in a survey of *Alu*I-digested DNA from unrelated people of British stock (Mrs.Rita Neumann, Mrs.Ila Patel and Prof.Alec Jeffreys, unpublished results) were chosen for further study. Under carefully controlled PCR conditions (see Table 2.2 and Figure 4.12) it was possible to recover small amounts (10-200ng) of products corresponding to the alleles at D7S21, although with considerably more difficulty than experienced at D1S8. During this procedure it was noted that the size of the PCR product bore an inconstant relation to the size found by Southern blot analysis of DNA digested with *Alu*I, and that the relations observed appeared to fall into two classes. Restriction mapping showed that one of the *Alu*I sites flanking D7S21 was polymorphic, with a heterozygosity of about 35% (data not shown).

Results of MVR mapping of two of the alleles chosen are shown in Figure 4.12b. These show the presence of variant repeats mutually interspersed along each allele. Furthermore,

Figure 4.12

MVR mapping of alleles at D7S21.

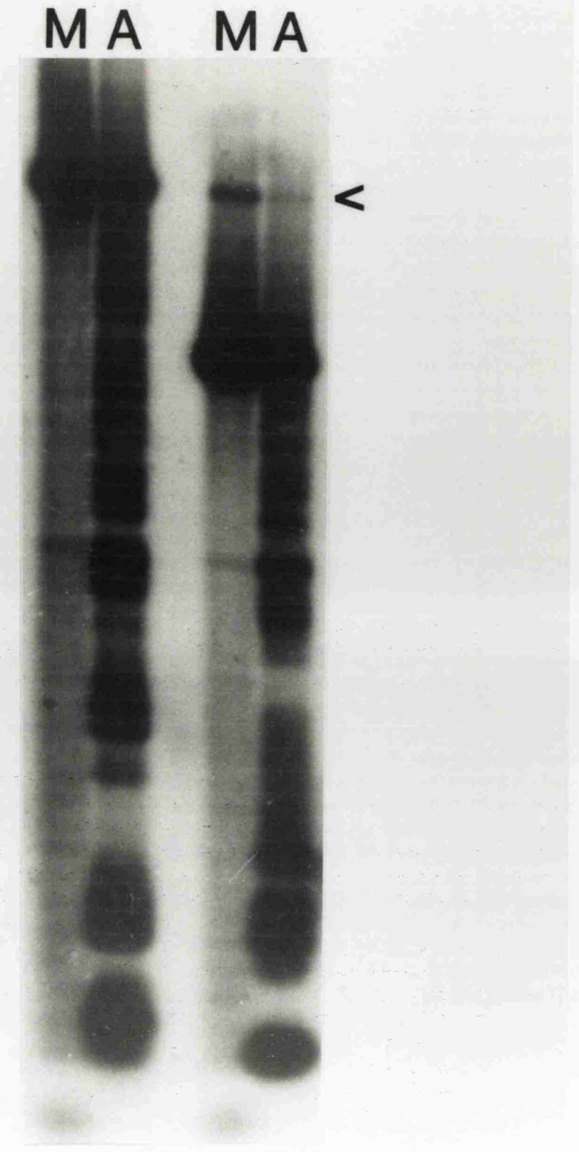
(a) shows the MS31 consensus repeat unit as determined by sequence analysis (Wong *et al.*,1987). All repeat units sequenced contained a site for *Mnl*I (GAGG); a site for *Alw*NI (CAGNNNCTG) is only present if the central pyrimidine (Y) is a cytosine.

(b) MVR mapping of two very short alleles at D7S21. Alleles at this locus were amplified from 500ng genomic DNA using the primers 31AE (5'**CCTAGGATCCGAATTC**TTTGCACGCTGGACGGTGGCG3') and 31B (5'CCCACACGCCCATCCGGCCGGCAG3') for 30 cycles (Table 2.2). 31AE has a 5' extension (bold) which does not match the target sequence; this allows the artificial creation of a site for *Eco*RI (underlined) during amplification. 200ng of genomic DNA was amplified for 26 cycles (Table 2.2). PCR products were digested with *Eco*RI and this end labelled by filling-in with [α^{32}P]dATP and AMV reverse transcriptase. End-labelled substrates were partially digested and separated on a 2% agarose gel. Note the apparent uniformity of intensity at adjacent *Mnl*I (M) sites, but the heterogeneity between *Alw*NI (A) sites. The DNA amplified from the smaller allele has been contaminated with some DNA from the larger allele (arrowhead).

(a)

```
             AlwNI
           cagnnnctg
           |||||||||
TGGGAGGTGGRYAGTGTCTG
   ||||
   gagg
   MnlI
```

(b)



some of the repeats containing a site for *Alw*NI appear to be cut more readily than others; since *Alw*NI has a recognition sequence which is an interrupted palindrome (CAGNNNCTG), it may be that the nature of the interrupting bases (NNN) has an effect on the efficiency of cleavage. In addition, less frequent variant repeats containing a recognition sequence for *Hae*III have been detected and mapped (data not shown).

Although these results suggested the possibility of highly informative MVR variation at D7S21, the practical difficulty posed by amplification of DNA from alleles at this locus precluded its use in further work. Even the very small alleles shown in Figure 4.12b required very careful tuning of PCR conditions to give products visible as ethidium stained bands on an agarose gel. At about the point at which the alleles became visible (1-10ng), even a few further PCR cycles resulted in the degeneration of the amplified DNA into a heterogeneous smear of products, presumably due to mispriming by high concentrations of amplified tandem repeats (Jeffreys *et al.*,1988b). Thus while it is possible to produce amounts of alleles at D7S21 directly visible after ethidium staining, it appears that the "window" of PCR conditions to produce these amounts is narrow, and if amplification is pushed even a little further the products collapse into a smear. Further work at this locus by Mrs.Rita Neumann and Prof.Alec Jeffreys has confirmed the difficulty of preparing substrates for MVR mapping at this locus.

The repeat unit at D7S21 is considerably shorter (20 bases) than at D1S8 (29 bases); this may account for the observed difficulty in the point at which PCR cycling ceases to result in amplification of fully-sized alleles. If, as suggested

(Jeffreys *et al.*,1988b), the "collapse" of minisatellite PCRs is due to out-of-register priming by partially extended products, then one would expect significant effects at a lower overall concentration of products if the repeat unit is short. The ease with which large amounts of alleles at D22S163, which has a 90 base repeat unit (*v.i.*), can be produced suggests that the repeat unit length has a strong influence on the practical ease with which large amounts of minisatellite alleles can be produced by PCR.

4.4.3.3 *MVR mapping at D22S163*

cMS607 detects a highly variable locus, D22S163 (Table 3.2). This locus was chosen for further study as it appeared to be highly variable (heterozygosity 90%) and unstable (two mutations were seen in 100 meioses scored). Furthermore, allele size at this locus appeared to be restricted to a range (2-5.9kb *Mbo*I alleles) which should allow MVR mapping of most alleles in the population. Sequence analysis suggested that a number of restriction enzymes would be found to cleave some repeat units but not others (Figure 4.14a).

Further genomic analysis showed that two length variable regions were present in the *Mbo*I fragments analysed by Southern hybridization with MS607. An example is shown in Figure 4.13(a), in which a father/mother/child trio is analysed using *Mbo*I alone or in combination with *Sst*I. The child has a new mutant allele which is seen after digestion with *Mbo*I alone, but not in combination with *Sst*I. Furthermore, the sizes of alleles seen after digestion with *Mbo*I bear an inconstant relation to the sizes seen with *Mbo*I plus *Sst*I, suggesting the presence of a second length-variable region. Subcloning allowed the main repeat block in the clone ("607A") to be analysed in

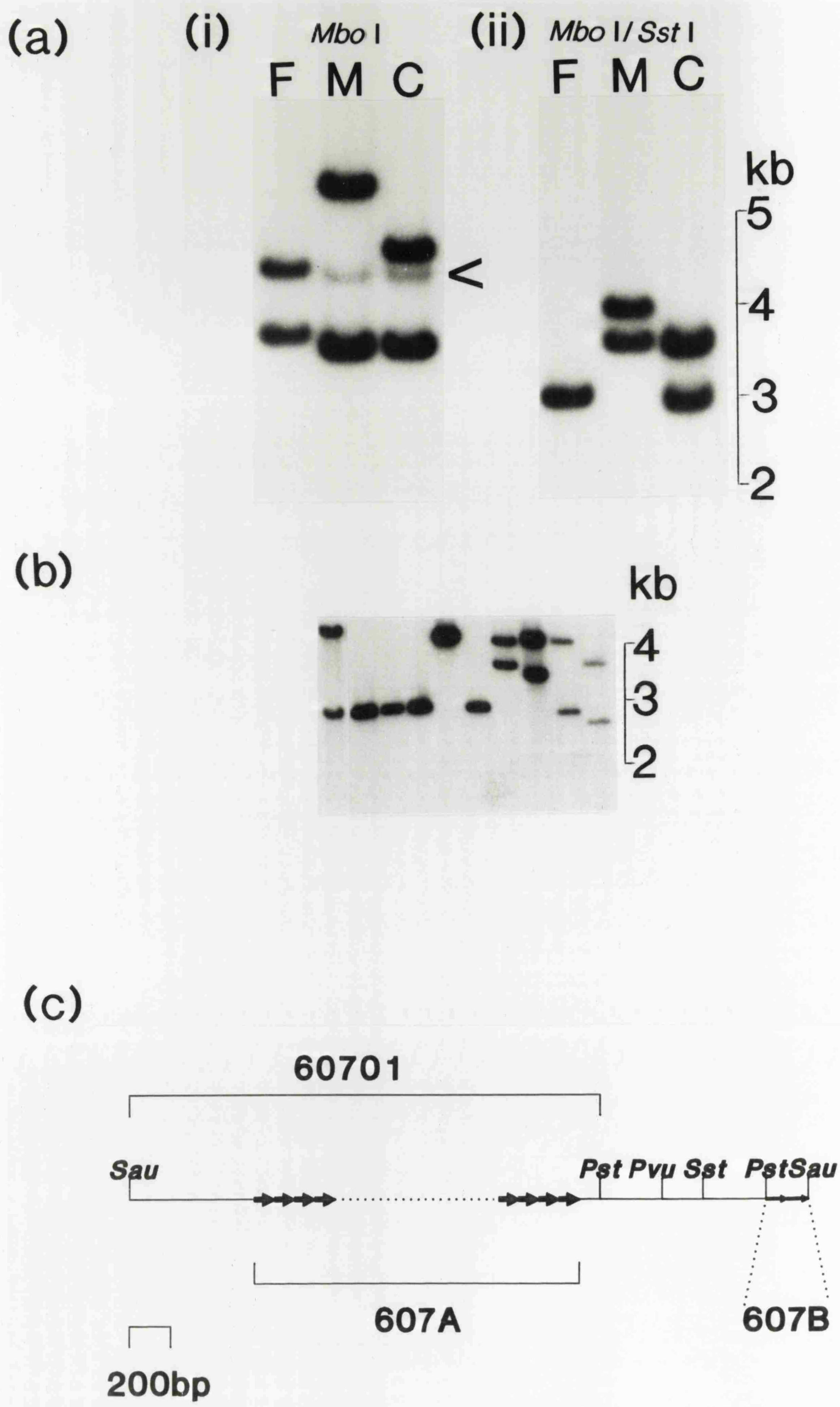
Figure 4.13

Contributions of minisatellites 607A and 607B to variation and mutation at D22S163.

(a) DNA from a father (F, CEPH individual 136201)/ mother (M, CEPH 136202)/ child (C, CEPH 136206) trio from the CEPH panel. A mutation of paternal origin is observed in the child when the DNA is digested with *(i) Mbo*I alone; the mutation is not observed when the DNA is cut with (ii) *Mbo*I plus *Sst*I. Note also that when *Sst*I is added, the fragments detected bear an inconstant relation to those seen with *Mbo*I alone. For example, the father is a heterozygote with *Mbo*I alone but an apparent homozygote with *Mbo*I and *Sst*I. The origin of the faint bands (arrowhead) in mother and child ~~are~~ is unknown.

(b) DNA from 10 unrelated individuals digested with *Pvu*II and probed with a *Sau*3AI-*Pst*I subclone ("60701" - see (c)) containing only the 607A region. As shown in (c), *Pvu*II would be expected to separate the two minisatellite arrays, and thus allow 607A to be studied in isolation. The fragments detected are variable, but the variation seen is clearly not sufficient to account for the overall heterozygosity of 90% observed with *Mbo*I fragments.

(c) Summary of subcloning, restriction mapping and sequencing data at D22S163. The 607A minisatellite is variable, but the extremely high levels of variation, and the mutations observed, appear to be attributable to the second minisatellite 607B. Sites for the enzymes *Sau*3AI, *Pst*I, *Pvu*II and *Sst*I have been abbreviated to their first three letters.


(a)
(i)
Mbo I
F
M
C
<
(ii)
Mbo I/ Sst I
F
M
C
kb
5
4
3
2
(b)
kb
4
3
2
(c)
60701
Sau
Pst
Pvu
Sst
Pst
Sau
607A
607B
200bp

isolation (Figure 4.13b). This showed that this repeat block was of limited variability (heterozygosity 50% in 16 unrelated individuals studied), and that the high level of overall variability in the population, as well as the new mutations observed, were in fact due to a smaller repeat array ("607B") which extended to one end of the cloned *Sau*3AI insert. A summary of sequence data from this clone, which confirmed the relative positions of the tandem arrays and restriction sites, is shown in Figure 4.13(c).

Although 607A was thereby shown not to be the source of the observed mutations, the presence of a limited number of alleles in the population suggested that it might be possible to analyse most or all alleles in the population by MVR analysis. Primers were designed to allow amplification of the 607A repeat array, but the primer at one end was deliberately located some 200bp from the repeat block. This was to allow MVR mapping by *indirect* end-labelling. In this procedure amplified alleles would be subjected to partial digestion, and those molecules containing one end in common identified by Southern blot hybridization using a probe from between the recessed primer and the tandem repeat array (Figure 4.14b). The practical advantages of this method are: (a) since the end-probe can be labelled to high specific activity (by oligo-labelling), the method is much more sensitive than end-labelling, in which a maximum of two ^{32}P atoms can be incorporated per target molecule; (b) the extensive handling of radioactive samples necessary during end-labelling, purification of the end-labelled fragment and partial digestion is reduced to labelling the end-probe and setting up a hybridization. The method, however, is not universally applicable, since it

Figure 4.14

Principles of MVR mapping at D22S163.

(a) shows the repeat unit sequences determined on sequence analysis of cMS607, a consensus (bold) derived from them, and examples of enzymes which would be predicted to cleave some repeat units but not all.

(b) shows the principle of the indirect end-labelling method for MVR analysis. Alleles are amplified using primers 607A and 607B, subjected to partial digestion and separated by gel electrophoresis. Partial digestion products extending to the right hand end of the PCR product are identified by Southern blot hybridization using a subcloned fragment (indicated as "60714") between *Pst*I (P) and *Acc*I (A) sites as an "end-probe". This probe contains no sites for *Sph*I or *Bst*EII, but there are sites for *Mae*II and *Apa*LI. However, the site for *Apa*LI is close to the end of the probe, and in practice the presence of these sites does not result in the appearance of additional bands on MVR mapping.

(a)

```
                     ...TGCACCAGCGAGTGTGCATATACTTGCCTCTTCTGGGGGTGTATGTGTGT  GTGGGCAC CAAGTG
GACCCTGTACGGTGATTGCATGTGTGCACC   GAGTGTGGATATACTTGCCTGTTCTGGGGGTGTACGTGTGTTTGTGTGCACACAGGT
GACCCTGTACAGTGATTGCATGCGTGCACCAGGGAGTGTGGATATACTTGCCTGTTCTGGGGTTGTACATGT...
              ...GCATGCGTGCACCAGGGAGTGTGGAAATACTTGCCTGTTCTGGGGGTGTACACGTGT  GTGTGCACACATAT
GACCCTGTACAGTGATTG...

                        ApaLI                                       MaeII
                       gtgcac                                       acgt               T
                       ||||||             T          C        G     ||||       T       A
GACCCTGTACRGTGATTGCATGYGTGCACCAGGGAGTGTGGA ATACTTGCCT TTCTGGGG TGTACRYGTGTTTGTGGCACACAGRTgaccct
                 ||||||                   A          G        T   ||||       ||||||   |||||||
                 gcatgc                                           acgt       gtgcac   ggtnacc
                  SphI                                            MaeII      ApaLI    BstEII
```

(b)

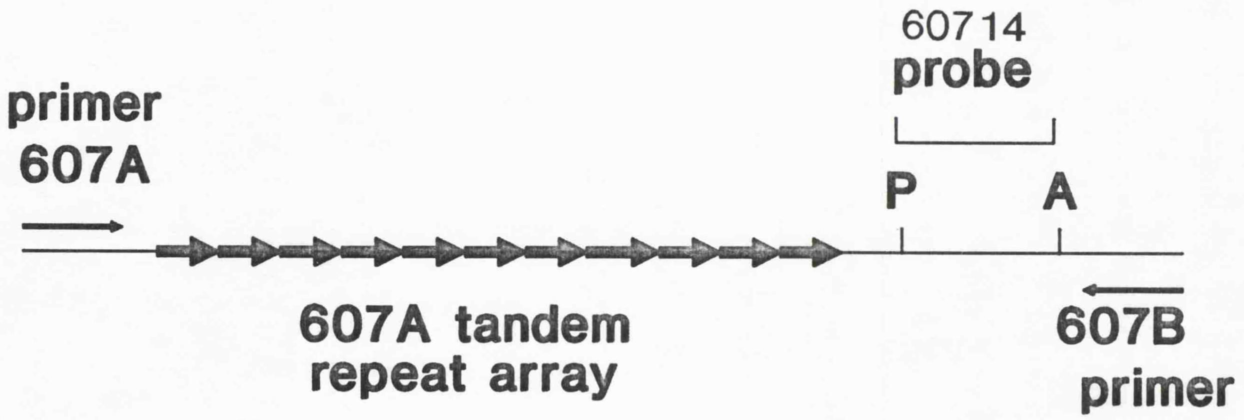
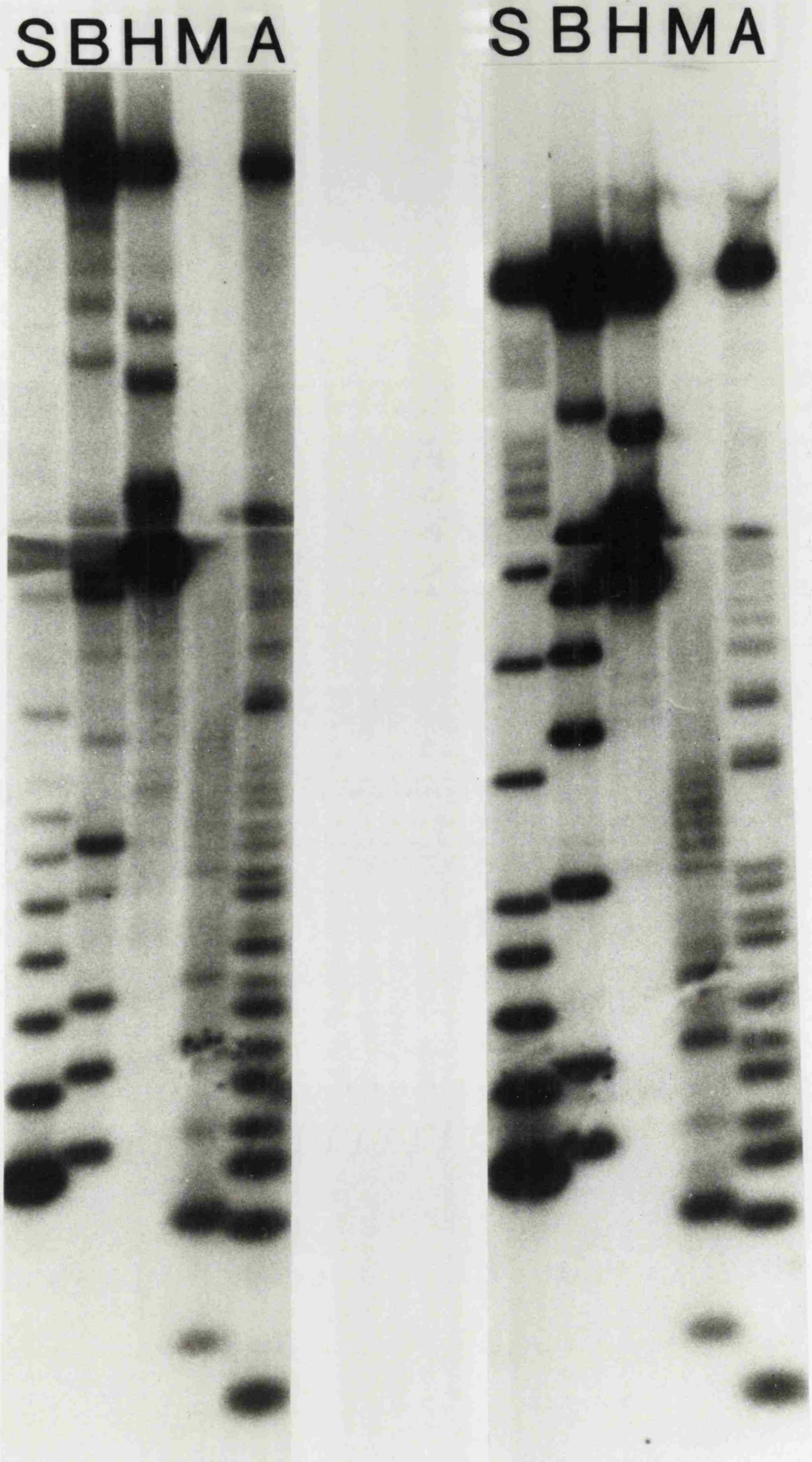


Figure 4.15

Initial work on MVR mapping at D22S163.

The figure shows MVR maps of the two alleles at D22S163 from CEPH individual 136202, using *Sph*I (S), *Bst*EII (B), *Hin*fI (H), *Mae*II (M) and *Apa*LI (A). The methods used to map these alleles by indirect end-labelling are outlined in section 4.4.3.3, and illustrated in Figure 4.14(b). Alleles were amplified from 400ng genomic DNA for 31 cycles (Table 2.2) in a total volume of 20μl (for some subsequent work involving larger alleles a total volume of 50μl was used). PCR products were partially digested and Southern blot hybridized using probe 60714 (Figure 4.14).


S B H M A
S B H M A

assumes that the locus in question has alleles so small that the luxury of intentionally lengthening the PCR product required can be tolerated without too great a reduction in yield.

