THE INVESTIGATION OF VARIATION AND *DE NOVO* MUTATION IN THE HUMAN GENOME

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Caroline Ruth Hollies

Department of Genetics

University of Leicester

June 1999

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

Contents

Abstract

Hypervariable minisatellites form a subset of tandem repeat arrays which show high rates of germline instability. At the hypervariable minisatellite MS31a, population studies indicated that there was no polarity of variability within the tandem repeat array. This contrasts with previous analysis of pedigree mutant alleles which demonstrated that mutation is polarised towards the 5′ end of the repeat array. Population studies on the flanking DNA also demonstrated high levels of recombination throughout the MS31a locus. It is suggested that the absence of any polarity of variability within the repeat array may be due to these elevated levels of recombination flanking the minisatellite. Further evidence to support this has arisen from studies to isolate and characterise crossover molecules from sperm DNA. This work identified a putative recombination hotspot in the 5′ flanking DNA adjacent to the most unstable part of the tandem repeat array. This correlates well with similar studies on the human minisatellite MS32, indicating that there may be a conserved mechanism of mutation at these minisatellites.

The second half of the project investigates instability of Alu elements. Alus are dispersed repeats which have been associated with recombination breakpoints in a significant number of genetic diseases. Two Alu rich regions were examined; the first is located upstream of the human hypervariable minisatellite MS32. The second is a region of the C1 inhibitor gene that has previously been associated with Alu-mediated mutation. Population and evolutionary studies predicted that rearrangement at these loci was very rare. Concentrating on the detection of deletion events in sperm DNA, the most sensitive techniques available at this time were unable to recover any mutant alleles at either locus. The implications are that mutation detection techniques must be improved before the molecular mechanisms of low level mutation can be elucidated.

Acknowledgements

Firstly, I would like to thank Alec for taking me on and putting up with me for the duration of my time in Leicester, and for all his necessary time and patience. I would also like to mention those other members of genetics without whom I could never have conquered these windmills. Most of all Celia who made things make sense, and who never seemed to mind when I contaminated her bench.....only jesting! Also to Dave Neil who introduced me to the wonders of MS31a. To the French Connection and Keiji for lots of help with that scientific stuff, and for being infinitely tease-able (especially Jerome's shirts and Keiji's moustache - or was that Saki's?!). To Yuri for endless dialogue on the weather, statistics and, of course, beer. To Ruth, Maria and Colin for gossip, moans, cups of tea and chocolate. To John Stead for never going home. Hooge amounts of appreciation go to all the technical staff for their help and support - Ila (how can I cope without you?). I would also like to thank everyone with whom I've climbed, drank, talked science, raved with, talked motorbikes. Particularly HelenV., Wendy, Vanessa and Gemma for living with me; JM for hacking my ankles to death at Unihoc and for shouting; Jeni, Noel, Emma, Glen, Tony, Zoe, Matt and all the PhD students who suffered with me. LUMC must also have a mention for taking my mind off things by scaring me stupid. For Linus and his gear, Matt and TP for "doing" Asia; all those that braved Magalluf and Benidorm for the grand climbing that was to be had (honest!). For an introduction to Scottish winter mountaineering (i.e. cold wet and sheep) - James, Kim, and Chris Abrams. To our illustrious leaders, Noel and Aiders that made everything run smoothly, not withstanding their dreadful gags and lunchtime crosswords. To Steph (she of the broken ankle) and that crowd for dragging me out and feeding me food, beer and cocktails (yes, at the same time!). To Leicester University "ladies" rugby team without whom I would have gone mad. To all those canoodlers that just won't let go - Karen W., Pigeon, George, Jackie, Pete Jeep, and Simon. Thanks also to everyone in Cardiff who made me welcome, especially DK and Duncan who ploughed through bits of my thesis. I would also like to give a very big thank you to Lisa Bostick who can always be relied on for food, festivals and walking in the road. Most thanks go to my parents, who have always driven me mad with talk of Scouts, boats, cows and potash, but who have always been fantastic and supported me throughout. Mention also to my brothers for being in different countries for most of the duration of this work, cheers guys! Finally I would like to say a very large thank you, thank you, thank you to Ali who has had to put up with it all. And is still alive!

Abbreviations

ASO	allele-specific oligonucleotide
C1NH	C1-inhibitor gene
CEPH	Centre d'Etude du Polymorphisme Humain
d.f.	degrees of freedom
DSB	double strand break
HAE/HANE	hereditary angioedema
HERV	human endogenous retrovirus
indel	insertion/deletion site
LINE	long interspersed nuclear element
mb, kb, bp	mega-, kilo- base pair
µl, ml, l	micro-, milli- litre
pM, µM, mM, M	pico-, micro-, milli- molar
MVR	minisatellite variant repeat
nt	nucleotide
PCR	polymerase chain reaction
pers. comm.	personal communication
RFLP	restriction fragment length polymorphism
SINE	short interspersed nuclear element
SNP(s)	single nucleotide polymorphism(s)
SP-PCR	small pool-PCR
SRP	signal recognition pathway

Chapter 1

Introduction

Variation must exist in order for evolution to occur, and evolution is essential for the origin and survival of species. Variation arises by mutation and can be shuffled by recombination to give increasing levels of diversity. The processes of random genetic drift, migration, mutation, selection and reduction (or founder effect) then manipulate these variants allowing existing species to adapt and new species to emerge. Variation is both necessary and hazardous to the survival of species, and it is the ability of an organism to maintain a balance between these two effects that allows a species to survive and proliferate. Variation must be maintained within a species to allow it to undergo genetic experimentation in order to cope with different eventualities. However, the instability required to give rise to mutation, recombination, and ultimately variation can be highly deleterious if not strictly controlled. It is the study of these deleterious effects that has elucidated many features of the genetic constitution, and also revealed new information concerning the inheritance of genetic material and cellular metabolism. These processes of mutation are, in general, highly non-random, so that it is ultimately the composition of the genome which dictates the stability of the genome. These inherent instabilities can be manipulated by proteins, different cellular contexts and external factors to give the bewildering array of mutational and recombination mechanisms which shapes the genomes of all organisms.

A brief history of mutation research

Explorations into the origin of variation have been documented by Hippocrates, Aristotle and others since about the fourth and fifth centuries B.C. (Bartsocas, 1984). However, the main interest up until about the 19th century lay in the classification of organisms. During this time, Nature was believed to be programmed to achieve the ideal organism, while variations were perceived as imperfections and were generally dismissed under the heading of Vitalism, or life force. This rather idealistic view was gradually abandoned towards the end of this period following the classification of most known organisms. These studies were then replaced by the more scientific examination of differences, both between and within species, to investigate the evolution of organisms. This shift from the descriptive to the explanatory was heralded, in 1809, with Lamarck's theory of the inheritance of acquired characteristics. This view was radically opposed in 1869 by Darwin, who put forward his theories of evolution as an adaptive process in which heritable variation is necessary for the survival of species. It was also around this time that Mendel derived his theory of hereditary by analysing variation in successive generations of pea plants (Bateson, 1894a).

In the 19th century the distinction between the inheritance of pre-existing mutation, and the occurrence of *de novo* mutation was beginning to be made (Bateson, 1894b; Korschinsky, 1901). This work was pioneered by de Vries (1901) who derived several prescient laws of mutation following the detailed examination of variation in the evening primrose. This work earned him the title of "the father of mutation". de Vries suggested that organisms could spontaneously change or mutate, and concluded that this type of change is different from the variation normally observed because it appears suddenly and periodically, and occurs within the hereditary material. He considered mutation to be dictated by internal causes, although the timing of mutation was determined by external causes. He also suggested that heredity was mediated by chemical molecules which are not created new but are handed down, largely in a preserved form. This was proved correct following Sutton and Boveri's independent observations (Sutton, 1903) that the segregation of chromosomes during meiosis paralleled the random assortment of Mendel's "factors" or genes, during the production of gametes in peas. Also at this time Bateson (1894b) put forward his theory that genetic diseases are inherited in an all-or-nothing manner. This was not refuted until Morgan *et al.* (1912) demonstrated that mutation could give rise to multiple phenotypes in the form of eye colour mutants in *Drosophila.* Sturtevant (1913) subsequently used the amount of recombination observed between different genes to map them linearly along the chromosome, thereby defining the gene as a unit of recombination. This is the first of many examples where analysis of mutation and recombination contributed to the understanding of more diverse cellular processes, in this case, heredity of the genetic material.