Alleles at D22S163 were amplified, gel purified and partially digested. After size separation by gel electrophoresis, partial digestion products were blotted onto a nylon membrane and hybridized with the end-probe "60714" (Figure 4.14b) labelled by random primer labelling (Feinberg and Vogelstein, 1984). Examples of MVR mapping at 607A, using a range of restriction enzymes, of two alleles from the same person are shown in Figure 4.15. It is clear that while there is extensive MVR variation at D22S163, some enzymes (for example *Apa*LI and *Mae*II) give highly complex patterns, while others (like *Sph*I and *Bst*EII) give rather simpler profiles. The former class of enzyme appears to have two potential sites per repeat unit, while the latter class has only one. Thus the choice of enzyme could be used to determine the degree of detail required of the MVR map; *Sph*I and *Bst*EII could be used to determine the pattern on a broad scale, and *Mae*II and *Apa*LI brought into use if finer detail were required.

In fact, *Sph*I and *Bst*EII alone were sufficient to give detailed maps of the alleles at the 607A minisatellite. The results of mapping the six alleles seen in a survey of 16 unrelated people are given in section 5.3.2.

4.4.4 *Summary and prospects*

Internal mapping of minisatellite alleles by MVR analysis makes available a powerful tool for the investigation of many unresolved issues concerning the mutation and evolution of minisatellites (Jeffreys *et al.*,1990a). While the pioneering work done at D1S8 has broken much new ground, the system is applicable to many other loci. The chief limitations on its use are that sufficient variation should occur between different repeats, that this variation should be identifiable using commercially available restriction enzymes, and that it should be possible to amplify large amounts (>5ng) of minisatellite alleles.

This section has investigated the application of MVR mapping to two other loci, D7S21 and D22S163. The former, although rich in both length and internal map variation, poses too great a practical difficulty during the amplification stage to be a useful system. D22S163, however, is amplifiable without difficulty, has alleles short enough to map by indirect end-labelling, and contains a highly informative fund of MVR variation. However, the major component of variation at this locus was due to variation at a second minisatellite, 607B, and the information of biological interest obtainable from this locus was limited to a survey of the relatively small number of different alleles in the population.

The broad applications of the system at other loci would depend on the population structure of the loci. Thus, for example, by MVR analysis at a moderately variable locus, which had fewer than 20 distinct alleles in the population, as exemplified by D22S163, it might be possible to map all or most of the alleles in the population (if short enough) relatively quickly. Thus the relationships between alleles at the more slowly evolving loci may be investigated. At the most variable loci, where mapping all the loci would probably be precluded not only by practical considerations but also by allele size distribution, population surveys of allele structure could be supplemented by structural analysis of spontaneous mutations, not only by size-selection of rare events from bulk DNA (Jeffreys *et al.*,1990a) but also by examining spontaneous events clonally amplified *in vivo*, either germline events clonally amplified from zygotes into children (Jeffreys *et al.*1988a) or somatic events, amplified by the clonal expansion of tumour cell populations (Armour *et al.*,1989a; see section 5.2.1.1).

CHAPTER 5

MUTATION AND EVOLUTIONARY CHANGE AT MINISATELLITE LOCI

Cats and monkeys, monkeys and cats - all human life is there.

Henry James

Summary

Somatic mutation at human minisatellite loci has been investigated using gastrointestinal tumours as a model system. Somatic mutations within these clonal cell populations were detected at the most variable minisatellite loci, and many of the properties of somatic mutations, including the high frequency of small length changes and symmetry of direction of length changes, closely parallel those of germline mutations. Minisatellite probes were also used to survey for somatic change in human breast cancer. Both DNA fingerprints and locus-specific minisatellites show that genetic change at minisatellite loci is relatively uncommon in these tumours. The most frequently detected change was deletion of an allele from the D17S134 locus on the short arm of chromosome 17. Other, less frequent molecular lesions were also investigated, including what proved to be a bizarre transposition of a minisatellite segment. Initial studies on some aspects of minisatellite evolution were carried out, including analysis of the alleles present in modern human populations at two loci, as well as the analysis of allele size distribution between humans and chimpanzees by PCR at D17S134.

5.1 INTRODUCTION

5.1.1 *Hypervariability and germline mutation*

A genetic system at which alleles were selectively neutral but with high levels of population variability would be predicted to maintain that variation by high rates of mutation to new length alleles in the germline. Experimentally, a high rate of germline mutation has been measured directly for some human minisatellite loci by pedigree analysis (Jeffreys *et al.*,1988a). The loci most variable in the population also have the highest rate of mutation, and the relation between heterozygosity and mutation rate closely follows that predicted by the neutral mutation-random drift hypothesis (Kimura,1983). No evidence of mosaicism for germline mutants was detected, and the rates of production of mutations in the male and female germlines were not significantly different. Since the stem cells of the male and female germlines have very different pre-meiotic cellular histories (*v.i.*, section 5.2.1.3), the equivalence of mutation rates suggested that the mutation process might be restricted to one stage of gametogenesis, possibly meiosis itself.

The resemblance of the minisatellite "core" sequence to the *chi* recombination signal of *E.coli* prompted the initial suggestion that new length alleles may arise by unequal exchange between chromosomes (Jeffreys *et al.*,1985a). However, studies of sequence polymorphisms in DNA flanking a new mutant allele at the minisatellite YNZ22 (Wolff *et al.*,1988) and of the linkage phase of genetic markers flanking D1S7 (Wolff *et al.*,1989) have shown that interallelic recombination does not necessarily, and probably does not commonly, accompany the generation of new length alleles by mutation. The direct analysis of new mutant molecules at D1S8 by amplification from bulk germline and somatic DNA (Jeffreys *et al.*,1990a) confirms that interallelic exchange only rarely accompanies the appearance of deletion mutants at D1S8.

Among the other candidate mechanisms for mutation at minisatellites, intramolecular recombination, which can only shorten alleles, is ruled out as a common mechanism by the equally frequent occurrence of mutations which increase or decrease allele length (Jeffreys *et al.*,1988a). Both unequal sister chromatid exchange and polymerase slippage at replication are feasible as mechanisms for the origin of new length alleles without exchange of flanking markers.

5.1.2 *Somatic mutation*

Somatic tissues, which do not undergo meiosis, might be predicted to show a different pattern of mutational change from the germline, where some mutational events may be occurring late in gametogenesis. Although meiotic recombination does not occur in the soma, mitotic recombination may still accompany mutation events, and replication slippage and unequal sister chromatid exchange may generate new mutants somatically. In addition to shedding light on the mutation process at minisatellites, somatic mutations may have practical utility as markers for the cellular lineages bearing them. In section 5.2, somatic mutation at human minisatellite loci is investigated in gastrointestinal and breast cancers, together with preliminary experiments on the detection of somatic mutation in normal tissues.

5.1.3 *Evolutionary change*

While individual germline mutations are the quanta of evolutionary change, their summation or diffusion into observed inter- and intra-species variation requires a different level of analysis. Rather than comparing parents and offspring in pedigree analysis, states can be compared within and between species, and the mutational events separating them inferred. Preliminary work on the evolution of some minisatellite loci is presented in section 5.3, and includes comparisons both within and between species.

5.2 MUTATIONAL ANALYSIS AT MINISATELLITE LOCI

5.2.1 *Somatic mutation*

5.2.1.1 *Introduction: use of clonal cell populations*

At the most variable minisatellite loci, the relatively high frequency of mutation would result in germline and somatic DNA containing a minor population of mutant molecules. The problem experimentally is that each individual mutant is present at a very low level, and all the mutant molecules are in turn vastly outnumbered by non-mutant molecules. In the germline, each gamete will contain a single allele at each locus, and will contribute that allele to the zygote. Thus the screening of offspring for mutation at minisatellite loci (Jeffreys *et al.*,1988a) is in effect a survey of the gametes which gave rise to the corresponding zygotes.

Similarly, one would predict that somatic DNA would be mosaic for normal alleles and a wide variety of somatic mutant alleles, each present at very low frequency. Just as in bulk germline DNA, the mutant molecules would be present at levels too low to detect by direct Southern blot hybridization.

The next two sections detail the investigation of somatic changes at human minisatellite loci by the analysis of tumour DNA. In those malignancies known to arise from a single cell, such as colorectal adenocarcinomas (Fearon *et al.*,1987), tumours can be used to provide a source of DNA clonally amplified from a single precursor cell. Just as the formation of a zygote, and subsequently of a child, clonally amplifies the DNA contributed by each gamete, so the malignant expansion of the tumour clonally amplifies the DNA present in the cell which originally underwent transformation to give rise to the tumour. Comparison of DNA from the tumour with normal polyclonal tissues, such as peripheral blood, acts as a screen for somatic mutation.

Differences observed in the tumour, however, may have arisen in a number of ways. Any mutation which had already occurred before the clonal expansion of the tumour would be predicted to be present in all the tumour cells. A change occurring during the malignant expansion of the tumour would be present at a level which would depend on the timing of the mutation relative to the expansion. Furthermore, if a subpopulation of cells in the tumour, in which a mutation had occurred, were to gain a growth advantage, mutation-bearing cells would come to predominate in the final tumour despite a late common cellular ancestor. Changes occurring in such "takeover" cell populations would thus appear in most cells in the final tumour, and so give a similar appearance to mutations which were already present in the normal epithelial cell prior to malignant change. For this reason the incidence of mutations observed in tumours provides an upper estimate of the incidence in cells of the corresponding normal epithelium.

Compounding these difficulties of interpretation is the consideration that a tumour is not composed solely of malignant cells. The lump of tissue from which DNA is extracted will include contributions from the various non-malignant cells present, including blood and blood vessels, fibrous stromal tissue and inflammatory cells.

5.2.1.2 *Somatic mutation in gastrointestinal tumours*

The high rate of appearance of new mutant bands in DNA fingerprints in pedigree analysis (Jeffreys *et al.*,1985a,b) prompted the screening of a number of different human tumours for somatic change (Thein *et al.*,1987). DNA fingerprinting probes were used to compare DNA from a tumour with normal DNA from the same individual; the use of multi-locus probes allowed the simultaneous screening of many hypervariable loci in the genome, but consequently did not specify the genomic location of the changes seen. A high incidence of somatic changes was detected, and included both band losses and the appearance of new mutant bands. The rate of appearance of new mutant bands was particularly high in adenocarcinomas of gastrointestinal origin. This class of tumour was therefore chosen for the study of somatic mutation at individual minisatellite loci.

This work was carried out in collaboration with Dr. Swee Lay Thein (Nuffield Department of Clinical Medicine, University of Oxford) and Dr. Martin Fey (Department of Medicine, Inselspital, Bern, Switzerland). They collected the clinical samples, extracted the DNA and prepared the Southern blot filters. DNA was extracted from each sample, digested with *Hin*fI or *Alu*I and gel electrophoresis performed under the same conditions as those used for the study of germline mutation at minisatellite loci (Jeffreys *et al.*,1988a), so that the electrophoretic resolution of mutant alleles would be directly comparable between the studies. From most patients, samples from the tumour, blood and from normal mucosa from the affected tissue were used. These were loaded onto the gels in the order [blood - tumour - normal mucosa], so that the tumour DNA was flanked by two representatives of normal tissue DNA from the same individual. In some cases, only two samples were available, from the tumour and either blood or normal mucosa.

51 patients were studied; 39 had colorectal adenocarcinomas, 11 had gastric carcinomas and one had a transitional cell

bladder tumour. The filters bearing DNA from the corresponding clinical samples were probed using the locus-specific minisatellite probes pλg3 (Wong *et al.*,1986), λMS1, λMS8, λMS31, λMS32, and λMS43 (Wong *et al.*,1987). All these probes except λMS32 detected hybridizing fragments of similar size with either *Hin*fI or *Alu*I. Any *bona fide* change in the length of a tandemly repeated allele should result in a novel fragment which should be detectable using either of these enzymes. The use of duplicate samples, one cleaved with *Hin*fI and one with *Alu*I, therefore acts as a check for putative mutations (for example, see Figure 5.1). Since the repeat unit of D1S8 detected by λMS32 contains a site for *Hin*fI, this locus could only be assessed using the *Alu*I filters.

In all, 18 mutant alleles were observed among the 102 alleles screened in this way. Ten mutants were seen in colorectal cancers, and eight in gastric carcinomas. Examples are shown in Figures 5.1 and 5.2. The appearance of novel hybridizing fragments in tumour DNA is not simply due to sample contamination; the filters were probed with six highly informative minisatellite probes, and any contaminating DNA would be observed with most or all of them. In fact, novel fragments were detected by one or, in one patient, two probes. As a further control, the novel band was seen in both of pairs of duplicate *Hin*fI/*Alu*I samples, with the novel fragment differing in length from the germline alleles by the same amount whichever enzyme was used (Figure 5.1a,b,c).

Nine mutations were detected by λMS1, six by λMS32 and one each by pλg3, λMS8 and λMS43. In many instances the appearance of a new hybridizing fragment was accompanied by a relative diminution in the intensity of one of the germline alleles, or

Figure 5.1

Examples of minisatellite mutations (arrowed) detected in gastrointestinal carcinomas. DNA from blood (B), tumour (T) and normal mucosa (N) from the same patient were digested with *Alu*I or *Hin*fI and Southern blot hybridized using minisatellite probes. (a) and (b) show DNA from patients 26 and 29 respectively probed with λMS1. In the tumour DNA from both these patients a new hybridizing fragment is seen, which differs from the unmutated alleles by the same amount after digestion with *Alu*I or *Hin*fI. In patient 26 (a), the lower germline allele appears to have been completely replaced by the new mutant allele, whereas in patient 29 (b), the hybridization intensities of the new mutant and lower unmutated alleles appear to be about equal in the tumour DNA. (c) shows DNA from patient 110 probed with λMS43. The size change to the mutant allele is the same with *Hin*fI or *Alu*I, and the intensity of the upper unmutated allele is diminished in the tumour DNA, suggesting that this allele is the precursor of the mutant. (d) shows DNA from patient 71 (no blood DNA sample) probed with pλg3, again illustrating the diminution in intensity of one allele, in this case the lower, accompanying the appearance of the mutant allele in the tumour sample.

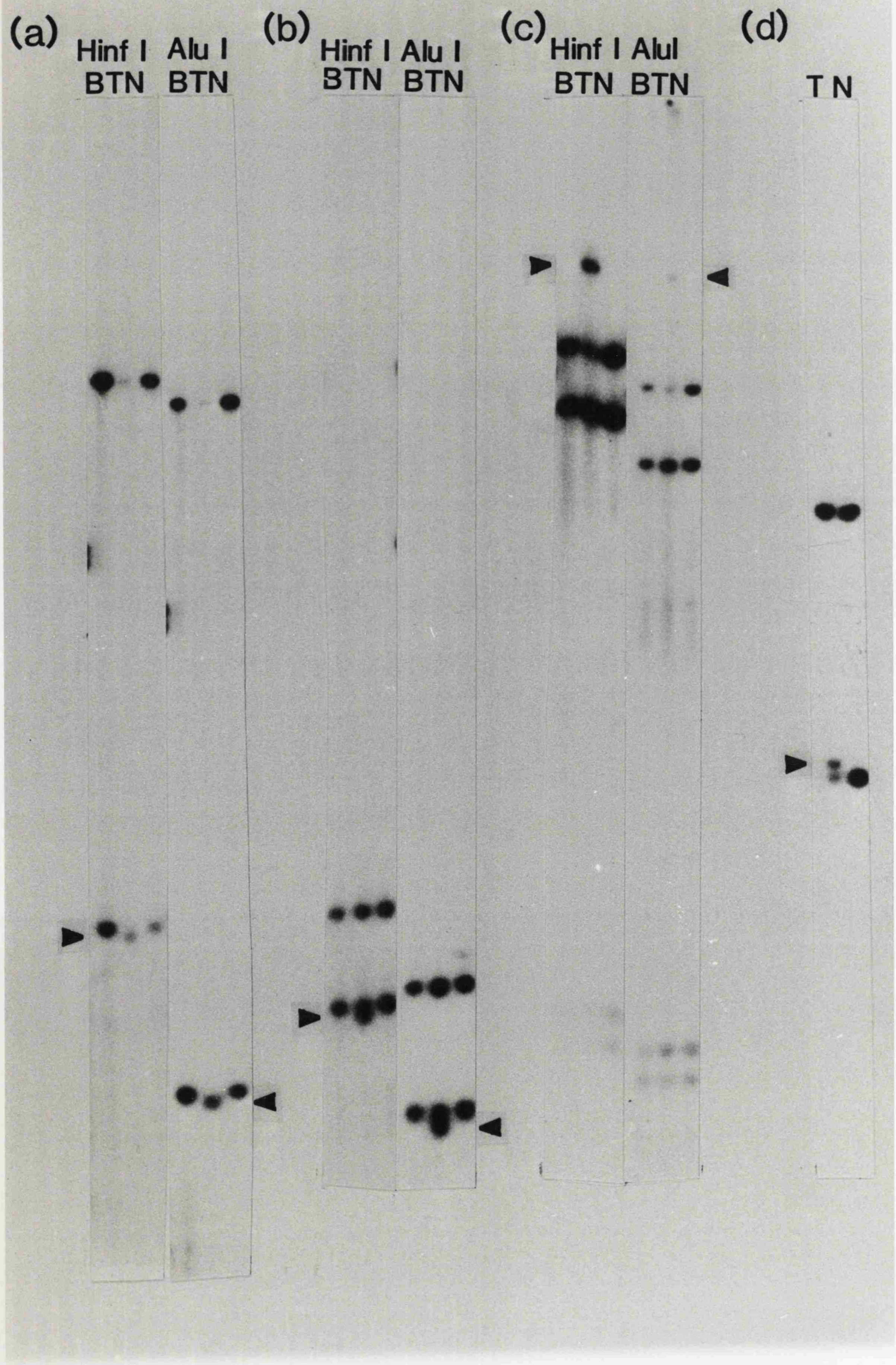
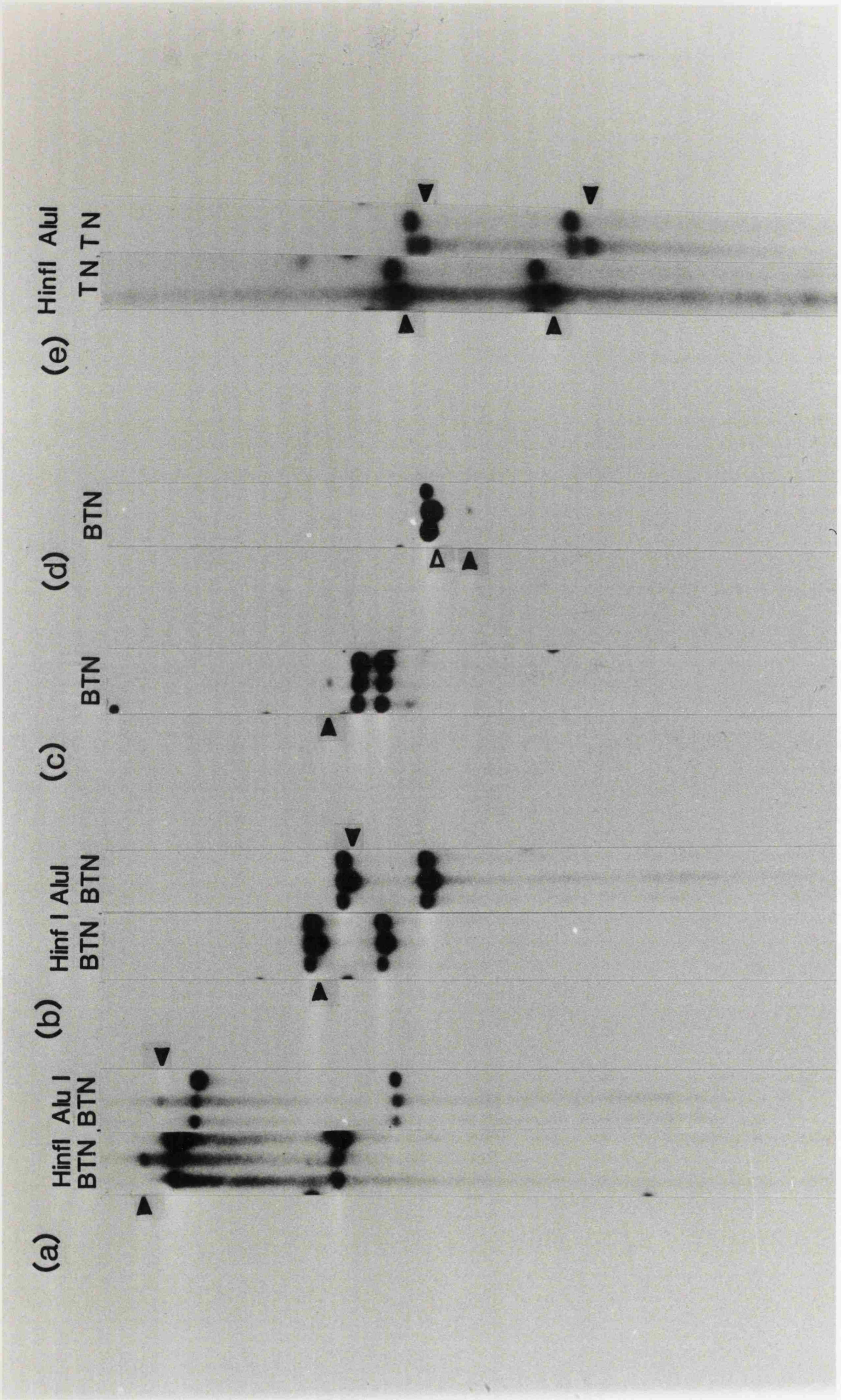

(a)
Hinf I
BTN
Alu I
BTN
(b)
Hinf I
BTN
Alu I
BTN
(c)
Hinf I
BTN
AluI
BTN
(d)
T N

Figure 5.2

Further examples of minisatellite mutations in gastrointestinal tumours. In (a) and (b) DNA from patients 101 and 107 respectively is probed with λMS1, and the appearance of the mutant alleles is not accompanied by any significant diminution in the hybridization intensity of either germline allele. In the analysis presented in Table 5.1, these mutations are assumed to derive from the germline allele closer in size. (c) and (d) show DNA from patients 218 and 420 probed with λMS32; arrowheads show mutant alleles apparently present at a fraction of the dosage of the germline alleles. The tumour from patient 420 contains a mutant present at approximately equal dosage (open arrowhead) as well as the less intense mutant. Since 420 is a homozygote at this locus, there is no indication of whether the two mutations are on different chromosomes or whether they are two mutants of the same allele. (e) shows DNA from patient 119 probed with λMS1; this tumour bears two mutations at this locus.[B=blood, T=tumour, N=normal mucosa].