Artificial induction of gene mutation

This next major period in the study of mutation began with the discovery that X-rays could induce mutation in *Drosophila* (Muller, 1927) and maize (Stadler, 1928). The main contribution of this work was in understanding the non-randomness of mutation, and the different spectra of mutations arising from different types of mutagen, such as X-rays and various chemicals (Auerbach, 1979). It was postulated that the specific and characteristic patterns of mutation of the different mutagens were influenced by the DNA flanking the mutated lesion (Benzer & Freese, 1958; Brenner *et al*, 1958; Benzer, 1961; Koch, 1971). This work also demonstrated the existence of pre-mutational lesions, and it was later demonstrated that these lesions could be fixed as mutations, or restored by cellular repair mechanisms. This period was associated with a relatively rapid increase in the understanding of genetic mutation because it removed the constraint imposed by the labour-intensive and time-consuming search for spontaneous mutations. However, progress was hindered by the absence of a molecular model of the hereditary material to explain these specific patterns and non-randomness of mutation. In fact, there was speculation that, because X-rays induced both mutation and chromosomal rearrangement, all mutations were a result of chromosomal abnormalities (Goldschmidt, 1946). It was not until the structure of DNA was resolved by Watson and Crick in 1953 that the nature of the mechanisms employed by the different mutagenic agents

could finally be speculated upon. Unfortunately, these studies revealed little about spontaneous mutations, and for a long time it was thought that genes were very sensitive to heat and that spontaneous mutation was caused by fluctuations in micro-temperature around the DNA.

Molecular biology and recombinant DNA technology

Pre-1960's mutation research was based on the visual identification of differences between parents and offspring, and the examination of cytogenetic lesions in chromosomes. However, this revealed little about the molecular mechanisms of mutation, and often failed to reveal the genetic cause of many diseases. The molecular era started with the discovery that protein electrophoresis could be used to identify mutant polypeptide sequences (Ingram, 1957). By determining the amino acid sequence of the dysfunctional protein, the nature of the mutation at the DNA level could also be inferred (Crick, 1961). This technique revealed that variation at the molecular level was far greater than expected, and it was estimated that around a third of all coding loci are polymorphic (Lewontin & Hubby, 1966; Nevo *et al.*, 1984). However, this electrophoretic technique can only reveal variation in the form of amino acid substitutions which alter the overall charge of the protein. It also fails to detect synonymous changes (those which code for the same amino acid) that may exist within the coding sequence, and mutations in the untranslated regions of the gene, such as the introns and the regulatory regions.

The study of mutation was revolutionised in the 1970's with the advent of recombinant DNA technology. These techniques included restriction enzyme digestion (Meselson & Yuan, 1968) and ligation (Jackson *et al*, 1972), Southern blotting (Southern, 1975), DNA sequencing (Maxam & Gilbert, 1977; Sanger *et al.*, 1977), and the polymerase chain reaction or PCR (Saiki *et al.*, 1985; Kogan *et al*, 1987). These techniques allowed unlimited amounts of DNA to be generated, either by bacterial cloning or by PCR, which could then be examined in detail. The first human gene was cloned by Shine *et al* in 1977 and today, approximately 10% of the human genome has been sequenced, and thousands of genes have been cloned and non-coding regions have been sequenced (the progress of the human genome sequencing projects can be monitored at the NCBI website on http://www.ncbi.nlm.gov/genome/seq/). Using a combination of Southern blotting and restriction enzyme analysis variation in specific sequences could be examined directly at the DNA level. This revealed the presence of multiple restriction fragment length polymorphisms, or RFLPs (Kan & Dozy, 1978; Jeffreys, 1979) demonstrating that there was a considerably higher level of diversity at the DNA level than could be identified by protein mobility assays. However, the full extent of this variability was only realised following DNA sequencing which revealed additional polymorphisms that did not affect restriction sites. Southern blotting also allowed the direct detection of deletions and duplications within genes (Orkin, 1978). This era provided a variety of formidable tools for the analysis of mutation and genetic disease, and was the main impetus for the shift back to the investigation into the causes of spontaneous mutation.

Spontaneous mutation

It was soon realised that the levels of mutagen required to induce mutation were far higher than existing physiological levels (Sparrow, 1950; Muller & Mott-Smith, 1970). This raised the possibility that artificial mutation might be induced through different mechanisms, and may not therefore be a useful model for spontaneous mutation. It is difficult to investigate spontaneous mutation because of the low frequencies at which it arises. Much that is known about spontaneous mutation in humans comes from the molecular and biochemical characterisation of disease, particularly cancer, and extrapolation of studies on smaller organisms such as bacteria and yeast. Spontaneous mutation arises in a highly non-random fashion. As with artificially induced mutation this is largely dependent on both the local DNA sequence, for example sequence repetivity and DNA methylation status; and enzymatic processes such as transcription, and higher order chromatin structure. Spontaneous mutation can alter the primary sequence of the DNA in a variety of ways. These can be split into two major classes, replication-associated mutation and recombination. These two classes are not definitive and there are many overlaps between them.

Replication-associated mutation

In 1900 it was first speculated that DNA replication may play a causative role in mutation (Auerbach, 1979). The extent of this involvement was not fully realised until 1967, following the discovery of a mutator phenotype that showed increased frequency of mutation, in bacteriophage T4. This phenotype is caused by mutations in the DNA polymerase which increase the frequency of point mutation throughout the bacteriophage genome (Speyer, 1965). Streisinger *et al.* (1967) further demonstrated that the non-mutated form of this enzyme could be error prone, particularly during the replication of runs of identical bases. This provided the first evidence that a substantial proportion of spontaneous mutations could be replication-dependent and endogenous in origin. There are a number of different types of DNA polymerase, each of which has a specific function and each also has a specific mutational fingerprint (Kunkel, 1985a & b; Lai & Beattie, 1988; Sanderson & Mosbaugh, 1998). The *in vitro* error frequency of DNA synthesis is between 10^{-9} and 10^{-11} per base replicated due to correct dNTP selection, proof-reading and mismatch repair during polymerisation (Kunkel, 1992). However, the accuracy of this process *in vivo* is much higher due to post-replicative DNA repair systems.

Repair mechanisms were first identified in 1949 when Kelner and Dulbecco independently demonstrated that short-wave visible light greatly reduced the lethal effect of short-wave UV. This phenomenon, called photoreactivation, is due to enzymatic repair mechanisms which use the complementary DNA sequence as the template for repair during replication (Rupert *et al.*, 1958; Pettijohn & Hanawalt, 1964). Subsequently, many different repair pathways have been identified in bacteria and yeast by studying mutants defective in repair; and in humans by studying inherited repair defects such as xeroderma pigmentosum, and by work on mutagen-resistant cell lines

(Kimball, 1987). These repair pathways are essential for maintaining the integrity of the DNA; this can be undermined firstly, because of the inherent inaccuracies within the cellular replication and transcription machinery; and secondly because of DNA damage caused by endogenous metabolites such as reactive oxygen species (Fujikawa *et al.*, 1998; Inoue *et al.*, 1998), and exogenous agents such as ionising radiation (Lehrman, 1995). For example, it has been estimated that DNA hydrolysis alone can affect 10,000 bases per human cell per day (Seeburg *et al.*, 1995). These repair systems can make "intelligent" decisions, for example certain systems will incorrectly repair badly damaged DNA rather than allow the existing lesion to persist (Walker, 1995). The best characterised repair system in most species is the mismatch repair system. Absence of mismatch repair results in tolerance to methylation damage resulting in strand breaks and cell death (Ciotta *et al.*, 1998); pre-disposition to cancer; increases in the instability of microsatellite sequences (Strand, 1993; Liu *et al.*, 1995a) and increases in homologous recombination and subsequent chromosome rearrangements (de Wind *et al.*, 1995). The mismatch repair system is also thought to cause cell cycle arrest if it recognises certain types of DNA damage (Hawn *et al.*, 1995; Anthoney *et al.*, 1996).

Replication-associated mutation is expected to occur at a higher frequency following paternal transmission because sperm has undergone fifteen times more cell divisions before maturation than oocytes (Vogel & Motulsky, 1986). This predication has been borne out in diseases such as Duchenne muscular dystrophy (DMD) where paternal transmission increases the rate of mutation by 14 times that of maternal transmission (Cooper & Krawczak, 1993). However, this is not true of all loci, for example mutation in the F9 gene which results in haemophilia B occurs at roughly the same rate following transmission from either parent (Ludwig *et al.*, 1992), and fragile X syndrome where the transmission of the full mutation always occurs through the female germline (Rousseau *et al.*, 1991).

Non-randomness of mutation

The majority of replication-associated mutations are single nucleotide polymorphisms (SNPs), accompanied by a much lower frequency of small deletions, inversions and duplications (Cooper & Schmidtke, 1984). Mechanisms of replication-associated mutation include chemical methods such as methylation and transition of CpG dinucleotides; physical methods such as slippage; and enzymatic processes such as misincorporation of nucleotides during replication. The sequence dependency of these processes means that mutation is non-randomly distributed (Cooper *et al*, 1985). In the human genome, the frequency of neutral mutation is higher than phenotypically expressed mutation. For example, the frequency of mutation which gives rise to synonymous codon substitutions is similar to that in non-coding DNA (~1 - 2 x 10^{-9} / bp / year), whereas the mutation rate which gives rise to non-synonymous codon substitutions is much lower (Strachan, 1992). Only those coding regions which require a polymorphic product (such as the HLA genes) show any significant level of heterozygosity (Strachan, 1992).

Repetitive sequences and unusual DNA structures

Replication-associated mutation appears to occur at inverted and direct repeats, symmetrical elements and selected DNA sequences more often than expected by chance (Cooper & Krawczak, 1993). This is potentially due to the difficulties associated with enzymatic replication of these regions, either because of the sequence repetivity, or due to the formation of DNA secondary structures obstructive to the polymerase. Palindromic or quasi-palindromic sequences, for example, can form hairpin secondary structures which cause the DNA polymerase to undergo strand switching, and may result in deletions (Ripley, 1982). These hairpin structures require large amounts of energy to form, and may therefore be more instrumental in larger deletions. Symmetric elements (e.g. 5′ CTGAAGTC 3′) are also significantly over-represented at deletion sites and, like palindromes, may also facilitate mutation by the formation of secondary structures.

Microsatellites and DNA slippage

Microsatellites generally have small repeats of 1-5 bp and array sizes ranging from ~10 bp to 1 kb. They are highly abundant throughout the genome and are generally highly polymorphic in terms of repeat array length (Mant *et al.*, 1991; Beckmann & Weber, 1992; Polymeropoulos *et al.*, 1992; Mahtani & Willard, 1993; Armour *et al.*, 1995). Replication slippage is recognised as the predominant mode of mutation generating new size alleles at human microsatellite loci (Figure 1.1A), with a 2:1 bias in favour of gain of repeats (Weber & Wong, 1993). A yeast-based system has also demonstrated this bias towards gain of repeats (Sia *et al.*, 1997), although some prokaryotic and *in vitro* systems show a bias towards deletion (Kunkel, 1990; Hite, 1996). PolyA rich microsatellites (which make up the majority of microsatellite repeats) and, to a lesser extent minisatellites are thought to originate by slipped mispairing during the replication of poly(A) tails associated with dispersed repeats such as Alus, L1s and HERV-L elements (Arcot *et al.*, 1995; Yandava *et al.*, 1997; Smit, 1996). Slipped mispairing increases with the amount of homology between the repeats, and is inversely proportional to the distance between them (Albertini *et al.*, 1982; Singer & Westlye, 1988). There is also a positive correlation between the size of the deletion and the length of the direct repeat (Cooper & Krawczak, 1993). Rates of replication slippage at microsatellites can be drastically increased by mutations affecting mismatch repair (Strand, 1993; de Wind *et al.*, 1995; Liu *et al.*, 1995a; Ciotta *et al.*, 1998). The normal rate of mutation of microsatellite repeats in human fibroblast cells has been estimated at 3.1×10^{-8} to 44.8×10^{-8}, consisting mostly of small mutations of around 4 bp (Boyer *et al.*, 1998). Most microsatellite repeats appear to have a minimum threshold size necessary for a repeat sequence to undergo mutation. In *S. cerevisiae* this threshold seems to be determined by the number of nucleotides (~ 8 nt) in the array rather than the number of repeats (Rose & Falush, 1998). Evidence suggests that telomeres, which are the specialised structures that protect and maintain chromosome ends, may also mutate by replication slippage (Baird *et al.*, 1995). Telomeres comprise a hexameric repeat

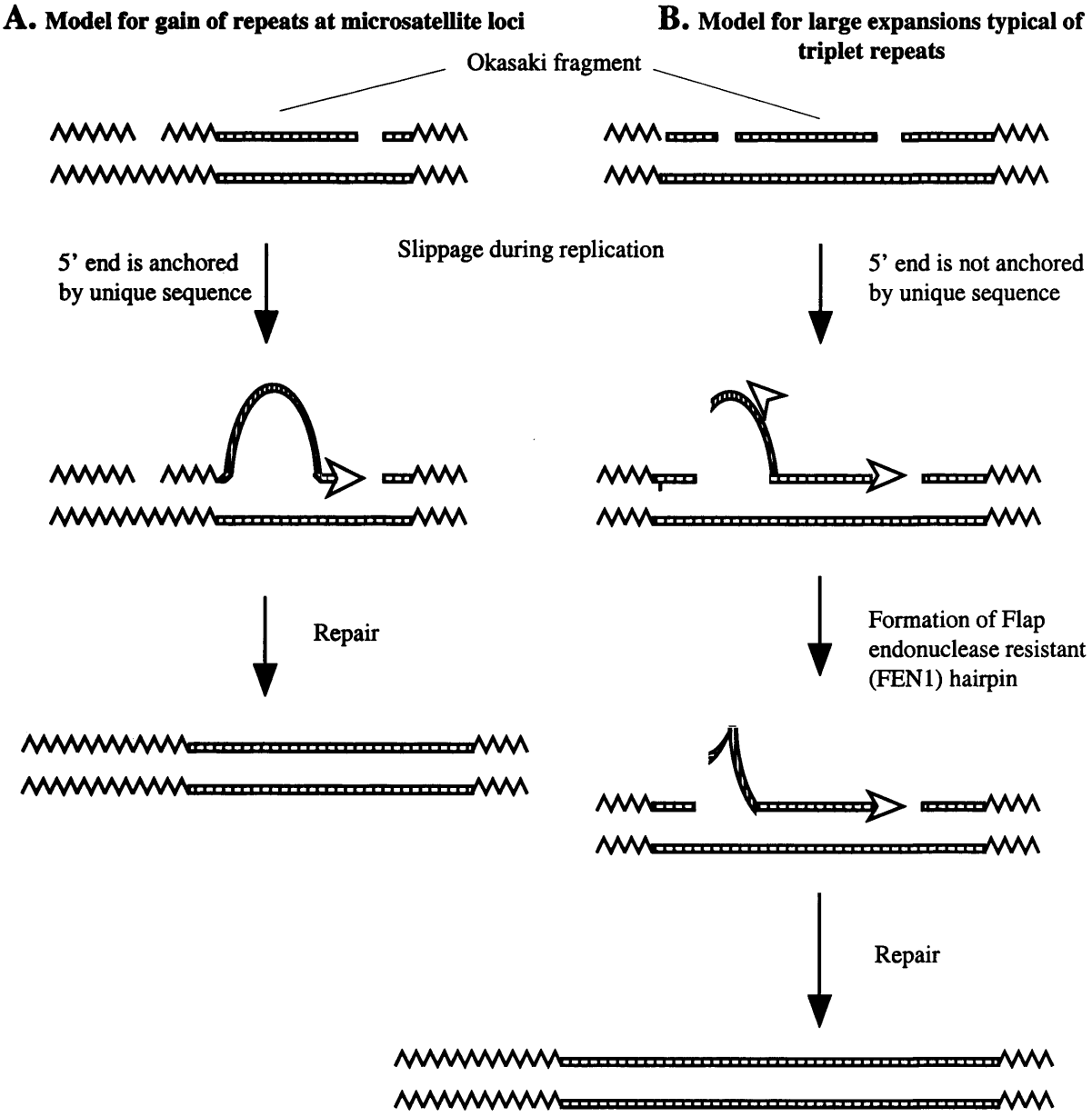


Figure 1.1. Okasaki fragment-based replication slippage model of repeat expansion.