(a)
HinfI
BTN
Alu I
BTN
(b)
Hinf I
BTN
AluI
BTN
(c)
BTN
(d)
BTN
(e)
HinfI
TN
AluI
TN

even its complete disappearance (Figure 5.1a,c,d). This suggests that the mutant allele arose from the now fainter allele. In other cases there was no such loss of signal from one of the alleles to indicate the progenitor allele (Figure 5.2a,b). In those cases where loss of signal could be used to identify the progenitor, it was always the allele closest in size to the mutant. For the purpose of the analysis presented in Table 5.1, therefore, if an assignment cannot be made on the grounds of loss of signal, the progenitor is assumed to be the germline allele closest in size to the mutant.

The incidence of minisatellite mutations among these gastrointestinal tumours is high, with 15 out of 50 tumours showing at least one minisatellite mutation. This contrasts with a lower incidence in breast cancers (see section 5.2.2.3); among the loci used here, only a single length change mutation has been detected (by λMS43) and a single, more unusual change (by λMS1) in 38 breast cancer patients studied ($p<0.01$).

In most mutations, the mutant allele has an intensity of about 60% or more of the presumed progenitor allele (for example, see Figure 5.1b,c,d), suggesting that cells bearing a mutant allele account for 40% or more of tumour cells (see Table 5.1 and legend). In some other cases the mutant alleles are seen at a much lower relative intensity (Figure 5.2c,d), suggesting an origin during the clonal expansion of the tumour.

One would predict that mutations already present in the cell which gave rise to the tumour, or arising early in the malignant phase, might be seen in secondary tumours derived from such primaries. If a mutant allele were detected in a primary tumour removed at surgery, its subsequent detection in, say, a biopsied lymph node would complement histological

Table 5.1

Summary of mutations at minisatellite loci detected in gastrointestinal tumours.

Knowledge of the direction and magnitude of the size changes is dependent on knowledge of the progenitor allele, which in most cases can be inferred from relative loss of signal from one of the alleles (Figures 5.1 and 5.2); where this is not possible, it is assumed that the mutant derives from the germline allele closest in size. Tumours in which double mutations were detected are marked with an asterisk. The proportion of cells within the tumour bearing the mutation was estimated from scanning densitometry of the autoradiographs, shown here in the final column as a percentage (given to one significant figure), as the intensity (m) of the mutant allele as a percentage of the sum of the mutant intensity and of any remaining progenitor (p) allele, i.e. 100 x m/(m+p). Where the mutant arose in a homozygote (§), it is assumed that only one progenitor allele has mutated to give rise to each mutant, and so the percentage given is 100 x 2m/(m+p). Since the hybridization signal of a minisatellite allele is linearly related to its length, a correction has been made in these calculations for the effect of length change on the hybridization intensity of the mutant.

Table 5.1

	Tumour type	Progenitor	Size change			% cells
Probe	(patient no.)	allele size(kb)	kb	%	N	mutant
λMS1	Gastric(26)	1.65	-0.08	-4.5	-8	100
λMS1	Gastric(29)	1.07	-0.07	-6.2	-7	40
λMS1	Colorectal(52)	11.33	-0.66	-5.8	-73	40
λMS1	Colorectal(68)	2.89	-0.08	-2.7	-9	30
λMS1	Colorectal(101)	15.20	+5.0	+32.9	+550	30
λMS1	Colorectal(107)	6.83	-0.19	-2.8	-21	50
λMS1	Gastric(116)	4.13	+1.19	+28.8	+130	40
λMS1	Gastric(119)*	9.81	-0.53	-5.4	-59	50
λMS1	Gastric(119)*	5.76	-0.25	-4.3	-28	50
λMS32	Colorectal(128)	7.04	-0.38	-5.4	-13	30
λMS32	Colorectal(134)	7.20	+0.26	+3.6	+9	20
λMS32	Colorectal(157)	7.12	+3.18	+44.6	+110	40
λMS32	Colorectal(218)	5.83	+0.58	+10.0	+20	7
λMS32	Gastric(420)*§	5.49	-0.18	-3.3	-6	100
λMS32	Gastric(420)*§	5.49	-0.48	-8.7	-16	7
pλg3	Colorectal(71)	3.66	+0.11	+3.0	+3	50
λMS8	Gastric(119)§	5.01	-0.2	-4.0	-6	50
λMS43	Colorectal(110)	10.80	+7.7	+71.0	+170	30

analysis by providing a specific link with the cellular lineage defined by the excised primary. Such a link might prove invaluable in those clinical situations where a second primary was suspected. Furthermore, this assay would not be subject to normal protein or antigen expression by the secondary, and would neither influence, nor be influenced by, the growth properties of the constituent cells. More generally, if a locus were found at which such lineage-specific mutations were to occur early in embryonic development, the mutation could be used to mark that developmental lineage. Such early embryonic somatic mutations, giving rise to adults mosaic for two cell types (mutant and non-mutant) has been described at a highly unstable minisatellite locus in mouse DNA (Kelly *et al.*,1989).

5.2.1.3 *Mutation mechanisms and rates*

In the analysis of germline mutations at minisatellite loci (Jeffreys *et al.*,1988a), most mutation events involved small (<0.5kb) length changes; in gastrointestinal tumours, the majority of length changes were small (Table 5.1). There was no apparent bias towards elongation or contraction of minisatellites in tumour mutations, with 11 decreases and 7 increases in size.

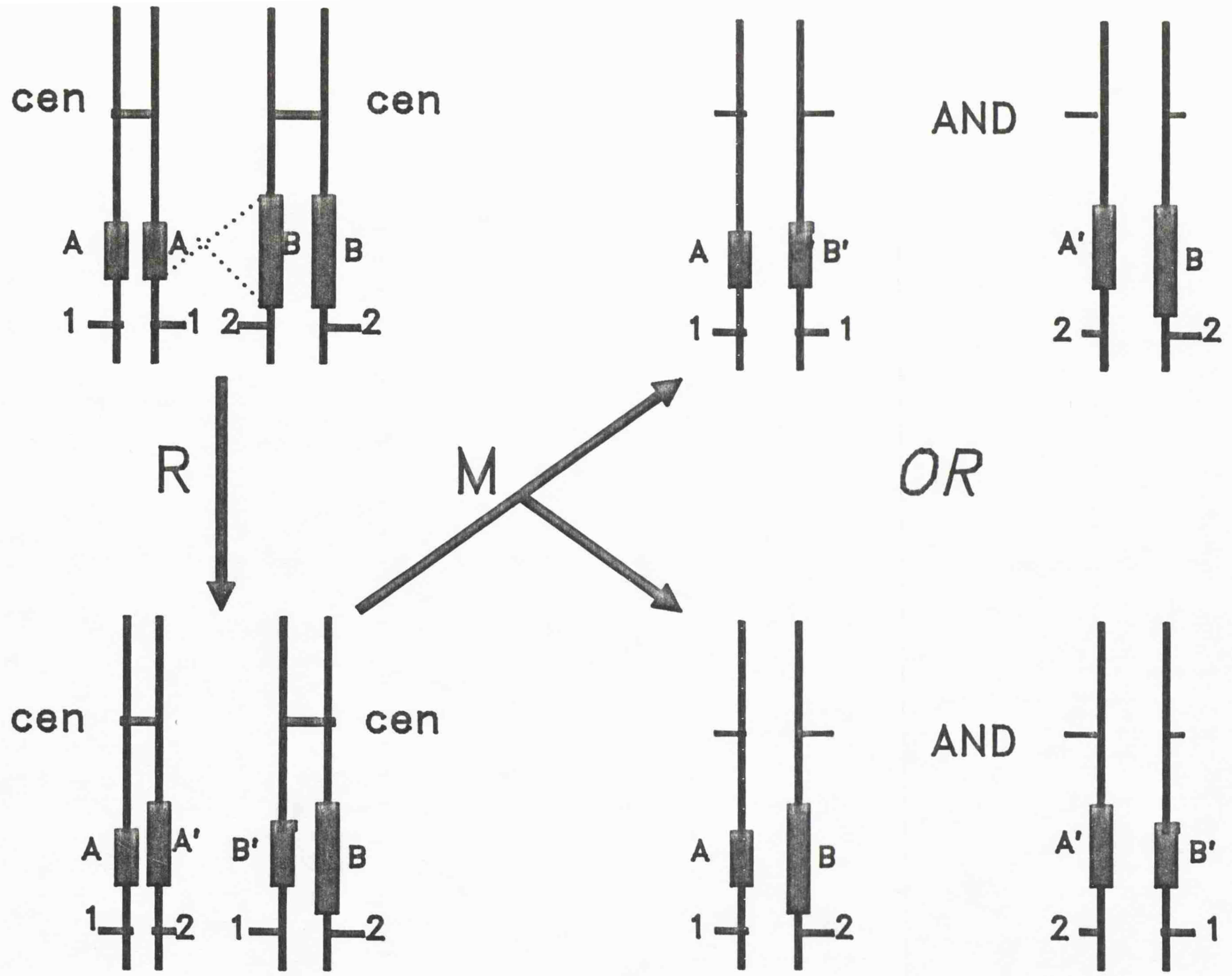
A number of mechanisms might be proposed to account for somatic mutation at minisatellite loci, including replication slippage, unequal sister chromatid exchange, unequal mitotic recombination, gene conversion and intramolecular recombination. As intramolecular recombination can only shorten alleles, it is ruled out as the only mechanism, and is unlikely to be a major contributor. The data also provide some evidence against unequal mitotic recombination as the predominant mechanism for these somatic mutations. In at least a quarter of

Figure 5.3

Predicted consequences of mutation by unequal mitotic recombination at a minisatellite locus.

The figure shows a mutation arising during mitotic recombination (R) between the alleles (A and B) at a minisatellite. After recombination (R), the products appearing in each daughter cell will depend upon the segregation at mitosis (M). Two dispositions are possible; one with each daughter cell receiving one mutant allele (A′ or B′) and one non-mutant (A or B) (above); the other (below) has one daughter cell receiving the two unmutated alleles A and B, and the other the two reciprocal mutants A′ and B′. Thus on approximately one in four occasions a mutant will be accompanied by its reciprocal product.

Note that this may be an underestimate, since those cells receiving only one mutant (above) will be homozygotized for markers (1/2) distal to the recombination, and so may be selected against *in vivo*. "cen" indicates the position of the centromere.


cen
cen
A
A
B
B
1
1
2
2
R
M
cen
cen
A
A'
B'
B
1
2
1
2
A
B'
1
1
AND
A'
B
2
2
OR
A
B
1
2
AND
A'
B'
2
1

the daughter cells resulting from such an unequal exchange, both reciprocal mutant products should partition into the cell which gives rise to the tumour (Figure 5.3 and legend). In no instance among the tumours studied was the appearance of a mutant allele accompanied by a reciprocal mutant. Where two mutations were seen (Figure 5.2d,e), they were both in the same direction.

In many respects the general properties of somatic mutations closely resemble those of germline mutations (Jeffreys *et al.*,1988a); increases and decreases in allele size seem about equally common, and small changes account for most mutations, although some very large changes do occur. Similarly, the loci most unstable in the germline are those which show most somatic mutation (Table 5.2). However, comparison of the relative frequencies of detection of mutant alleles by λMS1 and λMS32 in germline and soma suggest that the processes may not always be proportional; the germline mutation rate at D1S7 (detected by λMS1) is more than seven times the germline rate at D1S8 (detected by λMS32), whereas in the tumours studied the incidence of mutant alleles at D1S7 was only about 1.5 times that at D1S8. While some mutations in both germline and soma will be too small to resolve by gel electrophoresis, the electrophoretic resolution in the two studies was comparable, and so the relative rates measured are not due simply to differences in resolution.

It is of note that the incidences of mutant alleles in the gastrointestinal tumours are of the same order of magnitude as in gametes (Table 5.2). For example, D1S7 (detected by λMS1) has an incidence of mutant alleles of about 5% per gamete in the germline, and of about 9% per tumour allele in

Table 5.2

Summary of germline and somatic mutation data at minisatellite loci.

The heterozygosity levels and germline mutation data are taken from Jeffreys *et al.*(1988a), and (for λMS32) Armour *et al.*(1989a). The incidence of mutant alleles is shown per gamete (germline) or per tumour allele (gastrointestinal tumours), and the corresponding estimated germline mutation rates and frequencies in tumours are shown with 90% confidence intervals. Although the detection of mutants is limited by gel resolution, and the values shown here will be underestimates, the somatic and germline data are comparable, as the same gel electrophoresis conditions were used for each.

Table 5.2

			Germline		Somatic(GI tumours)	
Probe	Locus	%Heterozygosity	Incidence per n gametes	rate per gamete	Incidence per n alleles	frequency
λMS1	D1S7	99.4	36/686	0.052 (0.038-0.072)	9/102	0.088 (0.054-0.15)
λMS32	D1S8	97.5	5/684	0.007 (0.003-0.015)	6/102	0.058 (0.032-0.11)
λMS31	D7S21	98.0	5/684	0.007 (0.003-0.015)	0/102	0 (0-0.029)
pλg3	D7S22	97.4	2/671	0.003 (0.0006-0.009)	1/102	0.009 (0.003-0.046)
λMS43	D12S11	95.9	0/687	0 (0-0.003)	1/102	0.009 (0.003-0.046)
λMS8	D5S43	85.1	0/687	0 (0-0.003)	1/102	0.009 (0.003-0.046)

gastrointestinal tumours. This similar incidence of mutant alleles arises despite the very different cellular histories of the tumours and germ cells. It has been estimated that about 400 and 24 mitoses separate the zygote from the mature sperm and oocyte respectively (Vogel and Rathenberg,1975). By contrast, direct measurements of intestinal crypt cell turnover, if taken at face value, would suggest a cycle time for stem cells of about 24 hours or less (Lipkin *et al.*,1963). While the interpretation of these data relies upon assumptions about the number, recruitment and proliferative behaviour of the stem cells (Potten and Loeffler,1987), the available data would nevertheless suggest that a gastrointestinal epithelial cell giving rise to an adenocarcinoma in an adult would have a history of more than 10,000 postzygotic mitoses.

Some of the observed mutations in tumours will have arisen during their malignant expansion, and thus the observed incidence of mutations in all tumours will provide an upper estimate of their incidence in the corresponding epithelium. The rate of somatic mutation in gastrointestinal epithelial cells *per mitosis* must therefore be very low, of the order of 10^{-5} or less. If this low rate were also applicable to the mitoses occurring in the germline, then the observed rate of germline mutations at the most unstable loci cannot be accounted for by processes coupled to cell division. This may provide some explanation for the observation that the germline mutation rates in human sperm and oocytes are indistinguishable, despite the larger number of pre-meiotic mitoses leading to the formation of a sperm. However, the detection of germline mosaicism for deletion mutant alleles at D1S8 in bulk sperm DNA (Jeffreys *et al.*,1990a) would suggest, in contrast, that a significant fraction of these deletions at D1S8 arise pre-meiotically.

5.2.2 *Somatic changes in human breast cancer*

5.2.2.1 *Introduction*

The last section presented work in which the clonal properties of human gastrointestinal carcinomas were exploited to investigate somatic mutation processes at minisatellite loci. Inversely, minisatellite loci have ideal general properties for the investigation of somatic changes in tumour DNA. Their high population heterozygosity makes them ideal for the study of allele losses from tumours; moreover, if mitotic recombination is a predominant mechanism for allele loss, their preferential subtelomeric location is ideal for the detection of homozygotization of markers distal to the point of recombination. Their high informativeness may also be useful in pedigree analysis in the search for genes predisposing to the development of malignancy, and their high mutation rate to new length alleles optimises the detection of somatic mutations (v.s., section 5.2.1).

Family history has long been recognized as a risk factor for adenocarcinoma of the breast in humans. This familial clustering is not due simply to the coincident occurrence of a common disorder in families, and the familial bias appears to have a genetic, rather than an environmental, cause (King,1982). A model for inherited susceptibility based on one or more dominantly acting genes has been the most favoured explanation, and many large kindreds apparently segregating for susceptibility to breast cancer have been documented (Newman *et al.*,1988). The problems of finding the gene or genes by linkage are, however, formidable, since not only may there be a large number of genes of variable penetrance acting to determine susceptibility, but also the relatively high sporadic rate of breast cancer among western women may confound linkage analysis by disrupting the relationship between linked marker and phenotype.

The recognition that certain tumours, to which susceptibility is inherited by a simple autosomal dominant mechanism, had deleted one allele from a specific chromosomal region, suggested that the inherited susceptibility to these tumours might take the form of a defective tumour suppressor gene (Knudson,1971;Hansen and Cavenee,1987). Loss of the remaining (good) allele at that locus leaves the cell unprotected, and a tumour develops. Hence material from a specific chromosomal location is missing from one homologue in the tumour cells. Retinoblastoma is the paradigmatic example of such a tumour suppressor mechanism, and the detection of allele losses from the 13q14 region led ultimately to the isolation of the gene responsible for predisposition to these tumours (Lee *et al.*,1987; Fung *et al.*,1987). The loss of alleles on chromosome 5q from tumour tissue has also been demonstrated in Familial Adenomatous Polyposis (Solomon *et al.*,1987), and the autosomal dominant inheritance of the syndrome has also been linked to markers on 5q (Bodmer *et al.*,1987).

More complex relations have, however, been demonstrated. For example, while the gene responsible for the inheritance of Multiple Endocrine Neoplasia type IIA maps to chromosome 10 (Mathew et *al.*,1987a), specific deletions from the long arm of chromosome 1 in these tumours have been observed (Mathew et *al.*,1987b). Thus rather than having one copy of a tumour suppressor gene inactivated in the germline and one in the soma, both alleles at a tumour suppressor locus may be inactivated somatically; the gene for inherited susceptibility may act to promote these suppressor losses, or to induce tumour formation when they have occurred.

The following sections report the use of minisatellite probes to screen for somatic changes in human breast tumours. At the time of these investigations, there had been some reports of low frequency allele losses from the short arm of chromosome 11 (Theillet *et al.*,1986) and chromosome 13 (Lundberg *et al.*,1987); however, no chromosomal region frequently deleted from breast cancers had been defined. DNA fingerprinting probes were used to screen many loci for somatic change, and single locus probes were subsequently used to look for changes at defined loci.

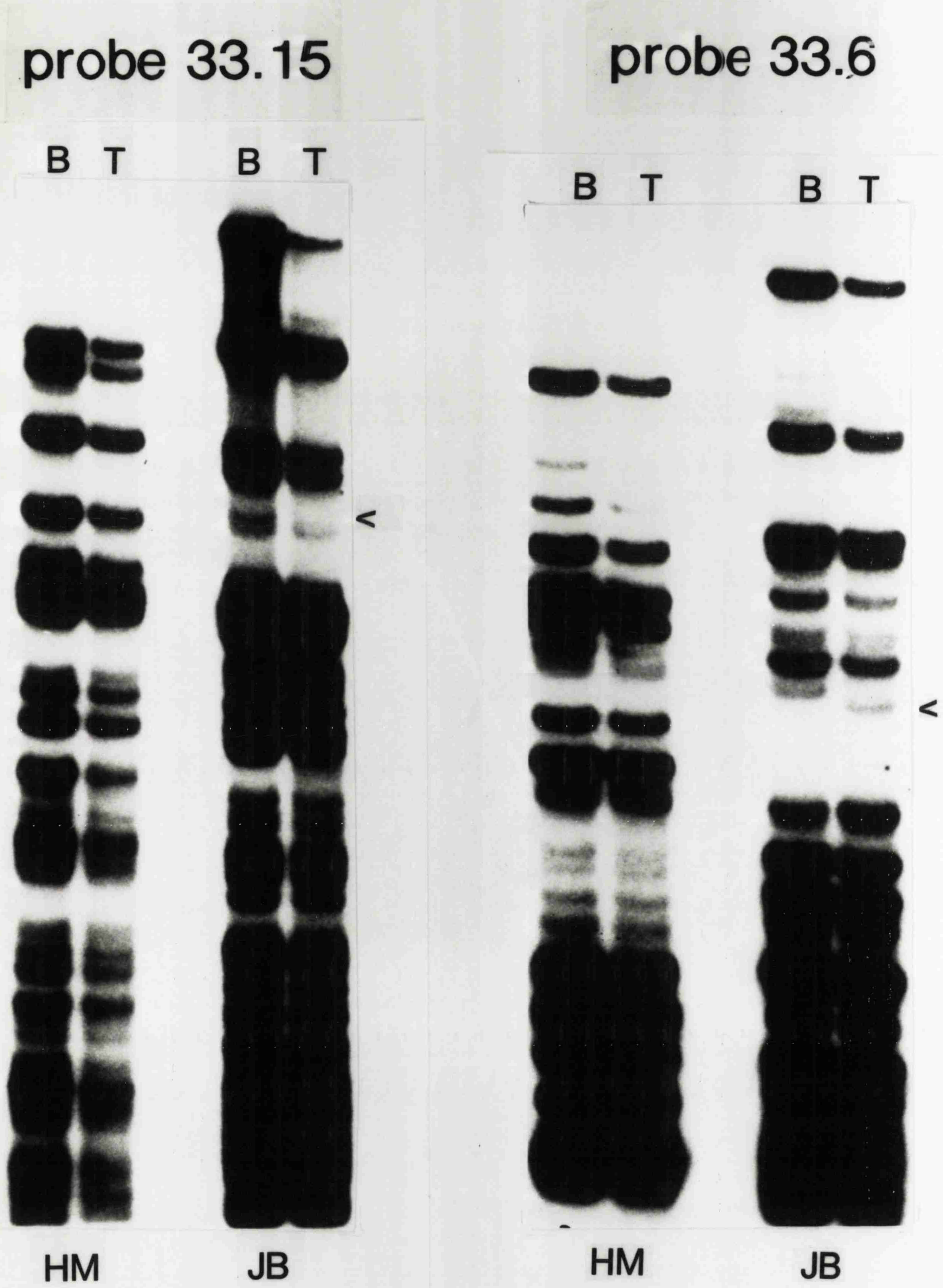
5.2.2.2 *DNA fingerprints of breast tumours*

Clinical samples were collected in Leicester by Dr.Rosemary Walker (Department of Pathology, Leicester Royal Infirmary) and at the Institute of Cancer Research (Royal Marsden Hospital) by Dr.Kristin Anderson and Dr.Bruce Ponder. DNA was extracted from the tumour, and from blood taken from the same patient (section 2.2.1.2). 5μg samples were digested with *Alu*I, and Southern blot hybridized using the DNA fingerprinting probes 33.6 and 33.15 (Jeffreys *et al.*,1985a,b). *Alu*I was used because its recognition sequence (AGCT) cannot contain or overlap with the dinucleotide CG, and so is unlikely to show changes due simply to tumour-specific methylation changes.

Examples of blood (B)-tumour (T) comparisons by DNA fingerprinting are shown in Figure 5.4. The most common result is that the DNA fingerprints of blood and tumour in the same

Figure 5.4

DNA fingerprints of breast cancers. DNA extracted from the blood (B) and tumour (T) of patients HM and JB were digested with *Alu*I and Southern blot hybridized using the DNA fingerprinting probes 33.6 and 33.15. Patient HM illustrates the nearly uniform finding that the DNA fingerprint profiles of blood and tumour are not distinguishable except for minor differences in band intensity attributable to inequality of DNA loading or degradation between the samples. Patient JB illustrates the only changes seen on the DNA fingerprint survey of breast cancers; a new mutant band detected by probe 33.6 and a band loss event detected by probe 33.15.



probe 33.15

probe 33.6

patient are indistinguishable (Figure 5.4, patient HM). This is a little surprising, given the well documented tendency of breast cancers to undergo aneuploid change (Owainati *et al.*,1987). However, although these DNA fingerprinting probes detect loci genetically dispersed in the human genome (Jeffreys *et al.*,1986), simple tetraploidy, for example, should not be apparent, since the relative contribution from each locus remains unchanged. In the two DNA fingerprints of the tumour from one patient (JB, Figure 5.4), two different abnormalities are seen. 33.6 detects a fragment present in the tumour but absent from the blood, suggesting that a new mutant allele at this locus has arisen in the tumour. A fragment of similar size and hybridization intensity is present in the blood but missing from the tumour, and may be the progenitor of the mutant allele. 33.15 detected a fragment in the blood which was absent from the tumour, suggesting allele loss from that locus in the tumour. The other, unchanged loci in these DNA fingerprints serve as internal controls for both equality of DNA loading between blood and tumour and for evenness of electrophoretic resolution.

These were the only lesions detected in 22 blood-tumour pairs each analysed using both probes 33.6 and 33.15, and illustrate both the power and limitations of DNA fingerprinting. The power of the method is to allow a large number of highly informative genetically dispersed loci to be screened for somatic change in a single test. The limitation is that since all the loci detected are hypervariable in size, the locus responsible for a particular DNA fingerprint band cannot be established without using cloned locus-specific probes. It is illustrative that the somatic changes in the tumour from patient JB shown in Figure 5.4 have not been seen using cloned locus-specific probes, and so the loci at which they occurred are still unknown.

5.2.2.3 *Analysis at single minisatellite loci*

The generalised screening of dispersed minisatellite loci by DNA fingerprinting was complemented by the analysis of single minisatellite loci. 5μg samples of DNA from the same patient's blood or tumour were digested with *Alu*I and Southern blot hybridized using cloned minisatellites under locus-specific conditions (Wong *et al.*,1987). In this survey, blood/tumour pairs from 26 patients were studied; an extended set of 38 blood/tumour pairs was used with λMS1 and λMS8. The probes used and the results obtained with them are summarised in Table 5.3.

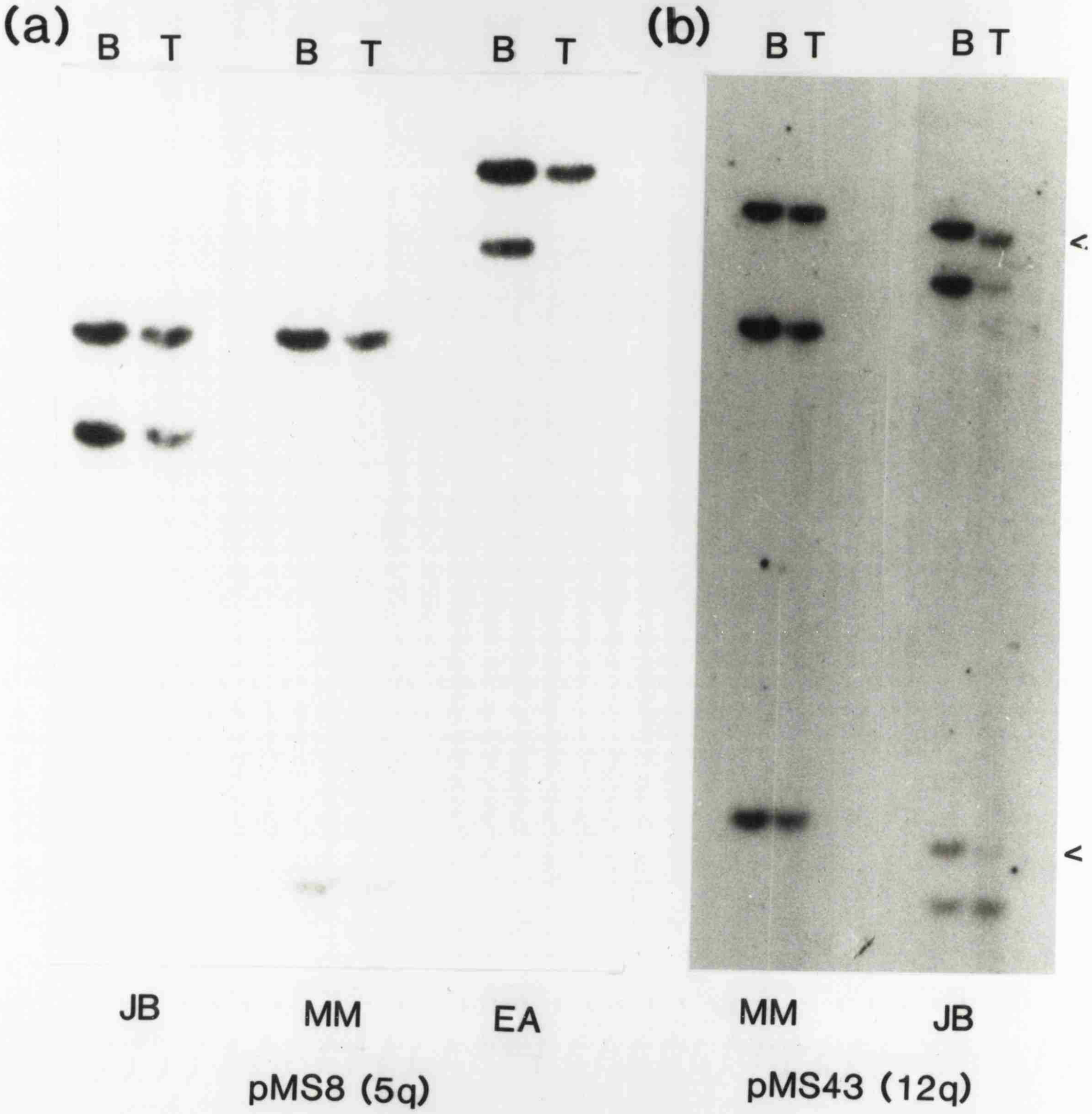
As with the DNA fingerprints of blood/tumour pairs, the profiles at individual minisatellite loci were generally preserved between the two samples. Some somatic changes were, however, defined, both as new mutations and as allele losses. Unlike the DNA fingerprint analyses, these single locus analyses contain no internal control for equality of DNA loading. This may constitute a particular problem when the two alleles at a locus are of very different sizes, and where there may be degradation of tumour DNA. For this reason, the criterion for defining allele loss is the relative loss of more than 50% of the signal from one allele in the tumour.

By this criterion, allele losses were defined at two loci. One of 38 tumours lost an allele at the locus D5S43 detected by λMS8 (Figure 5.5a). A rather higher rate of allele loss (8/34, about 24%) was detected by pMS228 at the locus D17S134. This locus comprises two minisatellites (section 4.3.1.2) which are separated when DNA is digested with *Alu*I, and there are often

Figure 5.5

Molecular lesions in breast cancers defined at single minisatellite loci. (a) shows DNA from the blood (B) and tumours (T) of patients JB, MM and EA digested with *Alu*I and Southern blot hybridized with λMS8 (chromosome 5q). The tumour DNA from patient EA has almost completely deleted the lower allele at this locus. (b) shows DNA from the blood (B) and tumours (T) of patients MM and JB digested with *Alu*I and Southern blot hybridized with λMS43, which contains two distinct minisatellite regions. In the tumour DNA from patient JB, the appearance of a new mutant allele at the larger minisatellite (43A) is accompanied by the deletion of one allele from the smaller (43B) minisatellite.

Breast tumours



large differences in size between the alleles at the major minisatellite 228A. In order to confirm the allele losses seen with *Alu*I, and to minimise the possible effects of DNA degradation, an additional eight blood/tumour pairs were analysed after digestion with *Mbo*I (Figure 5.6). *Mbo*I leaves both minisatellites at this locus on a single restriction fragment, and these larger alleles more frequently occupy a similar size class than alleles with those enzymes which separate the two minisatellites.

Frequent allele losses from the short arm of chromosome 17 had been documented from colorectal tumours (Fearon *et al.*,1987), and have since been reported in breast cancers (Mackay *et al.*,1988) using the probe YNZ22 (Nakamura *et al.*,1987a); YNZ22 appears to be very tightly linked to pMS228 (Ms.Annette MacLeod and Prof.Alec Jeffreys, unpublished work). Alleles were lost from a higher fraction of tumours (61%) than in this work, but Mackay *et al.* do not make clear what criteria are used to define significant allele loss. Further work on colorectal cancer has shown that the losses from the short arm of chromosome 17 centre on the p53 tumour antigen gene (Baker *et al.*,1989), and it remains to be seen whether this is also the fulcrum for the deletions observed in breast cancers, and whether it is one of the genes responsible for inherited predisposition to breast cancer.

In addition to these allele losses, two mutations were observed in DNA from breast cancers. Figure 5.5b shows the appearance of a novel hybridizing fragment detected by λMS43 in the tumour from patient JB. This probe detects two very closely linked minisatellites (section 4.3.1.2, Royle *et al.*,1988). The appearance of a novel fragment at the larger (MS43A)

Figure 5.6

Allele losses from D17S134 in breast cancers. DNA from the blood (B) and tumour (T) of four patients are shown after digestion with *Mbo*I and Southern blot hybridization with pMS228. In the first pattient there is no change in the tumour, but in the other three shown one of the tumour alleles is clearly diminished in intensity.

Breast tumour/blood comparisons
probe: pMS228 (chr.17p)

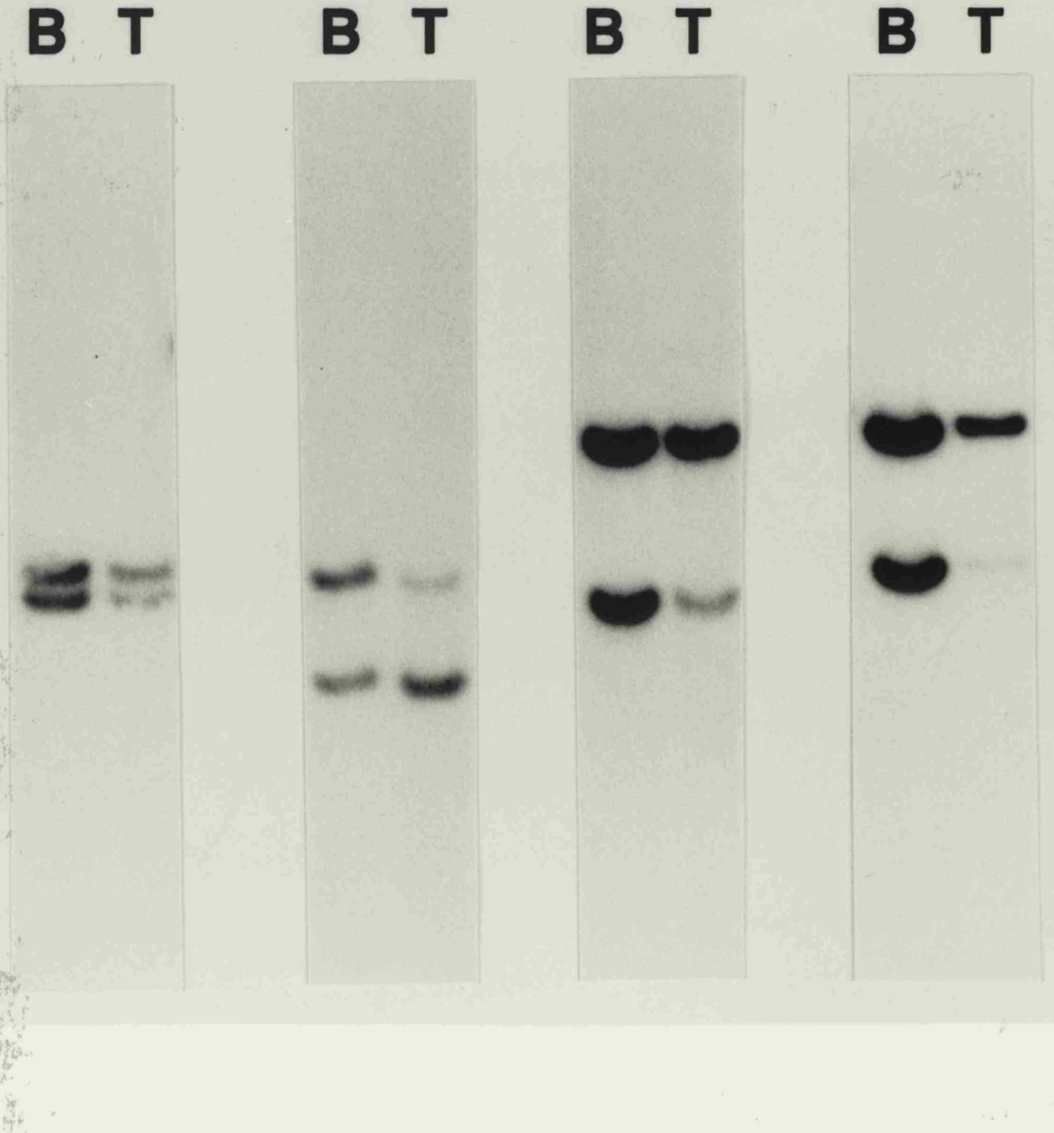


Table 5.3

Summary of somatic changes detected in breast cancers by minisatellite probes.

Where a locus has yet to be assigned a HGML D number, the chromosomal location is shown (column 2). The first four probes have been described by Wong *et al.*(1987). The others are pMS228 (Armour *et al.*,1989b); pHV82 is a hypervariable minisatellite isolated from a plasmid library of large human *Alu*I/*Hae*III/*Rsa*I fragments, and mapped by somatic cell hybrid analysis to chromosome 18 (S.Harris and A.Jeffreys, unpublished work); p33.1 is a plasmid subclone of the clone 33.1 (Jeffreys *et al.*,1985a), since mapped by somatic cell hybrid analysis and linkage to chromosome 9q (unpublished work of A.Jeffreys, R.Neumann and J.A.L.A.); pMS205 is a plasmid subclone from λMS205, which was isolated from further screening of the library described by Wong *et al.*(1987), as described in section 3.2.1.1 (unpublished work of N.Royle, R.Clarkson and A.Jeffreys). *pMS228 detects two variable minisatellites in human DNA digested with *Alu*I, and the heterozygosity level shown is the estimated proportion of the population heterozygous at at least one of the minisatellites.

Table 5.3

Probe	Locus (chromosome)	%Hetero-zygosity	no.informative/ no.tested	allele loss	mutation
λMS1	D1S7	99.4	36/38	0/36	1
λMS8	D5S43	85.1	24/38	1/24	0
λMS31	D7S21	98.0	24/26	0/24	0
λMS43	D12S11	95.9	22/26	0/22	1
pMS228*	D17S134	99*	34/34	8/34	0
pHV82	(18)	90	19/26	0/19	0
p33.1	(9q)	66	9/26	0/9	0
pMS205	(16)	97	22/26	0/22	0

minisatellite is accompanied by the apparent loss of an allele at the smaller (MS43B) minisatellite, suggesting the possible involvement of an unequal mitotic recombination mechanism. This straightforward model, however, is not acceptable as a complete explanation, since both alleles at the larger MS43A minisatellite appear to be reduced in intensity, and not just one as would be expected. Furthermore, the new hybridizing fragment is much closer in size to one of the possible progenitors than the other, and so cannot result simply from the superimposition of two reciprocal mutant alleles arising from an unequal mitotic exchange. The small amount of DNA available from this tumour, combined with the unavailability of PCR analysis at this locus, has meant that this mutation has not yet been further characterized.

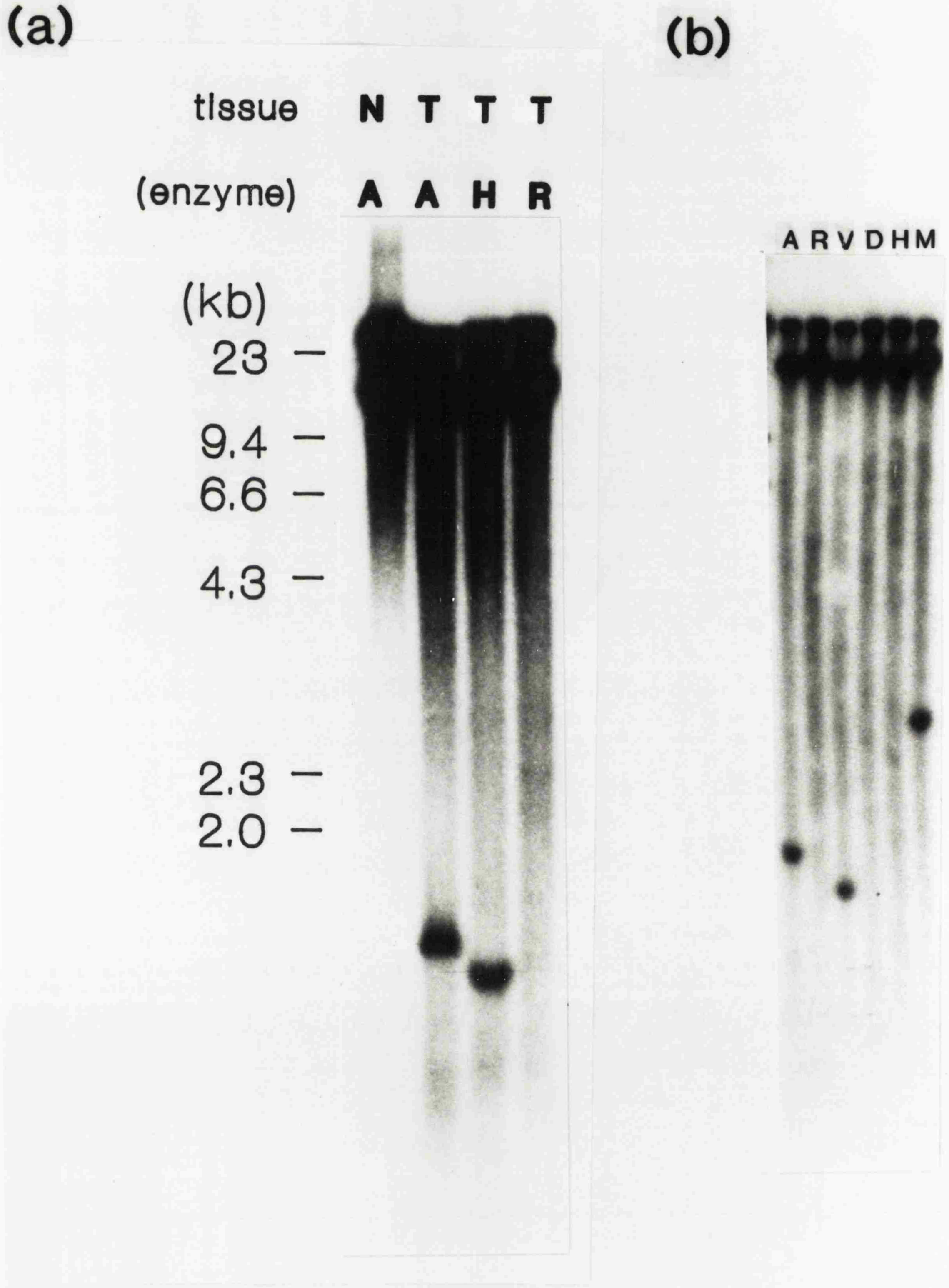
The only other example of somatic change observed was an unusual mutation detected by λMS1, which is described in the next section. In summary, while the overall level of somatic change at minisatellite loci in breast tumours is low, allele losses can be detected, most frequently in this study from the short arm of chromosome 17, as can somatic length change mutations at minisatellite loci, although at a much lower frequency than in gastrointestinal tumours.

5.2.2.4 *An unusual mutation event*

A novel hybridizing fragment was detected in tumour DNA from patient JB by λMS1 (Figure 5.7). The two alleles at D1S7 in this patient were both very large (about 16 and 30kb), and the newly appearing fragment was very much smaller, about 1kb with *AluI*. The new fragment appeared to contain tandem repeats very similar in sequence to those of λMS1, as it was detected even after washing filters to very high stringency (0.05xSSC at

Figure 5.7

Detection of an unusual mutation by λMS1 in tumour DNA from patient JB. In (a) DNA from blood (N) and tumour (T) has been digested with the enzymes *Alu*I (A), *Hin*fI (H) and *Rsa*I (R). A new hybridizing fragment is visible in the tumour DNA digested with *Alu*I or *Hin*fI. The fragment is much smaller than the unmutated alleles in this patient (both >15kb) and is not seen when tumour DNA is digested with *Rsa*I. (b) shows tumour DNA from JB digested with restriction enzymes *Alu*I (A), *Rsa*I (R), *Ava*II (V), *Dde*I (D), *Hae*III (H) or *Mbo*I (M). The new fragment is seen with *Alu*I, *Ava*II and *Mbo*I, but not with *Rsa*I, *Dde*I or *Hae*III.