Unique DNA is represented by jagged lines; repeat DNA is represented by the hatched boxes. (Figure and legend modified from Sutherland *et al.*, 1998).

A. In repeat arrays of less than ~240 bp the Okasaki fragment will be "anchored" in unique DNA at the 5' end. Slippage of this fragment can thus only occur in one direction, resulting in small additions of repeat sequences.

B. In repeat arrays of greater than about 240 bp, such as in the premutation state at trinucleotide repeat loci, the Okasaki fragment is not anchored and can slip in either direction along the repeat array. This may reveal sufficient DNA to loop back on itself to form secondary structures at the 5' end of the Okasaki fragment that renders it resistant to Flap endonuclease (FEN1). This enzyme is involved in ligation of these Okasaki fragments into a contiguous sequence. Inhibition of this activity further increases the instability of the array, resulting in the huge expansions typical of trinucleotide repeat arrays.

sequence based on the sequence $(TTAGGG)_n$, that is reminiscent of a microsatellite sequence. Slipped mispairing is also thought to play a causative role in the expansions of triplet repeat arrays associated with many human inherited genetic diseases (La Spada *et al.*, 1994; Zoghbi, 1996).

Triplet repeat disorders

These diseases all share the feature of anticipation which involves a premutation stage, where the number of repeats is intermediate between the normal and full mutation stages. The progression from pre- to full mutation in the fragile X syndrome, for example involves an increase in the number of repeats from between 52 - 200 to over 300 (Laird, 1987). This instability can be detected in both somatic and germline tissues of some individuals. In the case of fragile X syndrome, the pre-mutation will only progress to the full mutation by repeat copy number expansion on progression through the female germline. These large expansions remain unmethylated and appear to be unstable during embryonic development, giving an electrophoretic smear of variously sized deletion products in many tissues (Sutherland *et al.*, 1998). Other triplet repeat disorders include myogenic dystrophy, and X-linked spinal and bulbar muscular atrophy. These all posses $(CXG)_n$ repeats (Ashley & Warren, 1995) and, depending on the disease, size expansion can occur on progression through either the male or female germline (Chung *et al.*, 1993, Huntingdon's Disease Collaborative Research Group, 1993).

One theory to relate these huge expansions to the phenomena of anticipation at triplet disease loci has been put forward by Richards & Sutherland (1994). When the triplet repeat tracts are long there is a high possibility of an Okasaki fragment not containing any unique DNA (i.e. non-repetitive). This usually anchors the fragment, allowing slippage to occur in one direction only, as observed in "normal" microsatellite mutation and during the pre-mutation and the normal stages of these diseases (Figure 1.1A). The absence of this "anchor" means that the fragment can slip in either direction during polymerisation. This often exposes sufficient repeat DNA at the 5′ end of the Okasaki fragment to loop back and reanneal to itself, forming a hairpin. This prevents it becoming ligated to the previous, upstream fragment and increases the instability of the repeat tract, permitting the gross expansions typical of these diseases (Figure 1.1B).

Specific DNA sequences

The putative arrest site for DNA polymerase α has been implicated in causing small mutation events such as base substitution, deletion and insertion and is often referred to as the consensus site for deletion hotspots (Cooper & Krawczak, 1993). This deletion hotspot motif TG(A/G)(A/G)(G/T)(A/C) is similar to the core motifs of the immunoglobin switch regions and has been associated with mutation in different human genes (Purandare & Patel, 1997). These sequences may promote mutation by bringing together the primary and secondary DNA structures at the end of the replication fork following the arrest of DNA synthesis. 86% of these motifs in the

human genome are closely associated with palindromes (Weaver & DePamphilis, 1982), and it has been inferred that the two may work synergistically to promote mutation. The cleavage site for topoisomerase I (5′ CTT 3′), also appears to be associated with point mutation more often than expected by chance (Cooper & Krawczak, 1993).

Methylation and CpG dinucleotides

The most common modification in the human genome is the methylation of cytosine to 5-methylcytosine (5mC), 90% of which occurs at CpG dinucleotides (Cooper & Krawczak, 1993). The rate of mutation at CpG dinucleotides is 7.4 times higher than the base mutation rate in primates, which is about 2.2 x 10^{-9} per nucleotide per year (Bulmer *et al*, 1991). A substantial proportion of these dinucleotides (~15%) are clustered in gene coding regions and are largely unmethylated; the remainder are scattered throughout the genome and are largely methylated. 5mC residues are rapidly, spontaneously deaminated to thymine, resulting in the gradual attrition of CpG residues from the human genome. Accordingly, the amount of CpGs is lower than expected from genomic mononucleotide frequencies, and a large number (~35%) of RFLPs are located in enzyme recognition sequences which contain CpG dinucleotides (Barker *et al*, 1984; Cooper *et al*, 1985; Cooper & Youssoufian, 1988). The mutability of the CpG dinucleotide is dependent on various factors, such as the local DNA environment (Adams *et al.*, 1987), the population of origin (Pattinson *et al.*, 1990), and the protection conferred upon certain methylated residues thought to be important in genomic imprinting during development (Green *et al.*, 1990).

Tolerance of mutation in the human genome

Mutation in the eukaryotic genome is tolerated within the antibody variable (V) genes in order to maximise the number of potentially useful antibodies available to the immune system. Point mutation of the V genes is 6 - 7 times higher than the surrounding levels of spontaneous mutation, and it is highly localised to the region around expressed V genes (Shannon & Weigert, 1998). Cascalho *et al.* (1998) discovered that a protein normally involved in mutation repair is essential for V gene hypermutation. This occurs either 1.) by subversion of the normal function of this mismatch repair protein so that it repairs the parent strand using the mutant strand as template; or 2.) because the enzyme can no longer distinguish between the strands due to the removal of epigenetic markers so the template strand is randomly chosen. This subversion of a protein function which is normally involved in cellular repair processes has also been observed in V(D)J recombination (Kirchgessner *et al.*, 1995).

Other factors such as chromatin structure, various DNA-associated proteins, DNA topology and metabolism may also require consideration when examining the molecular mechanisms of mutation. For example, differences in local DNA sequence surrounding these motifs, and strand differences may also dictate mutation type and frequency. The probability that an incorrectly incorporated

nucleotide is fixed (base substitution), or realigns the DNA causing a deletion, is generally dependent on the flanking sequence. DNA base damage can also affect RNA synthesis and can prevent transcription, but in *E. coli* non-bulky lesions can easily be bypassed resulting in the synthesis of mutant RNA transcripts (Viswanathan & Doetsch, 1998). Different mismatches are repaired with different efficiencies (Jones *et al*, 1987) possibly because of the sequence-dependent proof-reading ability of the polymerase. DNA repair is also strand dependent (Wu & Maeda, 1987), so that the repair of the lagging strand is less efficient, resulting in a higher frequency of deletions in the lagging strand than in the leading strand (Kamiya *et al.*, 1998). This may be because the mutational lesion obstructs the formation of Okasaki fragments which results in gap formation and small deletions (Kamiya *et al.*, 1998). Strand differences in mutation have also been observed during transcription in *E. coli*. This is because the non-transcribed strand transiently exists in a single-stranded state, and is therefore more susceptible to point mutations (Beletskii & Bhagwat, 1998). Highly transcribed genes are therefore at higher risk of mutation than silent or weakly transcribed genes.