(a)
tissue N T T T
(enzyme) A A H R
(kb)
23
9.4
6.6
4.3
2.3
2.0
(b)
A R V D H M

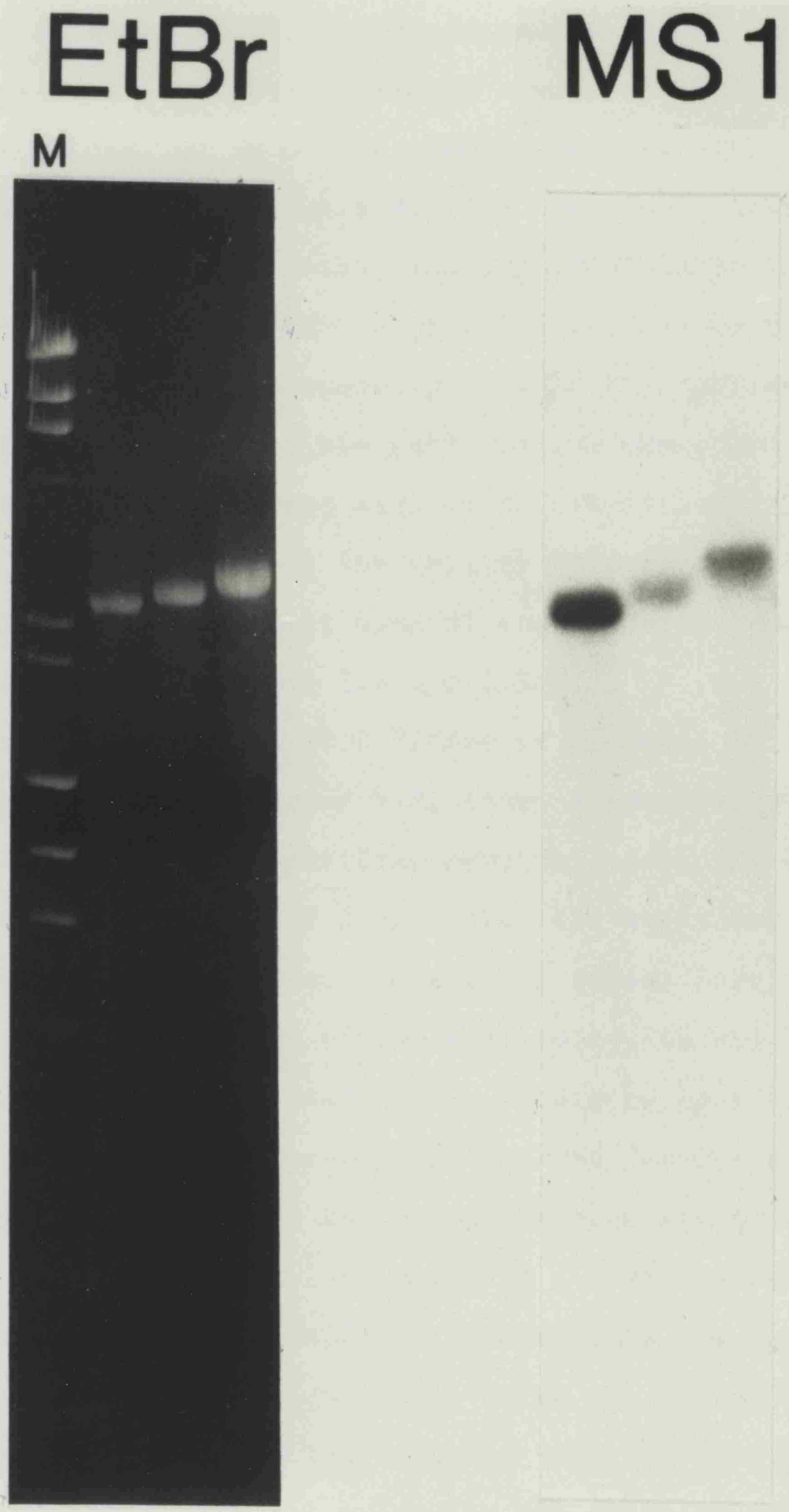
65°C). It was not a highly shortened length change mutant at this locus, as the relative sizes with different restriction enzymes (for example *Alu*I and *Hinf*I in Figure 5.7a) did not match those predicted from the flanking DNA sequence. Indeed, some enzymes, such as RsaI, failed to show the new fragment at all (Figure 5.7a).

A novel tumour-specific band which appears with some enzymes but not others could be explained by two main mechanisms; firstly, a rearrangement at D1S7 itself in which a block of tandem repeats had become separated from the rest of the allele by the interposition, near one end of the allele, of non-repetitive DNA. Any enzyme which cleaved within the inserting DNA would cut off a novel fragment, whereas enzymes which cut both blocks out on one fragment would fail to show the rearrangement. Alternatively, the "new" fragment could be due to events at a distinct locus, at which tumour-specific appearance of sequences similar to the λMS1 repeat unit had occurred. In this model one would predict that some enzymes might cleave in this repeat unit, and that any such enzymes would fail to show the new fragment. The novel fragment was observed with *Mbo*I but not with *Bam*HI. Since any repeat unit cleaved by *Bam*HI (GGATCC) would also be cleaved by *Mbo*I (GATC), this result suggested that the non-appearance of the novel fragment with some enzymes was not due to cleavage within a repeat unit.

The practical problem in determining the structure of the novel band was the small amount of DNA available from the tumour. For this reason, it was isolated using a modification of the "whole-genome PCR" method (Kinzler and Vogelstein,1989). The *Mbo*I fragment (about 2.4kb, Figure 5.7b) was chosen for

Figure 5.8

Initial isolation by whole genome PCR of a fraction containing the novel hybridizing fragment detected by λMS1 in the tumour from patient JB. 1.8 to 2.5kb *Mbo*I fragments were purified from tumour DNA and ligated to *Sau*3AI linkers, made by phosphorylation of the 24-mer oligonucleotide "SauLB" - 5'GATCCCCAAGCTTCCCGGGTACCGC3', followed by annealing to the 20-mer "SauLA" - 5'GCGGTACCCGGGAAGCTTGG3'. The tumour DNA/linker ligation products were separated from linker dimers by gel electrophoresis, during which three subfractions were prepared. These subfractions were amplified using PCR with SauLA as the primer, and the products analysed by Southern blot hybridization with λMS1. This figure shows the results of further amplification of three tighter subfractions prepared from the 2.3-2.5kb subfraction amplified in the initial experiment. The amplification products are shown after ethidium bromide staining (EtBr,left), which shows faithful amplification of the subfractions prepared (M = λ/*Hin*dIII and ΦX174 RF/*Hae*III size markers). To the right ("MS1") the results of hybridization analysis, in which a fragment (about 2.35kb) hybridizing strongly with λMS1 is shown in the smallest of the amplified size fractions.

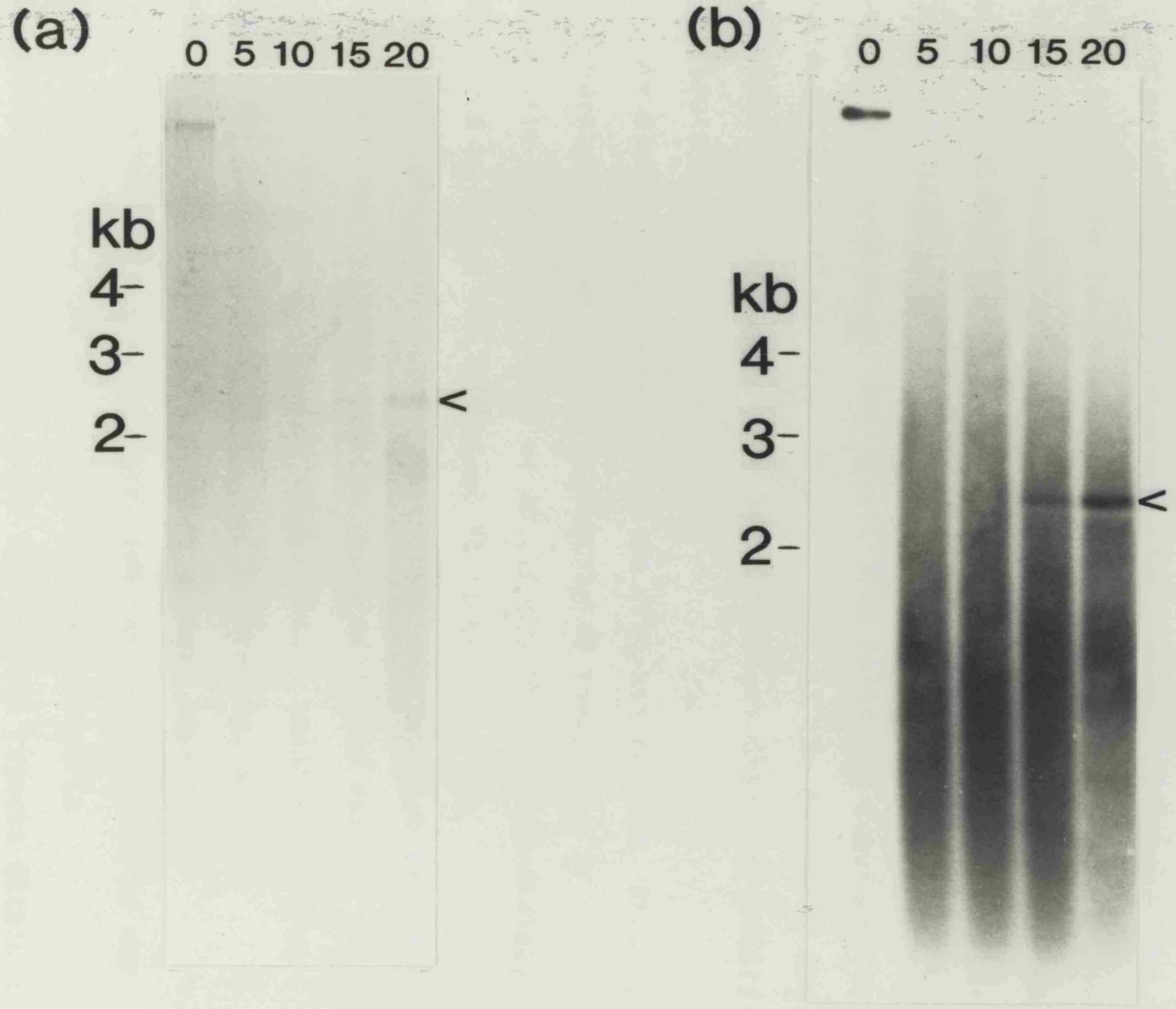

EtBr
MS1
M

isolation; 1.8 to 2.5kb *Mbo*I fragments were isolated from tumour DNA by gel electrophoresis, and *Sau*3AI linkers (Figure 5.8 legend) ligated onto the *Mbo*I ends at an initial molar ratio of about 1000:1. Large fragments (1.8-2.5kb) were separated from linker dimers by a second round of gel purification, at which the tumour DNA molecules, now with linkers at each end, were divided into three size subfractions. Each of these subfractions was amplified using the shorter "SauLA" oligonucleotide (Figure 5.8 legend) as primer. The amplified DNA corresponded in size to the subfractions used as input DNA, and one of the subfractions contained an amplified fragment of the correct size which hybridized strongly with λMS1. Figure 5.8 shows the results of a second round of amplification from this hybridizing fraction, using PCR products between about 2.3 and 2.5kb.

From an estimated 5000 different products in the hybridizing fraction shown in Figure 5.8, those molecules hybridizing with λMS1 were selected by filter hybridization. The details are given in section 2.3.2.3. Briefly, the amplified fraction was denatured and hybridized to a nylon filter bearing cloned pMS1; after washing to high stringency, molecules which had bound specifically to the filter were recovered by alkali treatment followed by ethanol precipitation. The results of PCR amplification of this recovered fraction are shown in Figure 5.9, and show the amplification of a 2.4kb product hybridizing strongly with λMS1. There is considerable non-specific background in this experiment, which may represent cloned MS1 sequences which have been removed from the filter; since, however, these are not tagged by the whole genome PCR linkers, they do not undergo amplification.

Figure 5.9

Purification of the fragment hybridizing with λMS1 by filter hybridization selection. DNA amplified from tumour DNA by whole genome PCR was alkali denatured and hybridized with cloned pMS1 attached to a small nylon filter (for details, see section 2.3.2.3). After hybridization at 65°C overnight, the filter was washed to 0.1 x SSC at 65°C. Molecules which had formed hybrids with pMS1 molecules on the filter were recovered by treatment with alkali, followed by ethanol precipitation of DNA from the alkali washes. This recovered DNA was amplified using the "SauLA" primer (see Figure 5.8 legend). The figure shows the appearance of amplification products after 0, 5, 10, 15 and 20 cycles of amplification. (a) shows the appearance after agarose gel electrophoresis and ethidium staining (photographic negative), while (b) shows the results of Southern blot hybridization with λMS1. Both analyses show a specific, amplified and hybridizing band at about 2.4kb, together with hybridizing material of heterogeneous size which does not undergo amplification. This non-specific material may be due to removal of cloned pMS1 from the filter by alkali treatment, but was simply separated from the required fragment by size fractionation and re-amplification.


(a)
0 5 10 15 20
kb
4-
3-
2-
<
(b)
0 5 10 15 20
kb
4-
3-
2-
<

The fragment was cloned by a round of gel purification followed by reamplification, cleavage with *Mbo*I and ligation into the *Bam*HI site of pBluescriptII KS$^+$. Plasmid transformants were screened by hybridization with λMS1; in fact, most recombinant clones appeared to contain cross-hybridizing sequences. DNA was prepared from six of these clones, and restriction digestion suggested that they all contained the same insert DNA in one of the two possible orientations. More detailed restriction mapping (Smith and Birnstiel,1976) of two of these clones (pJBT1 and pJBT2) demonstrated that they indeed contained inserts of indistinguishable structure but in opposite orientations. The insert from pJBT1 was subcloned into pBluescriptII SK$^+$ to give pJBT10; pJBT10 is the origin of the sequence shown in Figure 5.10 and summarized diagrammatically in Figure 5.11d.

The cloned insert contains three distinct regions. Firstly, a tandemly repeated region consisting of human type II satellite repeats (Prosser *et al.*,1986); this is immediately followed by about 750 bp (about 85 repeats) of MS1 minisatellite repeat units. This block has not been entirely sequenced, but clearly contains three different variant repeats. Variant MS1 repeats within normal human alleles at D1S7 have been analysed by PCR and sequencing, and characteristic patterns of variant repeats have been demonstrated for the extremities and central portions of MS1 alleles (Ian Gray and Alec Jeffreys, manuscript in preparation). The disposition of the variant repeats within the novel array in pJBT10 is typical of that seen within normal human alleles at MS1, but does not contain any patterns typical of the ends of arrays. The third region is a non-repetitive

Figure 5.10

Sequence of novel fragment detected in tumour JB by λMS1. The sequence shown was determined from a series of nested deletions prepared from pJBT10, and lacks about 300bp from the 3'end. The sequence contains three distinct domains: human type II satellite sequence (1-555) is shown in bold; the MS1 minisatellite repeats (558-804) are shown in capitals, and the remainder of the cloned insert appears to be single copy sequence. The MS1 repeat block has not been entirely sequenced, and "nnn" indicates the positions of the two gaps in the sequence. In this figure the *Sau*3AI site (GTAC) defining the start of the clone, and the *Pst*I site (CTGCAG) used to subclone pJBT11 (see Figure 5.11b and text), are shown underlined; satellite and MS1 repeats are shown with arrowed underlining. The main features of pJBT10 are summarized in Figure 5.11d.

```
   1  gatcacactggatttcattccataattctattcgattccattcgatgatg
  51  attccattcatttccatccgatgatgattccattcgattccgttcaatga
 101  ttattccattcgactccactcgatgattccattcgattccattcgatgat
 151  gattgcattcgagtccatggattattccattccattccattacatgattc
 201  cattcgggtccattcgatgattctcttcgattccattcgataattccgtt
 251  tttttccgtttgatgttgattccattcgattccattagatgataattcca
 301  ttcgattctatgcgatgataccattctattccatttgaagatgattccat
 351  tcgagaacattcgatgattgcattcaattcactcgatgacgattccattc
 401  aattccgttcaatgattccatttgattccatttcacgttgattccattcg
 451  attccattttatgatgattccatgcaattccattagatgacgactccttt
 501  catttccattcggtgacgattctatcgtttccatccgatgatgattccat
 551  tcgattccgttcaatgaAGAGGGTGGAGAGGGTGGACAGGGTGGACAGGG
 601  TGGAnnnGGATAGGGTGGACAGGGTGGACAGGGTGGACAGGGTGGACAGG
 651  GTGGACAGGGTGGATAGGGTGGACAGGGTGGACAGGGTGGACAGGGTGGA
 701  CAGGGTGGACAGGGTGGATAGGGTGGACAGGGTGGAnnnGGACAGGGTGG
 751  ATAGGGTGGACAGGGTGGACAGGGTGGATAGGGTGGACAGGGTGGACAGG
 801  GTGGcggctgggcagggctgctcctctacctgtggaccctggtagcccca
 851  ctcctctgcgcaaccgcgacttctgctgaggcagcctcacagcctgccat
 901  ctggtgcctcctgccacctggtgcctctcggctcggtgacagccaacctg
 951  cccccteccacaccaatcagccaggctgagcccccacccctgccccagc
1001  tccaggacctgcccctgagccgggccttctagtcgtagtgccttcagggt
1051  ccgaggagcatcaggctcctgcagccccatcccccgccacacccacacg
1101  gtggagctggctcttccttccctcctccctgttgcccatactcagcatct
1151  cggatgaaagggctcccttgtcctcaggctccacgggagcggggctgctg
1201  gagagagctgggaactcccaccacagtggggcatccggcactgaagccct
1251  ggtgttctggtcacgtccccaggggacccctgccccttcctggacttcg
1301  tgccttactgagtctctaagactttttctaataaacaagccagtgcgtgt
1351  accatgttctgtgcccctcaccctcagcacggagccccactgcatggggg
1401  ccggtgtgggggtttgggaatagaatgtttagggctgaggaggctgggac
1451  atcagggccagaccaggaggagcctcaaaggcagacagaatggcctgagt
1501  tctgtcttctgggtcatggagcgcctgaggga
```

sequence which bears no relation to the known DNA sequence flanking MS1.

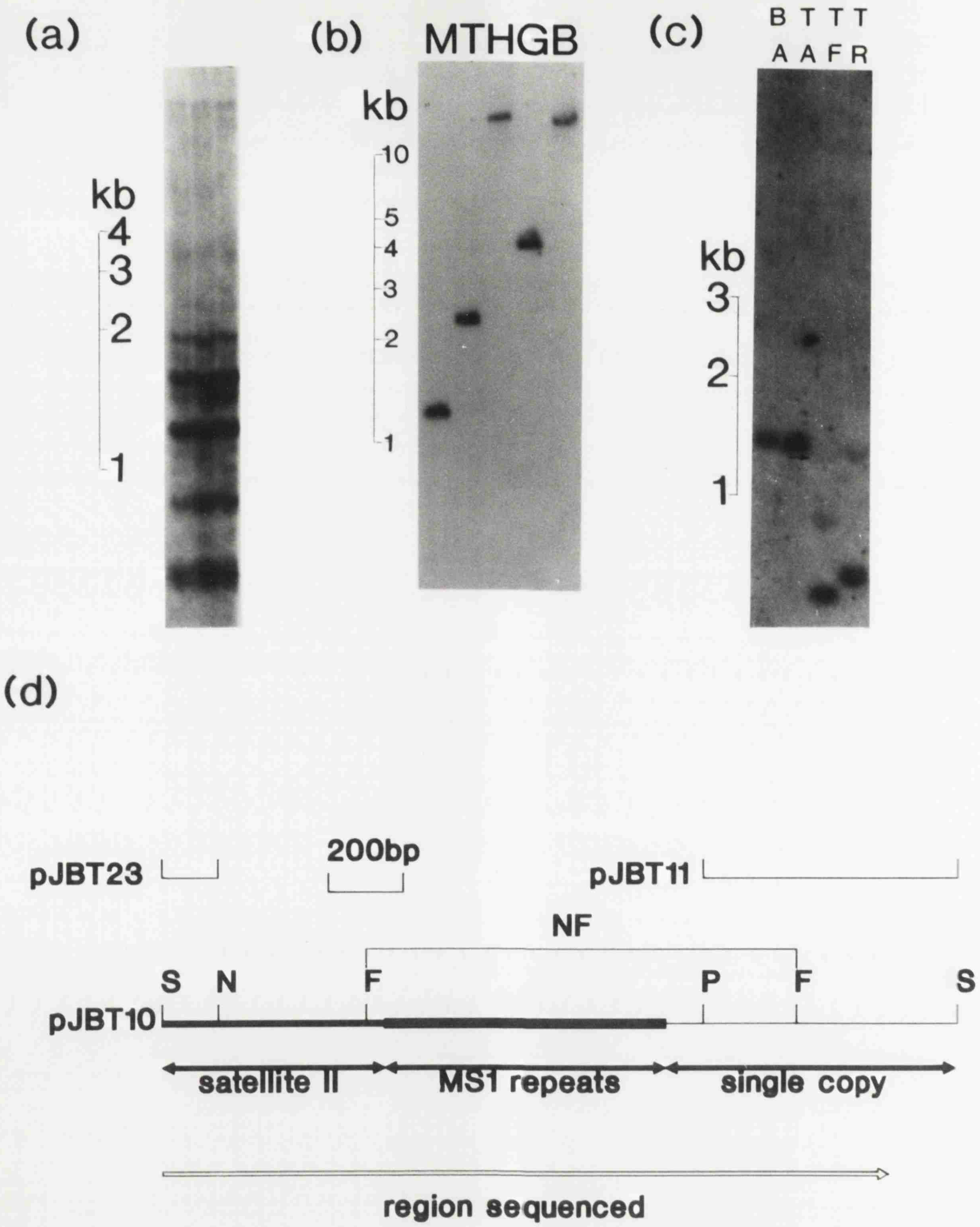
Probes were prepared from the satellite and non-repetitive regions, and used to probe human DNA. As predicted, the satellite sequence detected a monomorphic "ladder" of higher-order repeats in DNA digested with *Mbo*I (Figure 5.11a). pJBT11, from the non-repetitive region, detected a single copy sequence (Figure 5.11b) and analysis of somatic cell hybrids (section 2.1.5) suggested that this locus is, like D1S7 itself, on chromosome 1.

Faced with such a bizarre structure in this DNA fragment, it would be useful to have confirmatory evidence that this indeed is the structure present in the novel fragment in this tumour. From the sequence of pJBT10 it is possible to predict the expected sizes of the fragments cross-hybridizing with MS1; these predicted sizes closely match the sizes seen originally in genomic DNA from the tumour, with the possible exception of *Alu*I (>1.5kb predicted, 1.3kb observed). Furthermore, the nearby presence of a satellite array may explain why the novel fragment is not seen with some frequently-cutting enzymes, such as *Rsa*I. If the satellite were only very infrequently cleaved by *Rsa*I, then the novel MS1 repeat block would appear on that very large restriction fragment, and would be obscured by the presence of the progenitor alleles at D1S7 (Figure 5.7a).

The main evidence that this PCR product is a faithful copy of the tumour DNA comes from the analysis of the tumour DNA cleaved with *Hin*fI. This enzyme, unlike many of the others used to study the tumour DNA, does not cleave between the unique sequence probe (pJBT11) and the putative insertion site, and thus one would predict that pJBT11 would detect a novel

Figure 5.11

Restriction mapping using probes from pJBT10. (a) shows DNA from three unrelated people digested with *Mbo*I and probed with pJBT23, a subclone from pJBT10 which contains only satellite repeats (see (d)). As predicted, the probe detects a large number of fragments, in a "ladder" of higher-order repeats. (b) shows DNA from the same individual digested with *Mbo*I (M), *Taq*I (T), *Hin*dIII (H), *Bgl*II (G) and *Bam*HI (B), and probed with the subclone pJBT11, which contains a *Pst*I-*Sau*3AI fragment from the non-repetitive region of the cloned insert (see (d)). This detects a single-copy sequence in human DNA. (c) shows DNA from patient JB's blood (B) digested with *Alu*I (A), and DNA from the tumour (T) digested with *Alu*I (A), *Hin*fI (F) and *Rsa*I (R). This is the same filter as in Figure 5.7a, and was probed with pJBT11. The arrow highlights a novel *Hin*fI fragment which is the same size as the fragment seen with λMS1 (Figure 5.7a); all other hybridizing fragments seen are also detected in normal DNA (data not shown). (d) summarizes the structure of the cloned insert from pJBT10, and includes selected restriction sites. It shows the extent of the sequenced region, as well as the novel fragment ("NF") predicted with *Hin*fI digestion, and the position of DNA subcloned in pJBT11 and pJBT23. S = *Sau*3AI, N = *Nco*I, P = *Pst*I and F = *Hin*fI.


(a)
kb
4
3
2
1
(b)
MTHGB
kb
10
5
4
3
2
1
(c)
B T T T
A A F R
kb
3
2
1
(d)
pJBT23
200bp
pJBT11
NF
S
N
F
P
F
S
pJBT10
satellite II
MS1 repeats
single copy
region sequenced

fragment in tumour DNA cleaved with *Hin*fI ("NF" in Figure 5.11d). This novel fragment, moreover, should be the same one as detected by MS1 in tumour DNA cut with *Hin*fI. The appearance of this predicted diagnostic fragment in tumour DNA cleaved with *Hin*fI is shown in Figure 5.11c (compare Figure 5.7a, in which the same filter is probed with λMS1). DNA from this tumour is now in very short supply, and further confirmation may depend on the demonstration of a novel fragment in tumour DNA by PCR, although the design of a PCR amplimer from the satellite side may be problematical.

If, then, we can take the structure amplified and cloned at face value, it is indeed extraordinary. Firstly, MS1 repeat units are found "ectopically", in completely the wrong context. Secondly, this incorrect context not only appears near a satellite block, but seems to be on the very edge of the satellite array. If in normal DNA the unique sequence in pJBT11 really occurs at the very edge of a satellite block, then this sequence may be of considerable interest in itself.

How might the minisatellite have appeared in this unusual location? The question of why this particular site was the target for the putative event will require analysis of this region in normal DNA, but if it is in fact the boundary of a large satellite block, it may adopt unusual chromatin structures. The fact that the ectopic MS1 contains variant repeats, intermingled just as in normal alleles, suggests that the MS1 block may have transposed to its new location from one of the established alleles. If the repeats had arisen by *de novo* reiteration at this locus, one would predict that the array would be invariant in repeat unit sequence, have a very simple higher-order structure, or carry variant repeats unlike those seen in normal alleles at D1S7. There are no sequence features suggestive of retroposition flanking the MS1 array, and thus a DNA-mediated jump, via a circular intermediate, by gene conversion or by illegitimate (double) recombination, is favoured. It may, though, be no coincidence that the target site for this insertion is on the same chromosome as the locus of origin, and this may indicate a more local event than circle-mediated transposition.

5.2.2.5 Summary: clustering of mutations

In section 5.2.2, a number of molecular lesions have been defined in the DNA from breast cancers. Deletions from the D17S134 locus detected by pMS228 occurred in about one quarter of the tumours studied. Among the less frequent changes observed were the loss of a hybridizing fragment from a DNA fingerprint profile with probe 33.15, a new mutant band in a DNA fingerprint using 33.6 (section 5.2.2.2), allele loss at the D5S43 locus detected by λMS8, a new hybridizing fragment detected by λMS43 (section 5.2.2.3) and an extraordinary transposition detected by λMS1 (section 5.2.2.4). It is remarkable that a single tumour sample, from patient JB, is the source of all but one of these less common lesions. The DNA fingerprint profiles provide a quality control check against contamination of the tumour sample, and it seems that this tumour, which is unusual among those tested in (a) being a lobular carcinoma and (b) showing calcification on histology, has a significant clustering of mutation events. This accords with the predictions of the "error-catastrophe" hypothesis (Orgel,1963), in which the rate of accumulation of mutations undergoes positive feedback, as mutations interfere with the cell's machinery for the detection and correction of mutational change.

5.2.3 *Somatic mutation in normal tissues?*

The analysis of somatic mutation in clonal tumour cell populations, while obviating the problem caused by the polyclonality of normal tissues, is open to the criticism that the mutations observed are simply the result of the malignant phenotype, and give no information about somatic mutation processes in normal tissues. Indeed, the clustering of mutations in the tumour from patient JB suggests that many events may occur during the malignant phase of the cellular history of the tumour.

Hair roots consist of small groups of normal cells, and in development may have a late common cellular ancestor. DNA was extracted from a series of hair roots from different parts of the body of a male volunteer. Not all yielded enough DNA to type using minisatellite probes, but among those that did, no evidence for somatic mutation or mosaicism was detected in 15 roots tested with λMS1, or in 8 roots tested with λMS32 (data not shown). A further 14 hair roots from a different male volunteer have been tested at these two loci in experiments of Prof.Alec Jeffreys without any evidence of mutation. Thus no evidence for gross somatic mosaicism in hair roots has been found using those minisatellite probes most successful in detecting somatic mutation in gastrointestinal tumours.

The high prevalence of somatic mutations in gastrointestinal, but not other types of carcinoma (Thein *et al.*,1987) suggested that among normal tissues, gastrointestinal epithelium might harbour the highest incidence of somatic mutations. DNA was extracted from normal ileal mucosa (collected at Leicester

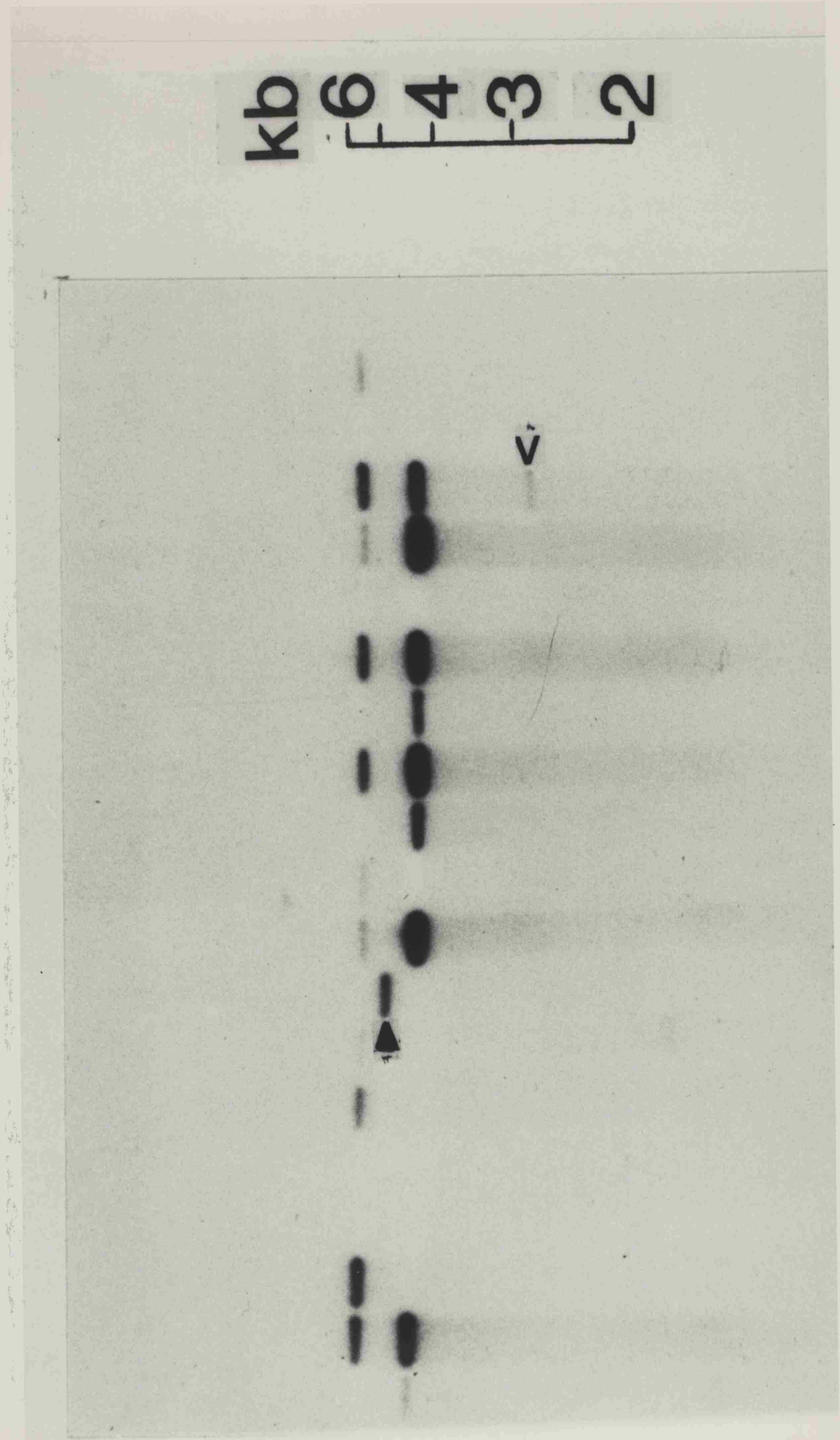
Royal Infirmary by Dr.Ian Talbot), and PCR amplification was used in an attempt to detect somatic mutation in normal cells. 20 samples, each of about 30pg DNA, were prepared. Each of these samples would be predicted to contain about 10 target molecules, 5 from each allele. However, there will be random variations in the number of amplifiable molecules from each allele between the samples. Furthermore, the efficiency of PCR amplification from single target molecules is much less than unity, and in practice is only about 40% (Li *et al.*,1988;Jeffreys *et al.*,1988b).

From a probable total of 200 target molecules, about 30 underwent PCR amplification. This very approximate figure was estimated from the number and intensities of amplified products (Figure 5.12). Among the amplified products were two which did not correspond to unmutated alleles for this individual. One (open arrow in Figure 5.12) was present at lower than expected intensity for the product of a single initial target molecule, and may have arisen as an aberrant product during the amplification. The other (closed arrow in Figure 5.12) cannot be so explained, since it is the only amplified product in that sample. It may therefore represent a genuine mutant allele from ileal mucosa; if so, the incidence (1 in 30) would be high enough to make recovery of large numbers of somatic mutant alleles from this tissue feasible without prior size-selection. This in turn would include in the analysis the large class of minisatellite length mutants which are too close in size to their progenitors to be recovered efficiently by size-selection methods (Jeffreys *et al.*,1990a).

However, it is also possible that this amplified product is due to contamination of the ileal DNA with a low level of

Figure 5.12

Amplification of alleles from single target molecules at D1S8 in ileal DNA. DNA was prepared from normal ileal mucosa and 20 diluted samples, each containing about 30pg of genomic DNA, were amplified at the locus D1S8, using primers 32A and 32B for 25 cycles (see Table 2.2). The figure shows the autoradiograph after hybridization of PCR products with λMS32. Two non-canonical products are indicated; a product present at less than unit dosage, which may have arisen during the amplification (open arrow), and a product (closed arrow) which is present at a level suggestive of a single initial target molecule, which may represent a pre-existing mutant allele.


kb
6
4
3
2
v

extraneous human DNA. Although it was possible to amplify single target molecules from these alleles at relatively high efficiency, their size made internal mapping by MVR analysis (section 4.4.1) impracticable. In an individual with alleles small enough to map efficiently, a correspondence between the internal maps of the progenitor alleles and the putative mutant would be strong evidence in favour of an origin *in vivo* as somatic mutations.

Firmer evidence for somatic mutation in normal tissues has been gained from studies on peripheral leukocytes, in which deletion mutant alleles at the minisatellite locus D1S8 were size-selected, amplified and internally mapped (Jeffreys *et al.*,1990a). While that study did not include the most frequent class of (small) length change mutations, it does show that PCR amplification from single target molecules can be used to demonstrate somatic mutations in bulk DNA from a non-neoplastic tissue.

5.3 *EVOLUTIONARY CHANGE AT HUMAN MINISATELLITE LOCI*

5.3.1 *Evolutionary comparisons*

While our knowledge of the individual events of minisatellite evolution comes from the analysis of germline mutation (section 5.1.1), a broader view of evolutionary change derives from comparisons of states assumed to differ by accumulated mutational events. In the following section three kinds of comparison are made: firstly, between the different alleles at a slowly-evolving minisatellite locus (section 5.3.2); secondly, between the tandemly repeated and "null" alleles at the D12S40 minisatellite locus (section 5.3.3); and thirdly, between human and primate DNA at the D17S134 locus (section 5.3.4).

5.3.2 *Relation between alleles at D22S163*

The internal mapping of alleles at the "607A" minisatellite at D22S163 has been described in section 4.4.3.3. This minisatellite is not very variable (heterozygosity 50%), and only 6 differently sized alleles were detected in a survey of 16 unrelated individuals. This small repertoire of alleles, however, provides the opportunity to survey most or all of the alleles in the population, and thereby to determine whether different alleles are of similar internal structure.

The results of these analyses are shown in Figure 5.13. The letters A, B, C and D are used here as symbols for the four different types of repeat unit defined by MVR mapping using *Sph*I and *Bst*EII (Figure 5.13 legend). Ten alleles were successfully mapped, and belonged to five different types. In all cases alleles were mapped from known heterozygotes, so that

Figure 5.13

Internal maps of alleles at the 607A minisatellite on chromosome 22.

(a) Alleles were amplified and internally mapped using *Sph*I and *Bst*EII. The four kinds of repeat unit seen are here designated A, B, C, and D. These symbols correspond to the repeat unit cleavage patterns:

	*Sph*I	*Bst*EII
A	+	+
B	-	+
C	+	-
D	-	-

The ambiguity codes used for repeats of uncertain type in allele 1 and 6 are: Y=C or D; N=any type. The alleles are numbered according to the size classes seen on Southern blot hybridization, and internal mapping shows that identity of size for the alleles mapped corresponds to true isoallelism.

(b) Relations between alleles at 607A. The introduction of a gap into the map of allele 5 makes clearer the close relation between alleles 4 and 5. Below is shown allele 1 aligned with the 5′ end of allele 6 and the 3′ end of allele 4 or 5.

(a)

Allele	Number of examples	Structure
1	1	AAACCACYDCDDDDNBDCDDDCADDCBCCBCCCCBDAN
2	2	AACCADDBCDDADDBDCBDCCCCBDCCCBB
3	(N.D.)	
4	5	AACDDDCCCCCCBCCCCCBDAB
5	1	AACDCCCCCCBCCCCCBDAB
6	1	AAACCACDDCDDBCCCBCCCCBNBB

(B)

```
    4       AACDDDCCCCCCBCCCCCBDAB
            ||||  ||||||||||||||||
    5       AACD--CCCCCCBCCCCCBDAB


    1       AAACCACYDCDDDDNBDCDDDCADDCBCCB-CCCCBDAN
            ||||||||||||             | ||| ||||||||
    6       AAACCACDDCDD...       ...CCCCBCCCCCBDAB 4 or 5
(5′ end)                                           (3′ end)
```

a mixture of two alleles of identical length would not be inadvertently mapped. In fact, in those cases where alleles of the same length were mapped (five copies of allele 4, two of allele 2), they had identical internal maps. Thus length identity appears to imply isoallelism at this locus.

All the alleles mapped begin with two A-type repeat units, and there is a strong tendency for clusters of B- and C-type repeats to appear near the 3′ end of alleles (bold in Figure 5.13a). Furthermore, there is a clear relation between the structures of alleles 4 and 5 (Figure 5.13b). This suggests the recent divergence of allele 5 (rare) as a variant of allele 4 (common). There is also a possible relationship between alleles 1, 4 (or 5) and 6 (Figure 5.13b); allele 1 is similar to allele 6 at the 5′ end, but to allele 4 and 5 at the 3′ end, suggesting the possible origin of allele 1 as a recombinant between the alleles ancestral to 4 and 6, or indeed of allele 4 or 6 by recombination between allele 1 and the precursor to 4 or 6. It is also possible that alleles 4, 5 and 6 arose by divergence from allele 1 as a common progenitor, of which the ends held in common are the vestiges. However, there is very little overall internal similarity preserved between the alleles mapped, and it may be that these main groups of alleles at 607A have long been established, and have since been evolving largely in independent haploid lineages.

5.3.3 *Analysis at D12S40*

A "null" allele (section 3.3.6.2, Figure 3.7b) was detected at the locus D12S40 detected by cMS608. This unusual allele provided the opportunity to investigate a minisatellite at an unusual stage; in the "null" state the locus is either in a

Figure 5.14

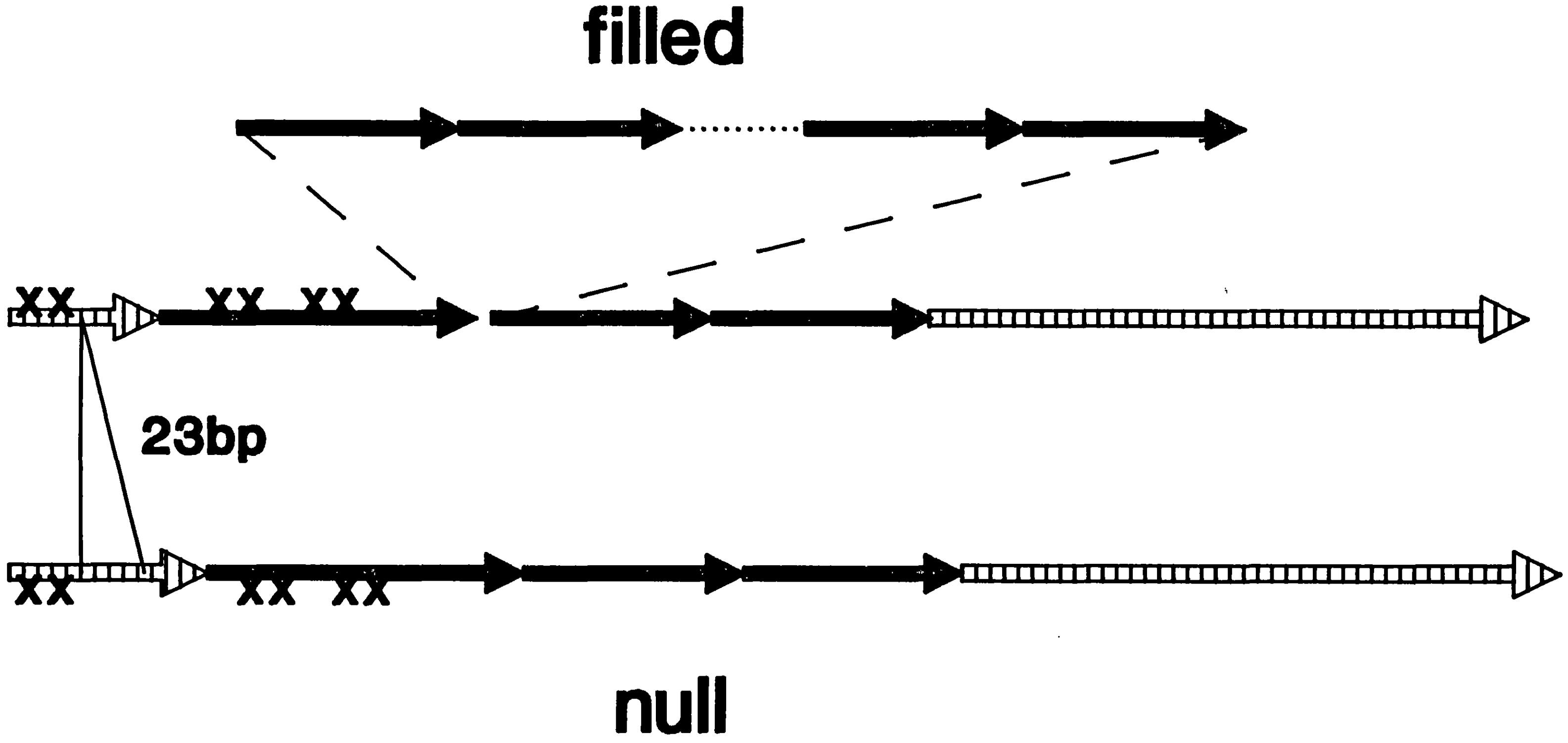
DNA sequence of the null allele and comparison with that of longer tandemly repeated alleles at D12S40 (two pages). The null allele sequence shown (above) is from a 510bp *Sma*I-*Pst*I fragment prepared by digestion of PCR amplified DNA from the null allele of CEPH individual 134101 (see text, section 5.3.3). Minisatellite repeat units are shown in capitals, and the numbering of the filled allele sequence follows that of Figure 4.8. As in longer filled alleles, the first repeat unit of the null allele is atypical, with an extended region of $(AAY)_n$ repeats. Mismatched regions, which include a 23bp segment of an Alu element deleted from filled alleles (underlined), are shown in bold. The A at position 221 in the null allele is a variant also seen in repeats in filled alleles, and so this does not represent a significant difference; all the other mismatches, however, represent definite differences between the sequences of the null and filled alleles.

```
              .         .         .         .         .
   1 gggaggtggaggttacaatgagccaagatcgcgccactgcactccagcct 50
     ||||||||||||||||||||||||||||| |||||| |||||||||||||
2776 gggaggtggaggttacaatgagccaagattgcgccagtgcactccagcct 2825

              .         .         .         .         .
  51 gggacagagcaaagctctgtcacaaatgaTAATAATAATAATAATAATAATAACA 105
     ||||||                       |||||||||||||||||||   ||||
2826 gggaca.......................TAATAATAATAATAATAAT...AACA 2875

         .         .         .         .         .         .
 106 ACAACAACAACAACAACAAACAACAATAATAATAAT...GGGC.AGGCATAGGCATATTGCC 163
     ||||||||||||||||||| ||| ||||||||||||   |||| ||||||||||||||||||
2876 ACAACAACAACAACAACAA.CAATAATAATAATAATAACGGGCCAGGCATAGGCATATTGCC 2925

       .         .         .         .         .         .
 164 TGTAATCTCAGCACTTTCAACAACAACAACAATAATAATAATAACGGGCCAGGCATAGG. 223
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||
2926 TGTAATCTCAGCACTTTCAACAACAACAACAATAATAATAATAACGGGCCAGGCAT.GGC 2974

2975 ATATTGCCTGTAATCTCAGCACTTTCAACAACAACAATAATAATAAT 3024

                                  .         .         .
 224 .............................CATATTGCCTGTAATCTCAGC 243
                                  |||||||||||||||||||||
3025 AACGGGCCAGGCATAGGCATATTGCCTGTNNNNNGCCTGTAATCTCAGC 3074

              .         .         .         .         .
 244 ACTTTCAACAACAACAACAATAATAATAATAATGGGCCAGGCATAGGCAT 293
     ||||||||||||||||||||||||||||||||||||||||||||| ||||
3075 ACTTTCAACAACAACAACAATAATAATAATAATGGGCCAGGCATA.GCAT 3120
```

```
              .         .         .         .         .
 294 ATTGCCTGTAATCTCAGCACTTTcagaagccaaggagggaggattgcttg 343
     ||||||||||||||||||||||||||||||||||||||||||||||||||
3121 ATTGCCTGTAATCTCAGCACTTTcagaagccaaggagggaggattgcttg 3170

              .         .         .         .         .
 344 aggccaggagttcaagaccagcctaggcaacatagggagactctgtctct 393
     ||||||||||||||||||||||||||||||||||||||||||||||||||
3171 aggccaggagttcaagaccagcctaggcaacatagggagactctgtctct 3220

              .         .         .         .         .
 394 acaaaattttttttttaatttaaaaattaacaatgcatggtggcatgcac 443
     ||||||||||||||||||||||||||||||||||||||||||||||||||
3221 acaaaattttttttttaatttaaaaattaacaatgcatggtggcatgcac 3270

              .         .         .         .         .
 444 ctgtagacctacctactagggaggctaaggcagaaggctcacctaagccc 493
     ||||||||||||||||||||||||||||||||||||||||||||||||||
3271 ctgtagacctacctactagggaggctaaggcagaaggctcacctaagccc 3320

              .
 494 aggatttcaagctgcag 510
     |||||||||||||||||
3321 aggatttcaagctgcag 3338
```