Recombination

Recombination is a fundamental process in many organisms, with essential roles in both meiosis and mitosis. However, it has to be carefully controlled to prevent deleterious genomic rearrangements. During **mitosis**, recombination is important for the repair of DNA double strand breaks (DSBs) which would otherwise block DNA replication, ultimately leading to a loss of genomic integrity and death. During **meiosis**, recombination exchanges alleles along the linear chromosome to bring new combinations together. In addition, crossing over between two homologues (non-sister chromatids) creates a physical link between two chromosomes to ensure proper disjunction during meiosis. Defects in this process can lead to abnormal chromosome segregation and the production of non-viable progeny. The disadvantages of recombination are that it can separate beneficial allele combinations, and create lethal or deleterious genetic rearrangements which can contribute to neoplastic transformation. It is probably these contradictory effects which engenders the non-random distribution of recombination, in the form of distinct hot and cold spots throughout the genomes of all organisms studied. Meiotic recombination has been intensively investigated in the yeasts *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*. Mitotic recombination, because it occurs at a lower rate (Magni, 1963), has mainly been studied by examining the products of artificially introduced and mutagen-induced DSBs in yeast, *E. coli* and various eukaryotic cell lines (Osman & Subramani, 1998).

Recombination is defined either as homologous or non-homologous (illegitimate). Homologous recombination can occur anywhere within a region of identical sequence, although often it is targeted to specific sequences. Non-homologous recombination occurs within sequences sharing little or no identity; this is generally non-conservative, involving loss or gain of DNA sequences at the site of recombination. Non-homologous recombination also encompasses site-specific

recombination. This originates at specific sites in two DNA sequences which otherwise show no homology e.g. protein binding sites. Most recombination events require homology of at least fourteen consecutive nucleotides between chromatids (Rubnitz & Subramani, 1984), although many large deletions can have as little as 2 bp of homology at the breakpoints (Morris & Thacker, 1993). Recombination is probably mediated by the formation of highly stable recombination intermediates, which may explain the predominance of these events in GC rich regions, which are thought to form stable secondary and tertiary structures (Singer & Westlye, 1988). Unusual sequences such as palindromes and repeat DNA promote genetic instability and possibly recombination for example in V(D)J recombination (Akgün *et al.*, 1997).

Mitotic repair of DSBs

Mitotic recombination in both human culture cells and *Saccharomyces* is significantly increased following the introduction of artificial DSBs (Osman & Subramani, 1998; Liang *et al.*, 1996). DSB-stimulated mitotic recombination can lead to conversions, crossovers, deletions, duplications, inversions and translocations, although the most common are conversions and deletions (Osman & Subramani, 1998). Mammals and *S. cerevisiae* utilise different pathways of recombinational repair of artificially introduced and mutagen-induced DSBs. In *S. cerevisiae*, DSBs are mainly repaired by homologous recombination, while in mammals DSBs are repaired by non-homologous recombination (Liang *et al.*, 1996). This may reflect the genomic context of the different organisms. Yeast mainly consists of coding DNA so that non-homologous repair will generally be highly disruptive and is used only as a last resort. Conversely, while homologous recombination may be preferable to conserve sequence, in mammalian cells it can be deleterious in the absence of the sister chromatid, potentially causing events such as the loss of heterozygosity in tumour cells (Lasko *et al.*, 1991), and gross rearrangements between distant repeats (e.g. Jeffs *et al.*, 1998).

Within the eukaryotic cell there are many legitimate, tightly regulated recombination mechanisms which are initiated by mitotic DSBs e.g. *Saccharomyces* mating type switching and mammalian V(D)J recombination. In mating-type switching in *Saccharomyces*, conversion at the *MAT* locus is stimulated by a DSB to replace DNA from one of two other donor loci located on different arms of chromosome 3 (Klar, 1992). This process is stimulated by an enhancer element which is thought to expose or sequester the entire left arm of chromosome 3, as dictated by the allelic state of the *MAT* locus (Haber, 1998). In humans, V(D)J recombination gives rise to the tremendous variability essential for antibody and T-cell receptor specificities. Recombination rearranges the variable (V), diverse (D) and joining (J) elements, which are at distant locations in germ cells, into a contiguous sequence. This process probably involves the formation of hairpins by palindromic sequences which are then nicked by DSBs to stimulate recombination (Akgün *et al.*, 1997). It therefore shows many resemblances to nonhomologous end joining mechanisms used for processing DSBs in many other cell types (Jeggo *et al.*, 1995; Oettinger, 1996).

Meiotic recombination and chiasmata

The non-random distribution of meiotic recombination was first observed following the examination of chiasmata. These are the cytological manifestations of chromosome crossovers during prophase of meiosis. For proper chromosome disjunction during meiosis (the movement of chromosome homologues to opposite poles of the cell), it appears to be essential that each pair of homologous chromosomes undergoes at least one crossover. For example, if a crossover fails to occur between the pseudoautosomal region of the Y chromosome and its homologous X chromosome, disjunction will not occur and viable gametes are not produced (Rapp *et al.*, 1988). The distribution of chiasmata does not appear to be primarily dictated by the underlying DNA sequence. However, it has been demonstrated that altering the sites of crossover changes the fidelity of chromosome segregation in yeast (Ross *et al.*, 1996), this may also be the case in humans (Sherman *et al.*, 1994) and *Drosophila* (Hawley & Theurkauf, 1993). Sites of crossing over may also vary with sex, haplotype, overall chromosome structure, and position within the genome (Lichten & Goldman, 1995). For example, many chiasmata are located in sub-terminal regions of chromosomes, particularly in males (McInnis *et al.*, 1993).

The study of chiasmata distribution revealed two major features; firstly that chiasmata seem to repel each other along the length of the chromosome, a phenomena called interference; and secondly that recombination appears to be repressed at the centromere and telomeres. The molecular basis of interference is unknown but may be due to some feature of the synaptonemal complex. This has been suggested by studies on *S. pombe* which does not appear to form synaptonemal complexes, nor is there any crossover interference during meiosis (Roeder, 1997). Another curious feature is that the frequency of chiasmata is inversely proportional to the chromosome length, in *S. cerevisiae* the shortest chromosome recombines at 2 - 3 fold higher than the larger chromosomes. This is also true of humans (Kaback, 1996). This apparent increase of crossover interference on the larger chromosomes may prevent excess recombination on large chromosomes while ensuring at least one crossover on smaller chromosomes (Kaback, 1996).

The suppression of recombination at both telomeres and centromeres, despite their high sequence repetivity, may be due to specific meiotic functions of these chromosomal regions. Detailed analysis of the Xp/Yp telomeres and their adjacent sequences have demonstrated that linkage disequilibrium extends from markers in the flanking DNA into the start of the telomere (Baird *et al.*, 1995). This apparent suppression of recombination contrasts with the elevated levels of recombination generally found in the subtelomeric regions (McInnis *et al.*, 1993). One possibility is that these patterns of recombination are due to the attachment of the chromosome ends to the nuclear envelope, called bouquet formation (Dernburg *et al.*, 1995). This occurs at the same time as, and is thought to promote chromosome homologue pairing during meiosis (Scherthan *et al.*, 1996; Bass *et al.*, 1997). This function may protect the telomeres from crossing over during meiosis while promoting recombination between the subtelomeric regions. This also correlates with

the fact that synaptonemal complex formation initiates close to chromosome ends (Loidl, 1990). Human centromeres consist of heterochromatic DNA which is made up of highly repetitive tracts of satellite DNA, the most predominant of which is α-satellite DNA (Pardue & Gall, 1970; Choo, 1997). These regions are capable of recombination (Laurent *et al.*, 1997; Kapitonov *et al.*, 1998) but curiously, the meiotic exchange rate is significantly lower compared to the rest of the genome (Choo, 1998; Mahtani & Willard, 1998). For a long time this suppression was assumed to be due to the condensed state of centromeric heterochromatin during meiosis. However, this suppression still exists in centromeres with no visible heterochromatin, and it appears that recombination suppression is a result of the higher-order chromatin organisation of the centromere. This has been demonstrated by the fact that certain histones are only associated with active centromeres (Choo, 1998). Alternatively, this suppression may be a result of the presence of the kinetochore which may block access to the underlying DNA (Choo, 1997). This emphasises the important role played by DNA-associated proteins, and chromatin conformation in meiotic recombination.

Meiotic recombination and transcription

Transcriptional promoters appear to correlate very closely with hotspots of meiotic recombination in *S. pombe* and *S. cerevisiae* (Nicolas, 1998). In *S. pombe* the most well characterised hotspot is caused by the M26 point mutation at the *ADE6* locus. This mutation creates a heptamer which is a binding site for two transcription factor-like proteins. Disruption of this site reduces recombination, but if this site is relocated to a different part of the genome it does not always retain its recombinogenic activity, implicating other necessary factors. The extensive distribution of these heptameric sequences throughout the genome of *S. pombe* suggests that these sites may account for about 50% of meiotic recombination in this organism. It has also been suggested that this M26 site may represent a recombination enhancer that fulfils a similar function to transcriptional enhancers.

In *S. cerevisiae* there is a strong correlation between transcription, recombination and DSBs. In both native and artificial systems recombination hotspots are controlled by upstream *cis*-acting elements or initiation sites, and meiosis-specific DSBs are induced at these sites. Breakpoints represented by DSBs are not located in a specific sequence within these hotspots but are generally scattered within a region of approximately 50 - 200 nucleotides. These regions often represent transcriptional promoter sequences, and disruption of binding sites for transcription proteins decreases recombination frequencies. These transcription-activating protein-binding sites are reasonably interchangeable with regards to ability to initiate recombination in different parts of the genome; although recombination activity is always reduced if the binding site is mutated. Curiously however, hotspot activity does not appear to be correlated with transcriptional strength. DSBs in *S. cerevisiae* all occur in chromatin that is hypersensitive to DNase I and micrococcal nuclease, indicating that they only arise in accessible sites. It has also been shown that hypersensitivity to micrococcal nuclease increases specifically at hotspots during early meiotic prophase, prior to the appearance of DSBs. This suggests the chromatin is adapting to open up the DSB sites. Although it

is not clear whether the opening up of chromatin precedes the targeting of the DSB endonuclease, or whether the assembly of the endonuclease apparatus at the DSB site causes the opening of the chromatin, as in transcription. Variation as high as two-fold in the distribution of DSBs in the genome of *S. cerevisiae* suggests that there are multiple levels of regulation, for example chromatin structure, type of initiation and enhancer regions (Nicolas, 1998). Not all DNase hypersensitive sites are DSB sites, indicating that recombination initiation may be controlled by factors such as alterations in the chromatin structure by binding transcription factors, as demonstrated at the *HIS4* locus.

In humans, Alu-mediated meiotic recombination is targeted mainly to the region between the A and B boxes of the RNA polymerase III promoter. This suggests that there may also be a link between transcription and recombination in humans (Figure 6.1A, Chapter 6).

Molecular mechanisms of meiotic recombination

There are multiple models for the molecular events which occur during recombination (Osman & Subramani, 1998). It is likely that many of these could be functional at any one time, depending on different levels of control in different organisms. The following is based on the Meselson-Radding (1975) double strand gap repair model (Figure 1.2). This seems to be the most appropriate for meiotic recombination, at least in *S. cerevisiae*, and is able to explain the dual existence of reciprocal (crossing over) and non-reciprocal (gene conversion) recombination.

1. Initiation by the formation of double strand breaks or single strand gaps. Despite the plethora of evidence for initiation occurring by DSBs in *S. cerevisiae*, there is very little to confirm that this is also the case in other eukaryotes. DSBs can initiate recombination in other eukaryotes, as demonstrated by mitotic DSB repair, although it is not clear whether this is also true of meiotic crossover initiation. However, proteins involved in the formation and repair of meiotic DSBs are conserved across taxa; but it has been demonstrated that the homologue of the *S. cerevisiae* DSB-endonuclease identified in *C. elegans* is not necessary for chromosome synapsis in this organism (Dernburg, 1998). In addition, DSBs have not yet been physically associated with the M26 recombination hotspot in *S. pombe*, although it has been associated with single stranded breaks (Osman & Subramani, 1998).

2. Activation of the single stranded DNA. In yeast this is achieved by a 5′ to 3′ exonuclease activity which digests (resects) the DNA from the DSB to expose single-stranded overhangs (Smith & Nicolas, 1998). In human minisatellite mutation this has been postulated to occur by expansion of the break formed by staggered nicks to generate a gap flanked by single strand overhangs (Buard & Jeffreys, 1997).

3. Strand-exchange. This involves homologues of the *E. coli* RecA protein that catalyse the pairing of homologous molecules and the initiation of strand exchange, suggesting these features

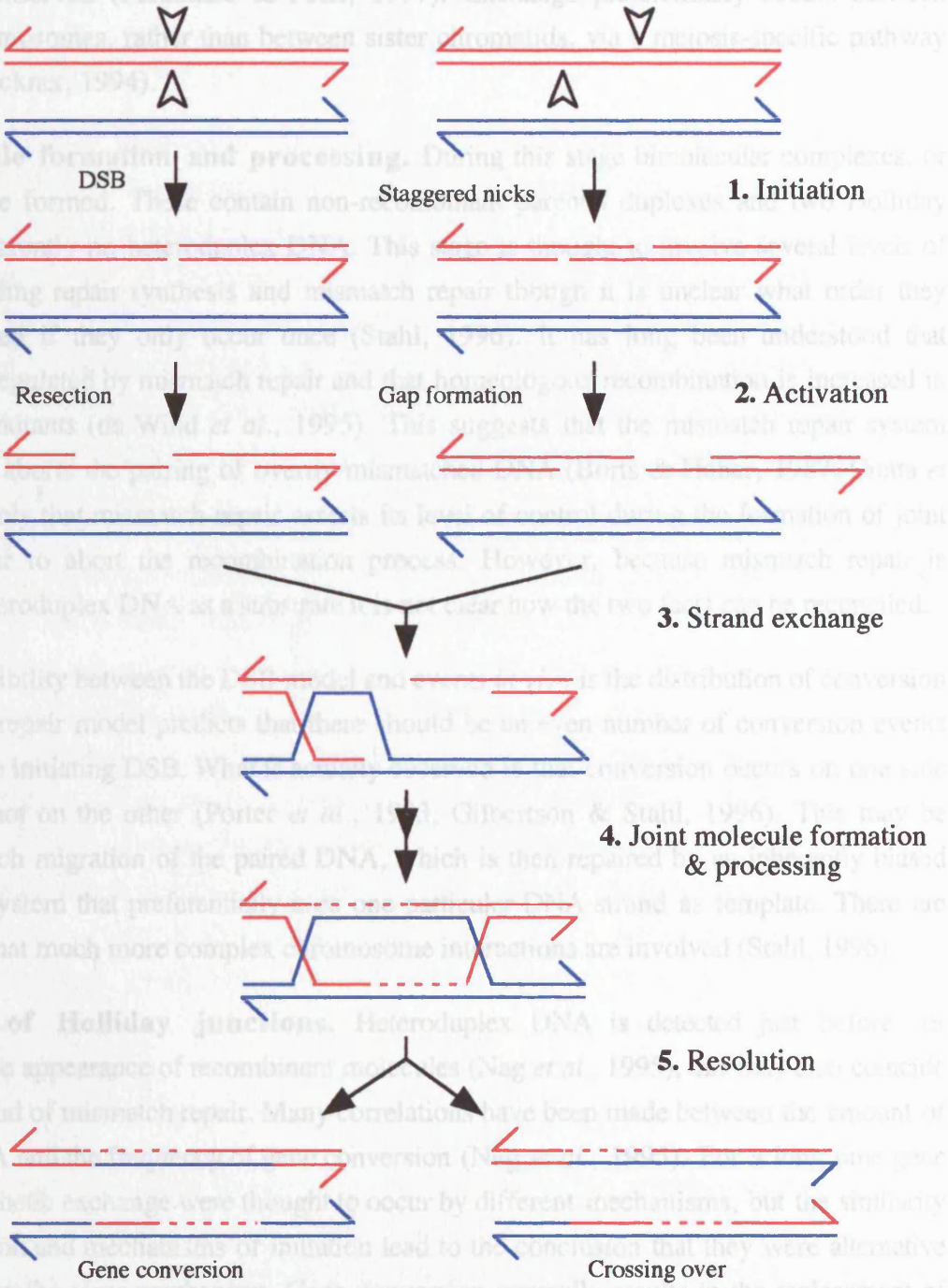


Figure 1.2. The double strand break gap repair model for meiotic recombination in eukaryotes.