Figure 5.15

Diagrammatic summary of sequence comparison between filled (above) and null (below) alleles at D12S40. Minisatellite repeats, including the first longer, atypical repeat, are shown as black arrows; sequences from the flanking Alu elements are shown as striped arrows. The additional repeats in the filled allele are shown "spliced out" to optimise alignment with the null allele. Regions of mismatch at the 5′ end are indicated: a 23bp region present in the null but absent from the filled alleles is indicated, and small regions of mismatch, including single base substitutions, are indicated by crosses.


filled
XX
XX
XX
23bp
XX
XX
XX
null

"ground state" from which a substantial block of tandem repeats has yet to be generated, or is the result of a mutation which has led to the loss of most or all repeats from a previously "filled" allele. In the latter model, the change could either be an extreme length change mutation, or a recombination between the two flanking Alu elements (Figure 4.9) deleting the minisatellite and some flanking non-reiterated DNA.

Since the minisatellite at D12S40 is immediately flanked on each side by an Alu dispersed repeat element, PCR amplification primers were designed using the DNA sequence from an extended, cosmid-derived clone from this locus (Figure 4.8 and legend).

In the amplification, cloning and sequencing of the null allele at D12S40 I am very pleased to acknowledge the technical assistance of Mrs.Moira Crosier. DNA from four members of the pedigree shown in Figure 3.5 (CEPH family 1341) was amplified under the conditions described in Table 2.2. The individuals were chosen so that any product corresponding to the null allele should be present in two, but absent from the other two. A single product of about 900bp was seen on ethidium staining, and segregated as expected for the null allele. Flanking DNA from the null allele of CEPH individual 134101 was removed by cleavage with *Pst*I and *Sma*I, which shortened the product to about 500bp as predicted from the known DNA sequence flanking filled alleles. This 500bp fragment was cloned into pBluescriptII vectors, and two independent clones in each orientation sequenced. This sequence, and the relation between the null allele sequence and that of a longer allele is presented in Figure 5.14, and is summarized diagrammatically in Figure 5.15.

The null allele appears to contain three repeat units; the first of these repeats, as in longer alleles, is atypical and contains an extended $(AAY)_n$ tract. The relation between the null allele and the longer alleles is, however, not simply one of repeat unit copy number. Both base substitutional and insertion/deletion differences are seen in the flanking DNA. These differences are very unlikely to be due to amplification or cloning artefacts, as all the null allele sequence was determined from two independent clones, and the "filled" allele sequence, although derived from a single cosmid clone, is confirmed at all the apparently discrepant positions by DNA sequence from the original Charomid clone (data not shown).

There are three single base substitutions and five single base insertions/deletions, as well as three larger discrepancies between the sequences at the 5′ end of the repeat array. The most striking of these is the presence in the null allele of 23 bases absent from the longer alleles. These 23 bases are part of the consensus Alu element sequence, but are missing from this position in filled alleles, suggesting that the state seen in the null allele may be ancestral to that in filled alleles. Furthermore, in the extended $(AAY)_n$ tract of the first repeat unit, the null allele has an extra AAT repeat, inserts an A to give an unusual AAAC repeat, and deletes an AAC repeat and a single base nearer the end of this extended $(AAY)_n$ tract.

The presence of three repeat units in the null allele suggests that it is not the result of an inter-Alu recombination, and may represent a ground state from which the filled alleles had expanded. However, the large number of differences between the null and filled allele flanking DNA suggests that the null allele has been independent of the filled for a considerable evolutionary time.

5.3.4 *Analysis at D17S134*

The minisatellite clone pMS228 contained two distinct polymorphic minisatellite regions (section 4.3.1.2). One of these, 228B, had the unusual combination of high variability (population heterozygosity 85%) and a restricted allele size range, 95% of alleles being smaller than 2kb (Figure 4.2b). This combination suggested that 228B would be an informative locus for analysis by PCR amplification, since unlike most highly variable minisatellites, at which many alleles in the population are too large (>10kb) to amplify, all the alleles at 228B should be small enough to amplify efficiently, at least to the point at which products can be detected by Southern blot hybridization.

Initial experiments were performed using a mixture of DNA samples from four individuals of known allele size at 228B. They were chosen to include eight alleles spanning the observed allele size range. The largest allele included (5.5kb) was the only allele larger than 2.5kb observed in a sample of 48 unrelated individuals (Figure 4.2b). Thus if this largest allele amplified efficiently, it would suggest that PCR amplification at 228B could be used to give a highly informative typing system from small amounts of starting material, and which could be relied upon to give a complete profile from nearly all individuals.

The results of analysis of this mixed DNA sample by PCR amplification and Southern blot hybridization (Figure 5.16a) show faithful amplification of all alleles, although rather inefficiently at the largest. Attempts to detect amplified

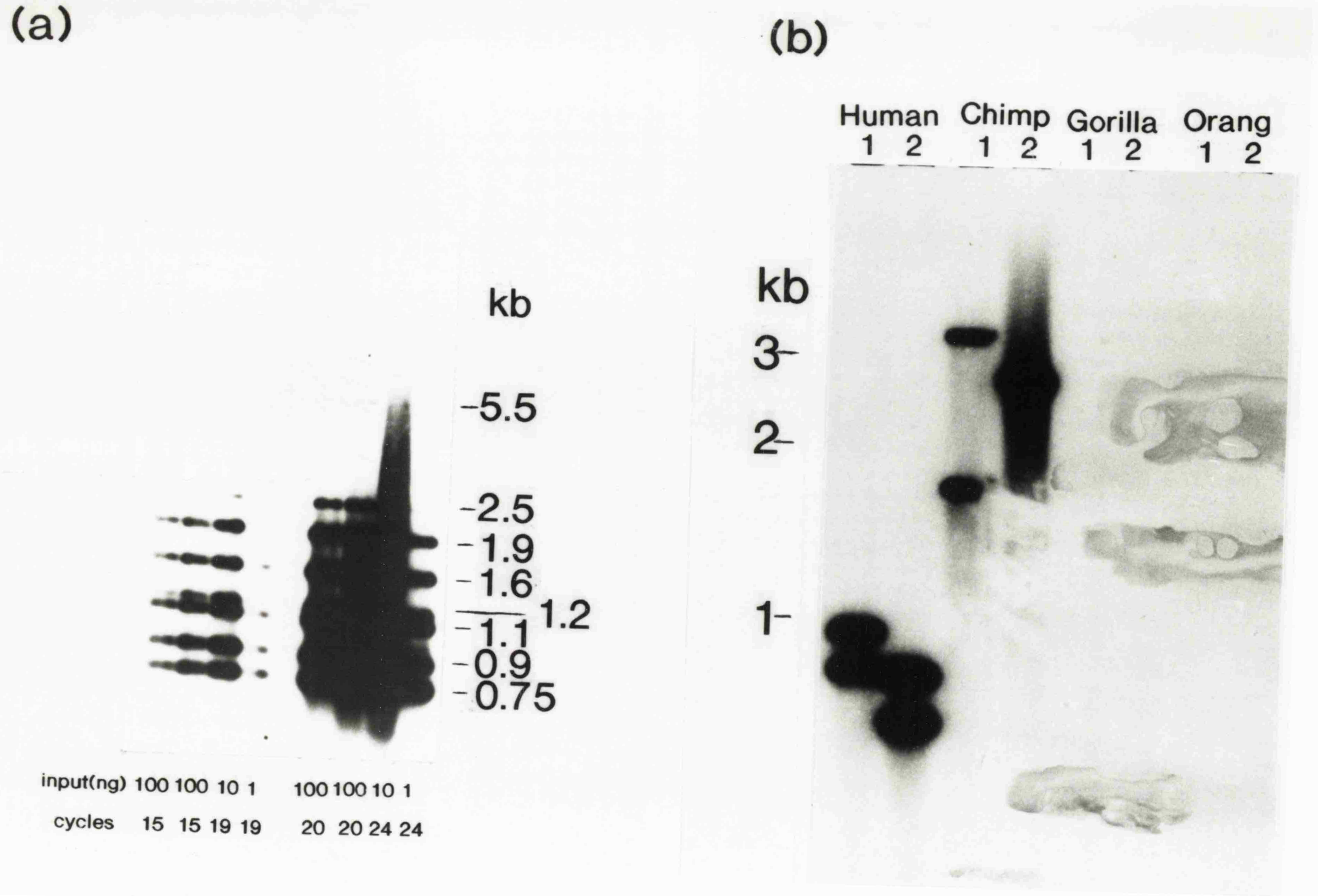
Figure 5.16

Analysis at the 228B minisatellite at D17S134 by PCR from human and primate DNA.

(a) shows the results of amplification of human DNA using a mixture of DNA from four unrelated individuals (CEPH individuals 2302, 133311, 134114 and 134510) to give a wide range (0.75-5.5kb) of target allele sizes. Primers 228BA and 228BC were used (Table 2.2). Alleles smaller than 2kb amplify most efficiently, although a signal can be obtained even from the 5.5kb allele from 100ng input DNA after 20 cycles. The "input" value is the amount of input DNA used *from each individual* in each reaction.

(b) shows the results of PCR analysis at this locus in two unrelated humans, chimpanzees, gorillas and orang-utans. 100ng input DNA was amplified using primers 228BA and 228BB for 15 cycles (Table 2.2). Of the non-human primates, only the chimpanzees show a discrete hybridizing product. There are at least three alleles amplified in the two chimpanzees (chimpanzee number 2 may be a homozygote for the product seen or a heterozygote for this allele and one too large to amplify).

PCR primers and cycle parameters are described in section 2.3.2.2. PCR products were detected by Southern blot hybridization using as probe a ≈700bp *Pvu*II-*Dra*I fragment from pMS228 bearing the 228B minisatellite (section 4.3.1.3)


(a)
kb
5.5
2.5
1.9
1.6
1.2
1.1
0.9
0.75
input(ng) 100 100 10 1 100 100 10 1
cycles 15 15 19 19 20 20 24 24
(b)
Human 1 2
Chimp 1 2
Gorilla 1 2
Orang 1 2
kb
3
2
1

products directly as ethidium stained bands on agarose gels were only partially successful; discrete amplification products could be obtained in large enough quantities, but these rapidly degenerated into a heterodisperse smear if further cycles of amplification were used. The difficulty in precise prediction of the number of cycles sufficient to give a visible product, complicated by the fact that smaller alleles will amplify more efficiently than large ones, makes direct visualization unsuitable for routine typing at 288B.

Attempts were also made to amplify DNA from this locus in a range of primate species. DNA samples from two unrelated humans, chimpanzees, gorillas and orang-utans were used as input for PCR amplification using the primers derived from the human DNA sequence (Figure 5.16b). Of the non-human species, only the chimpanzee DNA gave rise to a discrete amplified product, which appeared to be hypervariable in chimpanzees, with at least three different products in the two individuals studied. The sizes of the products, with at least two products larger than 2kb, suggest a difference in the allele size distributions between humans and chimpanzees at this locus ($p < 0.02$).

However, sequences at this locus failed to amplify from gorilla or orang-utan DNA; this may be due to loss of one or both amplimer binding sites by deletion or substitution, or because alleles at this locus in these species are too large to amplify efficiently. The absence of a specific amplified product from gorilla and orang-utan DNA precludes a conclusion about the evolutionary dynamics at this locus; without an outgroup for comparison it is not possible to infer whether the human state of hypervariability with a restricted allele size range is the ancestral or the derived state. In one model, the human-chimpanzee common ancestor would have had a human-like allele size distribution, to which the ancestral chimpanzee population added longer alleles by length change mutations after the lineages split. Another model would have a highly variable locus with a wide allele size range as the ancestral state, from which the sub-population ancestral to humans either drifted to small allele size or simply happened to include a high initial frequency of small alleles.

5.3.5 *Summary and perspectives*

This section has reported some initial investigations of evolutionary change at minisatellite loci. The analysis of the internal structure of minisatellite alleles present in modern populations has allowed conjectures to be made about at least some of the mutational events giving rise to them, and provide the most direct conceptual link with studies of germline mutation at minisatellite loci. The relations between the main groups of alleles at D22S163, however, cannot be inferred from structures held in common. This suggests an early origin for these main allele groups, possibly early enough to be studied in non-human primates.

The structure of the null allele at D12S40 is very different from that expected from simple contraction of the minisatellite array, and may testify to the antiquity of the split between null and longer tandemly repetitive alleles at this locus. It will be of great interest to analyse the structure at the cognate primate loci; if the null allele is a ground state, then the model predicts that many primate alleles will have a null structure, and if the split with filled alleles is indeed old, some filled alleles may also be detected in non-human primate DNA.

It is unfortunate that no outgroup could be assessed at the 228B (D17S134) minisatellite, and thus no inferences could be made about the evolution of allele size distributions at this locus. Minisatellite evolution has also been analysed by the determination of human and primate allele structure at D1S7 (by sequence analysis of amplified alleles) and of allele structure in primates and monkeys at D1S8 (Ian Gray and Alec Jeffreys, manuscript in preparation). These latter studies have demonstrated the extreme evolutionary transience of those minisatellites most variable in human populations; these loci are short, monomorphic arrays in non-human primates. By contrast, the loci studied in the work presented above may provide examples of less unstable minisatellites, at which alleles may survive for millions of years.

CHAPTER 6

GENERAL DISCUSSION AND SUMMING-UP

What I tell you three times is true.

Lewis Carroll

Much of what will be discussed in this final chapter, especially what concerns the relation between the structure of minisatellites and their general properties, could be appropriately dealt with under more than one of the three main headings. However, in order to preserve the general organization of the main body of this work, this discussion is divided, sometimes rather arbitrarily, into corresponding sections, concerned with the isolation of human minisatellites, their place in the genome, and evolutionary and mutational change.

6.1 *ISOLATING HUMAN MINISATELLITES*

The ordered array Charomid library described in chapter 3 appears to have provided a very useful tool for the isolation of human minisatellites, not least because of its practical simplicity and its yield of "free" information on the overlap between the loci detected by different multi-locus DNA fingerprinting probes. However, we are not yet in a position to predict how much useful service is left in the system; it may be that most loci which can be cloned from this library already have been, and we will soon enter a phase of diminishing returns as encountered with λ phage cloning. If isolation by cloning gets "stuck" again, what options are open?

In addition to the possible use of different *E.coli* host strains, an entirely unexplored avenue of enquiry stems from a consideration of sequence copy number. If the repeat unit of a minisatellite bears some similarity to an important regulatory unit in the *E.coli* genome, then the presence in the cell of a high copy number vector bearing tandem arrays of such sequences would be predicted to have a deleterious effect on cellular physiology. Indeed, the sequestration of recBC protein by large numbers of *chi*-like elements may be a contributing factor to the very poor growth of minisatellite phage recombinants (Wong *et al.*,1986,1987). One way to alleviate such effects would be to use a vector stably maintained at low copy number, but with a relaxed origin which may be induced when high yields of DNA are required. The recently developed P1 cloning system, for example, incorporates a single-copy P1 plasmid replicon with a runaway lytic replicon under the control of the *lac* operon (Sternberg,1990).

The polymorphic loci cloned from the Charomid library are extremely useful in a wide variety of human genetic analyses. What may be less obvious is that the monomorphic minisatellites isolated may also be of practical utility. Much has been written recently about the importance of quality control in the forensic use of minisatellite probes (Lander,1989). One recurring theme is the appearance of unpredictable "band-shifts", in which a DNA sample, usually a critical forensic specimen, migrates anomalously on agarose gel electrophoresis. One simple control for evenness of electrophoretic migration would be to re-probe the Southern blot filters with a probe known to recognise a monomorphic locus, such that any variation in migration must be due to "band-shift" effects. The practical problem has been to find a probe which recognizes a monomorphic locus in the relevant size-range, say 4-6kb *Hin*fI fragments. A monomorphic probe isolated from the Charomid library recognizes a monomorphic *Hin*fI fragment of 4.2kb, and is currently under investigation by Cellmark Diagnostics as a control probe for gel migration; initial results suggest that this probe recognizes a large minisatellite locus which is nevertheless truly invariant in a large population survey.

The Charomid library is also a source of information on the degree to which multi-locus DNA fingerprinting probes overlap, at least within the set of loci represented in the library. This in turn bears on the question of whether the variability at minisatellite loci is attributable simply to the propensity of tandemly-repeated structures to undergo length change mutation (Smith,1976), or whether sequence elements in the repeat units are at least partially responsible for *promoting* variability. In short, is the "core" real?

Inspection of a DNA fingerprint profile shows at once that tandem repeat instability alone will not do. The vast majority of loci detected on a DNA fingerprint are relatively short, invariant minisatellites, while nearly all the larger (>3kb) loci are extremely variable. Similarly, a recent survey of microsatellite loci (Weber,1990) shows that there is a strong correlation between the length of a dinucleotide repeat array and its informativeness. One might then propose that variability is an intrinsic property of tandem arrays *above a certain size*, irrespective of repeat unit sequence.

However, the idea of sequence dependence of variability cannot be entirely discarded, for the following reasons. Firstly, while many probes have now been defined which recognize multiple polymorphic loci in human DNA, the best are all of similar sequence structure, and include a G-rich "core"-like sequence. Secondly, some very short minisatellites are nevertheless highly variable, for example YNZ22 (Wolff *et al.*,1988) and MS228B (section 4.3.1.2). Thirdly, the difference in variability between minisatellites with similar allele size distributions, such as MS8 and MS31 (Wong *et al.*,1987), is not just a "frozen accident"; the two loci also have very different mutation rates. Thus MS31 does not just happen to be more variable than MS8 in modern human populations, but the difference is still being maintained by a higher current mutation rate to new length alleles. Fourthly, but with the proviso that the length of their *tandem repeated* portion is not yet known, the existence of long, monomorphic minisatellites would suggest that variability cannot be a simple function of allele length. Fifthly, while not yet fully characterised, the detection of a specific minisatellite binding protein (Collick and Jeffreys,1990) in eukaryotic nuclei also suggests roles for specific sequence motifs in minisatellite arrays.

In summary, then, and always bearing in mind that the properties of any one minisatellite array may to some extent be dependent upon its local genomic context (*v.i.*), minisatellite array length appears to be of some importance in determining mutation and variation, but it is equally clear that some tandem repeated sequences undergo more frequent mutation, and are thus more "intrinsically variable", than others.

6.2 *THEIR PLACE IN THE GENOME*

The studies presented in chapter 4 have extended the evidence for preferential clustering of minisatellite loci in subtelomeric locations, and have shown that minisatellite flanking sequences appear to contain dispersed repeat elements at a significantly elevated frequency. These correlations raise a number of questions, which are sometimes lumped together. In the absence of further evidence, however, an approach may at least be made by posing them separately. Firstly, were the subtelomeric location and/or nearby dispersed repeat elements instrumental in the initial generation of a tandem repeated array? Secondly, do minisatellites promote the apparently enhanced rate of recombination seen near the ends of chromosomes? Thirdly, are the dispersed repeats or subtelomeric location involved in the current maintenance of polymorphism by new mutation? The question of whether unequal recombination is involved in minisatellite mutation, which may be an entirely different matter, is discussed in section 6.3.

The answer to all three questions asked above is presently very simple, that very little evidence is available. On the second point, *in vitro* assays for recombination suggest that minisatellite repeats may indeed promote recombination (Wahls *et al.*,1990), although evidence for enhancement of recombination *in vivo* requires a different level of analysis. It may nevertheless be useful to point out that our attention has justifiably been centred almost exclusively on *polymorphic* minisatellites, almost entirely ignoring the probably very numerous monomorphic arrays. Thus our current impression that tandemly repeated minisatellites are prone to arise near chromosome ends may simply be the result of a higher frequency of polymorphism in these subtelomeric arrays.

6.3 *EVOLUTIONARY AND MUTATIONAL CHANGE*

Comparisons of minisatellite loci between species (Gray and Jeffreys, manuscript in preparation) have shown that the minisatellites most variable in human populations are extremely transient in evolution, such that the cognate locus in non-human primates is frequently short and monomorphic. However, the work presented in chapter 5 suggests that less rapidly evolving loci may be compared both within and between species, mainly in terms of allele frequency distributions. Because minisatellite loci can produce DNA fingerprints by cross-hybridizing to other loci of similar sequence, PCR amplification has been found to be the only reliable method for identifying the cognate locus in non-human species (Gray and Jeffreys, manuscript in preparation). This places two limitations on such comparisons; alleles at some loci may be too large in some species to amplify efficiently, and the cognate locus may fail to amplify from non-human DNA for other reasons, such as a small deletion at an amplimer binding site.

Minisatellite loci at which "null" alleles are seen may be of particular interest in evolutionary analysis. In addition to the D12S40 locus described in chapter 5, a second example of a locus with a "null" allele has recently been isolated from the Charomid library. At this new locus the null allele appears to have a population frequency of about 90%, such that most individuals in the population have no visible alleles at all. Nevertheless, those alleles which can be detected are hypervariable in length. These loci may also be of some interest to those concerned with validation of DNA fingerprinting in civil and forensic casework. One theoretical criticism levelled at DNA fingerprinting is that minisatellite probes detect an apparent excess of homozygotes, taken to be indicative of inbreeding within sub-population structures not taken into account by statistical analyses (Cohen,1990). However, the existence of an unsuspected null allele in a system would elevate the frequency of *apparent* homozygotes, and provide a simple explanation for some of the observed data, without invoking departure from Hardy-Weinberg equilibrium.

The process of germline mutation is the ultimate source of change in minisatellite evolution. The current evidence available on mutational mechanisms at minisatellite loci consists of detailed analysis of a single event at YNZ22 (Wolff *et al.*,1988), and of a range of mutations at D1S7 (Wolff *et al.*,1989) and D1S8 (Jeffreys *et al.*,1990a). None of these studies implicate interchromosomal recombination in the mutation process. However, the two loci for which we have detailed analysis are both interstitial, and a different pattern of mutational mechanisms may prevail at the more numerous subtelomeric loci.

The analysis of somatic mutation at minisatellite loci complements analyses of germline mutation and promises a number of interesting extensions. The "ectopic" appearance of MS1 minisatellite repeats in tumour DNA from patient JB may be an example of a DNA-mediated minisatellite transposition event. Evidence for another, much larger event has already been presented (Wong *et al.*,1990), and such transposition events could in theory be responsible for the generation of at least some new minisatellite loci.

Combination of MVR mapping with PCR amplification of low-level mutation events (Jeffreys *et al.*,1990a) may be of particular interest in the analysis of tissue from Bloom's syndrome patients. Cells from such patients have a markedly elevated level of sister chromatid exchange (German *et al.*,1965), and if this resulted in a similarly raised frequency of *unequal* sister chromatid exchanges, then one would predict a higher mutation rate in Bloom's syndrome. This is one of the few ways in which evidence could be adduced to distinguish mutational mechanisms involving replication slippage or unequal sister chromatid exchange, neither of which would result in the exchange of flanking markers.

REFERENCES

Albert,E.D., Baur,M.P. and Mayr,W.R. (Eds.) Histocompatability testing 1984, Springer-Verlag.

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

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

Armour,J.A.L., Povey,S., Jeremiah,S. and Jeffreys,A.J. (1990). Systematic cloning of human minisatellites from ordered array Charomid libraries. *Genomics* (in the press).

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

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

Baker,S.J., Fearon,E.R., Nigro,J.M., Hamilton,S.R., Preisinger,A.C., Jessup,J.M., van Tuinen,P., Ledbetter,D.H., Barker,D.F., Nakamura,Y., White,R. and Vogelstein,B. (1989). Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science* **244**,217-221.

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

Birnboim,H.C. and Doly,J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* 7,1513-1523.

Blattner,F.R., Williams,B.G., Blechl,A.E., Denniston-Thompson,K., Faber,H.E., Furlong,L.-A., Grunwald,D.J., Kiefer,D.O., Moore,D.D., Schumm,J.W., Sheldon,E.O. and Smithies,O. (1977). Charon phages: safer derivatives of bacteriophage lambda for DNA cloning. *Science* 196,161-169.

Bodmer,W.F., Bailey,C.J., Bodmer,J., Bussey,H.J.R., Ellis,A., Gorman,P., Lucibello,F.C., Murday,V.A., Rider,S.H., Scambler,P., Sheer,D., Solomon,E. and Spurr,N.K. (1987). Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature* 328,614-616.

Braman,J., Barker,D., Schumm,J., Knowlton,R. and Donis-Keller,H. (1985). Characterization of very highly polymorphic RFLP probes. (HGM8 abstract). *Cytogenet.Cell Genet.* 40,589.

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

BRL focus (1986). Tools for the molecular biologist. *BRL Focus* 8(2),8.

Brownstein,B.H., Silverman,G.A., Little,R.D., Burke,D.T., Korsmeyer,S.J., Schlessinger,D. and Olson,M.V. (1989). Isolation of single-copy human genes from a library of yeast artificial chromosome clones. *Science* 244,1348-1351.

Buluwela,L., Forster,A., Boehm,T. and Rabbitts,T.H. (1989). A rapid procedure for colony screening using nylon filters. *Nucleic Acids Res.* 17,452.

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

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

Chimini,G., Mattei,M.-G., Passage,E., Nguyen,C., Boretto,J., Mattei,J.-F. and Jordan,B.R. (1989). *In situ* hybridization and pulsed-field gel analysis define two major minisatellite loci: 1q23 for minisatellite 33.6 and 7q35-q36 for minisatellite 33.15. *Genomics* 5,316-324.

Church,G.M. and Gilbert,W. (1984). Genomic sequencing. *Proc.Nat.Acad.Sci.U.S.A.* 81,1991-1995.

Clarke,M.F., Westin,E., Schmidt,D., Josephs,S.F., Ratner,L., Wong-Staal,F., Gallo,R.C. and Reitz,M.S.Jr. (1984). Transformation of NIH 3T3 cells by a human *c*-sis cDNA clone. *Nature* 308,464-467.

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.

Colb,M., Yang-Feng,T., Francke,U., Mermer,B., Parkinson,D.R. and Krontiris,T.G. (1986). A variable number tandem repeat locus mapped to chromosome band 10q26 is amplified and rearranged in leucocyte DNAs of two cancer patients. *Nucleic Acids Res.* 14,7929.

Collick,A. and Jeffreys,A.J. (1990). Detection of a novel minisatellite-specific DNA-binding protein. *Nucleic Acids Res.* 18,625-629.

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

Cooper,D.N., Smith,B.A., Cooke,H.J., Niemann,S. and Schmidtke,J. (1985). An estimate of unique sequence heterozygosity in the human genome. *Hum.Genet.* 69,201-205.

Coulson,A., Sulston,J., Brenner,S. and Karn,J. (1986). Toward a physical map of the nematode *Caenorhabditis elegans*. *Proc.Nat.Acad.Sci.U.S.A.* 83,7821-7825.

Cross,S.H. and Little,P.F.R. (1986). A cosmid vector for systematic chromosome walking. *Gene* 49,9-22.

Demers,G.W., Brech,K. and Hardison,R.C. (1986). Long L1 interspersed repeats in rabbit DNA are homologous to L1 repeats of rodents and primates in an open-reading-frame region. *Mol.Biol.Evol.* 3,179-190.

Drayna,D., Davies,K., Hartley,D., Mandel,J.L., Camerino,G., Williamson,R. and White,R. (1984). Genetic mapping of the human X chromosome by using restriction fragment length polymorphisms. *Proc.Nat.Acad.Sci.U.S.A.* 81,2836-2839.

Donehower,L.A., Slagle,B.L., Wilde,M., Darlington,G. and Butel,J.S. (1989). Identification of a conserved sequence in the non-coding regions of many human genes. *Nucleic Acids Res.* 17,699-710.

Donis-Keller,H. and 31 others (1987). A genetic linkage map of the human genome. *Cell* 51,319-337.

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

Dower,W.J., Miller,J.F. and Ragsdale,C.W. (1988). High efficiency transformation of *E.coli* by high voltage electroporation. *Nucleic Acids Res.* 16,6127-6145.

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

Fearon,E.R., Hamilton,S.R. and Vogelstein,B. (1987). Clonal analysis of human colorectal tumours. *Science* 238,193-197.

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

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

Friezner Diegen,S.J., Rajput,B. and Reich,E. (1986). The human tissue plasminogen activator gene. *J.Biol.Chem.* 261,6972-6985.

Fung,Y.-K.T., Murphree,A.L., T'Ang,A., Qian,J., Hinrichs,S.H. and Benedict,W.F. (1987). Structural evidence for the authenticity of the human retinoblastoma gene. *Science* 236,1657-1661.

Georges,M., Lathrop,M., Hilbert,P., Marcotte,A., Schwers,A., Swillens,S., Vassart,G. and Hanset,R. (1990). On the use of DNA fingerprints for linkage studies in cattle. *Genomics* 6,461-474.

German,J., Archibald,R. and Bloom,D. (1965). Chromosomal breakage in a rare and probably genetically determined syndrome of man. *Science* 148,506-507.

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

Gillen,J.R., Willis,D.K. and Clark,A.J. (1981). Genetic analysis of the RecE pathway of genetic recombination in *Escherichia coli* K-12. *J.Bacteriol.* 145,521-532.

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

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

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

Hanauer,A. and Mandel,J.L. (1984). The glyceraldehyde 3 phosphate dehydrogenase gene family: structure of a human cDNA and of an X linked pseudogene; amazing complexity of the gene family in mouse. *EMBO J.* 3,2627-2633.

Hanotte,O., Burke,T., Armour,J.A.L. and Jeffreys,A.J. (1990). Hypervariable minisatellite DNA sequences in the peafowl *Pavo cristatus*. Manuscript submitted to *Genomics*.

Hansen,M.F. and Cavenee,W.K. (1987). Genetics of cancer predisposition. *Cancer Res.* 47,5518-5527.

Henikoff,S. (1984). Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* 28,351-359.

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

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

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

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

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

Jeffreys,A.J. and Flavell,R.A. (1977). The rabbit β-globin gene contains a large insert in the coding sequence. *Cell* 12,1097-1108.

Jeffreys,A.J. (1979). DNA sequence variants in the Gγ-, Aγ- δ-

and β-globin genes of man. *Cell* 18,1-10.

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

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

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

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

Jeffreys,A.J. (1987). 23rd. Colworth medal lecture: Highly variable minisatellites and DNA fingerprints. *Biochem.Soc.Trans.* 15,309-317.

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

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

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

Jeffreys,A.J., MacLeod,A., Neumann,R., Povey,S. and Royle,N.J. (1990b). "Major minisatellite loci" detected by minisatellite clones 33.6 and 33.15 correspond to the cognate loci D1S111 and

D7S437. *Genomics* 7,449-452.

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

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

Kielty,C.M., Povey,S. and 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.* 46,307-327.

Kimura,M. (1983). The neutral theory of molecular evolution. Cambridge University Press.

King,M.C. (1982). Genetic and epidemiological analysis of cancer in families: breast cancer as an example. *Cancer Surveys* 1,33-46.

Kinzler,K.W. and Vogelstein,B. (1989). Whole genome PCR: application to the identification of sequences bound by gene regulatory proteins. *Nucleic Acids Res.* 17,3645-3653.

Knott,T.J., Wallis,S.C., Pease,R.J., Powell,L.M. and Scott,J. (1986). A hypervariable region 3' to the human apolipoprotein B gene. *Nucleic Acids Res.* 14,9215-9216.

Knudson,A.G. (1971). Mutation and cancer: statistical study of retinblastoma. *Proc.Nat.Acad.Sci.U.S.A.* 68,820-823.

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

Lee,W.-H., Bookstein,R., Hong,F., Young,L.-H., Shew,J.-Y. and Lee,E.Y.H.P. (1987). Human retinoblastoma susceptibility gene: cloning, identification and sequence. *Science* 235,1394-1399.

Lehmann,H. and Kynoch,P.A.M. (1976). Human haemoglobin variants and their characteristics. North Holland,Amsterdam.

Leppert,M., Cavenee,W., Callahan,P., Holm,T., O'Connell,P., Thompson,K., Lathrop,G.M., Lalouel,J.-M. and White,R. (1986). A primary genetic linkage map of chromosome 13q. *Am.J.Hum.Genet.* 39,425-427.

Li,H., Gyllensten,U.B., Cui,X., Saiki,R.K., Erlich,H.A. and Arnheim,N. (1988). Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature* 335,414-417.

Lipkin,M., Sherlock,P. and Bell,B. (1963). Cell proliferation kinetics in the gastrointestinal tract of man. *Gastroenterology* 45,721-729.

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

Lloyd,R.G. and Buckman,C. (1985). Identification and genetic analysis of *sbcC* mutations in commonly used *recBCsbcB* strains of *Escherichia coli* K-12. *J.Bacteriol.* 164,836-844.

Loenen,W.A.M. and Blattner,F.R. (1983). Lambda Charon vectors (Ch32, 33, 34 and 35) adapted for DNA cloning in recombination-deficient hosts. *Gene* 26,171-179.

Loenen,W.A.M. and Brammar,W.J. (1980). A bacteriophage lambda vector for cloning large fragments made with several restriction enzymes. *Gene* 20,249-259.

Lundberg,C., Skoog,L., Cavenee,W.K. and Nordenskjold,M. (1987). Loss of heterozygosity in human ductal breast tumours indicates a recessive mutation on chromosome 13. *Proc.Nat.Acad.Sci.U.S.A.* 84,2372-2376.

Mackay,J., Elder,P.A., Steel,C.M., Forrest,A.P.M. and Evans,H.J. (1987). Allele loss on short arm of chromosome 17 in breast cancers. *Lancet* ii,1384-1385.

Maeda,N. (1985). Nucleotide sequence of the haptoglobin and haptoglobin-related gene pair. *J.Biol.Chem.* 260,6698-6709.

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

Mathew,C.G.P., Chin,K.S., Easton,D.F., Thorpe,K., Carter,C., Liou,G.I., Fong,S.-L., Bridges,C.D.B., Haak,H., Nieuwenhuijzen Kruseman,A.C., Schifter,S., Hansen,H.H., Telenius,H., Telenius-Berg,M. and Ponder,B.A.J. (1987b). A linked genetic marker for multiple endocrine neoplasia type 2A on chromosome 10. *Nature* 328,527-528.

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

Messing,J., Crea,R. and Seeburg,P.H. (1981). A system for shotgun DNA sequencing. *Nucleic Acids Res.* 9,309-321.

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

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

Morton,N.E. (1962). Segregation and linkage. In: Burdette,W.J. (Ed.) Methodology in human genetics, pp.17-52. Holden Day, San Francisco.

Nabholz,M., Miggiano,V. and Bodmer,W.F. (1969). Genetic analysis with human-mouse somatic cell hybrids. *Nature* 223,358-363.

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

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

Nakamura,Y., Lathrop,M., O'Connell,P.,Leppert,M., Lalouel,J.-M. and White,R. (1988a). A primary map of ten DNA markers and two serological markers for human chromosome 19. *Genomics* 3,67-71.

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

Naom,I.S., Morton,S.J., Leach,D.F.R. and Lloyd,R.G. (1989). Molecular organization of *sbcC*, a gene that affects genetic recombination and the viability of DNA palindromes in *Escherichia coli* K-12. *Nucleic Acids Res.* 17,8033-8045.

Newman,B., Austin,M.A., Lee,M. and King,M.-C. (1988). Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. *Proc.Nat.Acad.Sci.U.S.A.* 85,3044-3048.

Nicholls,R.D., Hill,A.V.S., Clegg,J.B. and Higgs,D.R. (1985). Direct cloning of specific genomic DNA sequences in plasmid libraries following fragment enrichment. *Nucleic Acids Res.* 13,7569-7578.

O'Connell,P., Lathrop,G.M., Law,M., Leppert,M., Nakamura,Y., Hoff,M., Kumlin,E., Thomas,W., Elsner,T., Ballard,L., Goodman,P., Azen,E., Sadler,J.E., Cai,G.Y., Lalouel,J.-M. and White,R. (1987). A primary genetic linkage map for human chromosome 12. *Genomics* 1,93-102.

O'Connell,P., Lathrop,G.M., Leppert,M., Nakamura,Y., Muller,U., Lalouel,J.-M. and White,R. (1988). Twelve loci form a continuous linkage map for human chromosome 18. *Genomics*

3,367-372.

Orgel,L.E. (1963). The maintenance of the accuracy of protein synthesis and its relevance to ageing. *Proc.Nat.Acad.Sci.U.S.A.* 49,517-521.

Owainiti,A.A.R., R.A.Robins, Hinton,C., Ellis,I.O., Dowle,C.S., Ferry,B., Elston,C.W., Blamey,R.W. and Baldwin,R.W. (1987). Tumour aneuploidy, prognostic parameters and survival in primary breast cancer. *Br.J.Cancer* 55,449-454.

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

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

Petit,C., Levilliers,J. and Weissenbach,J. (1988). Physical mapping of the human pseudo-autosomal region; comparison with the genetic linkage map. *EMBO J.* 7,2369-2376.

Potten,C.S. and Loeffler,M. (1987). A comprehensive model of the crypts of the small intestine of the mouse provides insight into the mechanisms of cell migration and the proliferative hierarchy. *J.Theor.Biol.* 127,381-391.

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

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

Raleigh,E.A., Murray,N.E., Revel,H., Blumental,R.M., Westaway,D., Reith,A.D., Rigby,P.W.J., Elhai,J. and Hanahan,D. (1988). McrA and McrB restriction phenotypes of some *E.coli* strains and implications for gene cloning. *Nucleic Acids Res.* 16,1563-1575.

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

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

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

Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989). Molecular cloning, a laboratory manual. 2nd.edition. Cold Spring Harbour Laboratory Press.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc.Nat.Acad.Sci.U.S.A.* 74,5463-5467.

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

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

Schumm,J., Knowlton,R., Braman,J., Barker,D., Vovis,G., Akots,G., Brown,V., Gravius,T., Helms,C., Hsiao,K., Rediker,K., Thurston,J., Botstein,D. and Donis-Keller,H. (1985). Detection of more than 500 single-copy RFLPs by random screening. (HGM8 abstract) *Cytogenet. Cell Genet.* 40,739.

Singer,M.F. and Skowronski,J. (1985). Making sense out of LINES: long interspersed repeat sequences in mammalian genomes. *Trends Biochem.Sci.* 10,119-122.

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

Smith,H.O. and Birnstiel,M.L. (1976). A simple method for DNA restriction site mapping. *Nucleic Acids Res.* 3,2387-2398.

Solomon,E., Bobrow,M., Goodfellow,P.N., Bodmer,W.F., Swallow,D.M., Povey,S. and Noel,B. (1976). Human gene mapping using an X/autosome translocation. *Somatic Cell Genet.* 2,125-140.

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

Southern,E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J.Mol.Biol.* 98,503-517.

Sternberg,N. (1990). Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. *Proc.Nat.Acad.Sci.U.S.A.* 87,103-107.

Stoker,N.G., Cheah,K.S.E., Griffin,J.R. and Solomon,E. (1985). A highly polymorphic region 3′ to the human type II collagen gene. *Nucleic Acids Res.* 13,4613-4622.

Suarez,B.K., Rice,J. and Reich,T. (1978). The generalized sib pair IBD distribution: its use in the detection of linkage. *Ann.Hum.Genet.* 42,87-94.

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

Tabor,S. and Richardson,C.C. (1987). DNA sequence analysis with a modifed bacteriophage T7 DNA polymerase.

Proc.Nat.Acad.Sci.U.S.A. 84,4767-4771.

Theillet,C., Lidereau,R., Escot,C., Hutzell,P., Brunet,M., Gest,J., Schlom,J. and Callahan,R. (1986). Loss of a c-H-*ras*-1 allele and aggressive human primary breast carcinomas. *Cancer Res.* 46,4776-4781.

Thein,S.L., Jeffreys,A.J., Gooi,H.C., Cotter,F., Flint,J., O'Connor,N.J.T. and Wainscoat,J.S. (1987). Detection of somatic changes in human cancer DNA by DNA fingerprint analysis. *Br.J.Cancer* 55,353-356.

Uitterlinden,A.G., Slagboom,E., Knook,D.L. and Vijg,J. (1989). Two-dimensional DNA fingerprinting of human individuals. *Proc.Nat.Acad.Sci.U.S.A.* 86,2742-2746.

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

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

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

Vieira,J. and Messing,J. (1982). The pUC plasmids: an M13mp7-derived system for insertion inactivation and sequencing with synthetic universal primers. *Gene* 19,259-268.

Vogel,F. (1964). Preliminary estimate of the number of human genes. *Nature* 201,847.

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

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

Washio,K, Misawa,S. and Ueda,S. (1989). Probe walking: development of novel probes for DNA fingerprinting. *Hum.Genet.* **83**,223-226.

Watson,J.D. (1972). Origins of concatemeric T7 DNA. *Nature (New Biology)* **239**,197-201.

Waye,J.S. and Willard,H.F. (1986). Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res.* **15**,7549-7568.

Weatherall,D.J. and Clegg,J.B. (1976). Molecular genetics of human haemoglobin. *Ann.Rev.Genet.* **10**,157-178.

Weber,J.L. (1990). Informativeness of human $(dA-dC)_n.(dG-dT)_n$ polymorphisms. *Genomics* **7**,524-530.

Weber,J.L. and May,P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am.J.Hum.Genet.* **44**,388-396.

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

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

Wolff,R.K., Plaetke,R., Jeffreys,A.J. and White,R. (1989). Unequal crossingover between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci. *Genomics* **5**,382-394.

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

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

Wong,Z., Royle,N.J. and Jeffreys,A.J. (1990). A novel human DNA polymorphism resulting from transfer of DNA from chromosome 6 to chromosome 16. *Genomics* 7,222-234.

Wyman,A.R. and White,R. (1980). A highly polymorphic locus in human DNA. *Proc.Nat.Acad.Sci.U.S.A.* 77,6754-6758.

Wyman,A.R., Wolfe,L.B. and Botstein,D. (1985). Propagation of some human DNA sequences in bacteriophage λ vectors requires mutant *Escherichia coli* hosts. *Proc.Nat.Acad.Sci.U.S.A.* 82,2880-2884.

Yanisch-Perron,C., Vieira,J. and 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. and Birnie,G.D. (1976). A new estimate of human ribosomal gene number. *Biochim.Biophys.Acta* 454,539-548.