1. Recombination is initiated either by DSBs in *S. cerevisiae*, or other models suggest that staggered nicks may be more appropriate (see text). **2.** Single strand DNA is exposed either by 5' to 3' exonuclease digestion from the DSB, or by expansion of the breaks formed by the staggered nicks. **3.** One of the single stranded DNAs invades the homologous duplex and displaces a D-loop. **4.** Several steps involving the formation of a joint molecule which contains both duplexes and two Holliday junctions. Repair synthesis (dashed lines) and mismatch repair are thought to occur in this stage of processing, but heteroduplex DNA has not been detected. **5.** Resolution of the joint molecule can give rise to two products, either gene-conversion (non-reciprocal recombination) or equal crossing over (reciprocal recombination).

may be highly conserved (Purandare & Patel, 1997). Exchange preferentially occurs between homologous chromosomes, rather than between sister chromatids, via a meiosis-specific pathway (Schwacha & Kleckner, 1994).

4. Joint molecule formation and processing. During this stage bimolecular complexes, or joint molecules are formed. These contain non-recombinant parental duplexes and two Holliday junctions, but apparently no heteroduplex DNA. This stage is thought to involve several levels of processing, including repair synthesis and mismatch repair though it is unclear what order they occur in, or indeed if they only occur once (Stahl, 1996). It has long been understood that recombination is regulated by mismatch repair and that homeologous recombination is increased in mismatch repair mutants (de Wind *et al.*, 1995). This suggests that the mismatch repair system either prevents or aborts the pairing of overtly mismatched DNA (Borts & Haber, 1987; Ciotta *et al.*, 1998). It is likely that mismatch repair asserts its level of control during the formation of joint molecules in order to abort the recombination process. However, because mismatch repair is thought to use heteroduplex DNA as a substrate it is not clear how the two facts can be reconciled.

Another incompatibility between the DSB model and events *in vivo* is the distribution of conversion events. The DSB repair model predicts that there should be an even number of conversion events on each side of the initiating DSB. What is actually observed is that conversion occurs on one side of the DSB, but not on the other (Porter *et al.*, 1993; Gilbertson & Stahl, 1996). This may be explained by branch migration of the paired DNA, which is then repaired by an inherently biased mismatch repair system that preferentially uses one particular DNA strand as template. There are also suggestions that much more complex chromosome interactions are involved (Stahl, 1996).

5. Resolution of Holliday junctions. Heteroduplex DNA is detected just before, or concurrent with the appearance of recombinant molecules (Nag *et al.*, 1995), this may also coincide with a second round of mismatch repair. Many correlations have been made between the amount of heteroduplex DNA and the frequency of gene conversion (Nag *et al.*, 1995). For a long time gene conversion and genetic exchange were thought to occur by different mechanisms, but the similarity between the position and mechanisms of initiation lead to the conclusion that they were alternative products derived by the same mechanism. Gene conversion generally results in the replacement of one set of information on one allele with information from the other. It is termed non-reciprocal because one allele acts as a donor of information and remains unchanged, while the recipient is always altered, often quite dramatically. In *S. cerevisiae* about 59% of meiotic crossovers are accompanied by gene-conversion. It is not clear how, or when the decision is made to resolve the Holliday junctions by gene conversion or crossing over, or whether gene conversion is simply the result of aborted crossover events. The answer may depend on whether gene conversion has a unique function within the human genome. It has been suggested that gene conversion may be a useful mechanism of experimenting with different combinations of large regions of DNA. This would allow extensive changes to take place without progressing through multiple single mutational

changes, any one of which may be deleterious (Ninio, 1996).

Recombination-mediated rearrangements

The main consequence of malfunction during meiotic recombination is the non-disjunction of chromosomes. In humans non-disjunction of chromosomes increases with age of the individual, this is most pronounced in reproducing females, and becomes a significant threat to the embryo with a maternal age of over 35 (Sherman *et al.*, 1994). In men, the frequency of non-disjunction does not reach the same level until they are over 55. Chromosome disjunction related to age (e.g. trisomy 21) appears to correlate with both a reduction in recombination and an alteration in the site of chromosomal exchange. It is thought to be due to the chromosomes becoming entangled at some stage during meiosis, and not being able to separate properly (Ross *et al.*, 1996).

The other major consequence of dysfunctional recombination involves gross rearrangements. The detrimental effects of these are evident in the progression of somatic cells to malignancy and the effects of *de novo* genetic disease. Rearrangements include deletions, duplications and inversions, the most common of which are deletions. This may be because many deletions occur by intra-allelic processes, so that not all are the products of unequal crossing over, only very occasionally are the reciprocal products of unequal crossing over observed. It is also possible that duplications may be more unstable and show a high rate of reversion, as observed in duplications at the HPRT gene locus (Monnat *et al*, 1992). Although this is not true of all duplications, some of which can persist through multiple generations. Inversions are also rare but curiously, large inversions generally do not produce phenotypic effects or alter viability, although they may interfere with crossing over (Auerbach, 1976; Therman, 1986). This is may be because the inverted region will often maintain intact coding regions that can still operate as normal, regardless of orientation. Large inversions are also non-random, for example inv(9) makes up about 40% of all inversions studied (Kaiser, 1980). The comparative rarity of duplications and inversions means that much less is known about them.

A large number of recombination events occur between or within repeated sequences in the genome. This is not surprising when you consider that the human genome mainly consists of non-coding sequence, only about 3% comprises coding DNA, the remainder consists of regulatory sequences, introns and intergenic spacers (Kazazian, 1998). Much of this non-coding DNA is made up of repetitive sequences which are ubiquitous in eukaryotes (Schmid & Jelinek, 1982). These repetitive sequences include dispersed repeats, such as Alus, L1s and pseudogenes, and tandemly repeated arrays, such as minisatellites.

Dispersed repeats

Dispersed (or interspersed) elements make up 35.5% of the human genome (Smit, 1996). There are three types of dispersed repeats; retrotransposons, retrovirus-like elements, and DNA transposons (Smit, 1996). Pseudogenes may also be classified as dispersed repeats. Retrotransposons are the

most abundant class of dispersed repeat and can be divided into short (100 - 400 bp) and long (6 - 8 kb) interspersed elements (SINEs and LINEs). The most prolific SINEs are Alu elements which comprise around 6% of the human genome (Schmid & Jelinek, 1982) and the most common LINE element is L1, or Kpn I repeats (Kazazian, 1998). Dispersed repeats contribute to recombination within the genome in two different ways, firstly by transposition to different genomic locations (e.g. Economou-Pachnis & Tsichlis, 1985; Vidaud *et al.*, 1988). Secondly, recombination between remote elements (particularly Alus) in the genome contributes to a significant proportion of genetic disease (for Alus see Chapter 6; for L1s see Burwinkle & Kimimann *et al.*, 1998; Schwartz *et al.*, 1998).

Transposition of dispersed repeats

The ability of some dispersed repeats to move around the genome was first documented by Barbara McClintock in maize (1951). She demonstrated that transposable elements could spontaneously change genomic location, both between and within chromosomes. These movements can not only insert into and disrupt coding sequences but they can also destabilise the surrounding DNA and cause chromosome breakage. Such increases in genomic mutation rates have also been observed following the integration of viruses and even random sequences (Taylor, 1963; Shapiro, 1969; Daniell *et al*, 1972; Shimada *et al*, 1973). It is mainly Alu and L1 repeats which are retrotranspositionally active, although these events are actually very rare. For example, only seven cases of *de novo* L1-mediated transposition (Kazazian *et al.*, 1988, Woods-Samuels *et al.*, 1991; Miki *et al.*, 1992; Narita *et al.*, 1993; Holmes *et al.*, 1994; Kazazian & Moran, 1998), and eleven examples of Alu-mediated transposition (Economou-Pachnis & Tsichlis, 1985; Stoppa-Lyonnet et al, 1990; Muratani *et al.*, 1991; Wallace *et al.*, 1991; Vidaud *et al.*, 1988; Goldberg *et al.*, 1993; Janicic *et al.*, 1995; Miki *et al.*, 1996; Kazazian & Moran, 1998) have been identified in humans. Retrotransposition occurs mostly in the germline or during early development, although one case of somatic L1 insertion has been described (Miki *et al.*, 1992). Retroviral-like elements also appear to have retained their ability to transpose in some mammals, but none have been reported in humans. However, some human endogenous retroviruses (HERVs) still contain intact genes encoding the reverse transcriptase, suggesting that they may still retain this activity (Kazazian, 1998). DNA transposons move around the genome by a "cut and paste" mechanism which is thought to stimulate recombination, and may be responsible for many chromosome rearrangements in plants, yeast and *Drosophila* (Auge-Gouillou *et al.*, 1995; Smit, 1996). It has been postulated that an inactive *mariner* transposon adjacent to the CMT1A-REPs on chromosome 17p11.2-p12 provides a target for a functional *trans*-acting transposase enzyme. Subsequent destabilisation of this region causes recombination, generating the reciprocal deletion/duplication products which give rise to two inherited neuropathies (Reiter *et al.*, 1996).

In species such as *Drosophila*, maize and yeast, mobilisation of dispersed repeats is important in shaping genomes, and may be actively involved in the creation of new species by promoting

karyotypic change (Voytas, 1996). In mammals it has been demonstrated that transposable elements are capable of promoting massive structural changes without disrupting the functioning of normal genes (Waugh O'Neill *et al.*, 1998). It has also been demonstrated that mobilisation of these elements may, occasionally, be beneficial. Firstly, dispersed repeats may be recruited to repair DSBs; this probably occurs by sequestration of the reverse transcriptase and may explain, in part, the large numbers of these elements in the mammalian genome (Boeke, 1997; Moore & Haber, 1996; Teng *et al.*, 1996). Secondly, L1 elements may be involved in exon shuffling to create new genes by co-mobilising a 3′ flanking segment of DNA during retrotransposition of its own sequence. It is able to do this because the poor polyadenation signal at the end of the element often permits sequence read-through, effectively enabling it to pick up different DNA sequences and carry them to different parts of the genome (Boeke & Pickeral, 1999; Moran *et al.*, 1999)

Recombination between dispersed repeats

Of the dispersed repeats, it is the Alu elements that play the largest role in mediating recombination between remote parts of the genome (this topic will be reviewed in more detail in Chapter 6). This may be due to their abundance and/or the homology between these elements. Meiotic recombination involving Alu repeats is targeted to particular elements within a cluster and to particular features within that element. This suggests that recombination between these elements is non-random, although the mechanisms behind this targeting are not clear. They may be related to features such as transcriptional ability of the Alu, homology between the two participating Alus, or chromatin conformation (Schmid, 1996). Most of the transposon-mediated recombination events that have been characterised result in deletion (e.g. Burwinkle & Kilimann, 1998; Tvrdik *et al.*, 1998) although some can cause inversions (Jennings *et al*, 1985; Schwartz *et al.*, 1998) and duplications (Stoppa-Lyonnet *et al.*, 1991). Pseudogenes cause mutation by gene conversion of part or all of the pseudogene into its active counterpart (Collier *et al.*, 1994; Eikenboom *et al.*, 1994; Tusie-Luna & White, 1995; Hulsebos *et al.*, 1996; Watnick *et al.*, 1998). In addition, dispersed elements may not always be directly involved in mutation but can promote recombination by stabilising secondary structures of recombination intermediates (Karathanasis *et al*, 1987).

Dispersed repeats - true genomic parasites or a hidden benefit?

The highly disruptive potential of these elements is not compatible with their abundance in the genome. To reconcile these features it has been suggested that these are truly "selfish" elements and that they are suppressed or inactivated by genomic mechanisms. This may be achieved by sequestering these elements in inaccessible regions of the genome, such as telomeres, telomeric chromatin, centromeres, and regions of low transcription (Prades *et al.*, 1996; Voytas, 1996; Junakovic *et al.*, 1998; Kidwell & Lisch, 1998). Transcription is necessary for transposition of these elements, and has also been implicated in promoting recombination between them (Chapter 6). It has been postulated that these elements may be inactivated by transcriptional silencing,

possibly by methylation (Yoder *et al.*, 1997; Kidwell & Lisch, 1998; Waugh O'Neill *et al.* 1998). The most convincing evidence in support of this was the observation that a ubiquitously undermethylated genome exhibited genome-wide amplification of these elements (Waugh O'Neill *et al.*, 1998). Furthermore, the majority of methylation sites in the human genome are found within inactivated dispersed repeats; and the CpG content of the Alus belonging to the younger subfamilies is nine times greater than in normal human DNA (Yoder *et al.*, 1997; Schmid, 1996).

The contrasting viewpoint is that Alus and other transposable elements have a functional role in the genome. It has long been a matter of dispute as to whether methylation is a genomic mechanism for the control of transcription (Bestor *et al.*, 1997; Bird, 1997). Schmid (1998) suggests that because Alus are differentially methylated in sperm and oocytes (Rubin *et al.*, 1994), they may be essential for genomic imprinting (differential expression of the parental genomes), or in chromatin organisation. It is argued that if methylation is necessary to prevent transcription and transposition of these elements, it is not logical that this suppression is lifted in tissues that contribute to the next generation (Bird, 1997). Transcription of Alu elements is also enhanced following viral infection, heat shock or cyclohexamide treatment (Russanova *et al.*, 1995; Liu *et al.*, 1995b). Alus may thus play some role in the cellular stress response. This may be mediated by Alu transcripts binding and inhibiting the protein kinase, PKR which phophorylates elongation factors to repress protein synthesis. Monopolising PKR by binding to Alu RNA would therefore up-regulate protein transcription, enabling the cell to respond quickly to the insult (Schmid, 1998).

Minisatellites

The repeat unit lengths of minisatellites range from 9 - 90 bp (Wong *et al.*, 1987; Armour, 1990) with array sizes in the order of 0.5 - 50 kb. Many minisatellites show extremely high variability as a result of high *de novo* germline mutation rates to new length alleles (Wong *et al.*, 1987; Jeffreys *et al.*, 1988; Vergnaud *et al.*, 1991). All hypervariable minisatellites studied to date also exhibit internal variation in the order of repeat types along the tandem array. This internal variation can be mapped by a technique called minisatellite variant repeat mapping by PCR (MVR-PCR). This is much more informative than typing these loci on the basis of size (Jeffreys *et al.*, 1990, 1991b; Neil & Jeffreys, 1993; Buard & Vergnaud, 1994; Armour *et al.*, 1996; Chapter 3). There are estimated to be between 1,500 - 3,000 hypervariable minisatellite loci per haploid genome (Braman *et al.*, 1985; Schumm *et al.*, 1985; Jeffreys *et al.*, 1987), the majority of which are clustered in the subtelomeric regions of the autosomal chromosomes (Royle *et al*, 1988; Amarger *et al.*, 1998).

<u>The effects of gene-associated minisatellite variation</u>

Coding minisatellites generally show much less variability than their non-coding counterparts, presumably due to selective constraints. In some cases, minisatellite length variability has been incorporated into protein function, for example within the coding region of the *per* gene of

Drosophila melanogaster (Costa *et al.*, 1992). In others, the changes in repeat number of a coding minisatellite can give rise to genetic disease, for example changes in the number of repeats in the prion protein gene that have been implicated as the cause of a number of cases of inherited prion disease such as Creutzfeldt-Jakob disease (Palmer & Collinge, 1993). The length changes in some minisatellites may influence neighbouring gene function, possibly by altering the number of binding sites for various transcription factors (Krontiris *et al*, 1993). For example, length changes in the HRAS-1 minisatellite may result in over-expression of the downstream proto-oncogene which contributes to several common human cancers (Kasperczyk *et al*, 1990; Krontiris *et al*, 1993). Similarly, at the insulin locus, alleles of a certain size range have been associated with insulin-dependent diabetes mellitus (Lucassen *et al*, 1993). Again, probably because the allelic state of the minisatellite influences the level of transcription of the insulin gene (Bennett *et al.*, 1995).

Minisatellites and recombination

This is reviewed in more detail in Chapter 5. Initial observations made following the isolation of the first GC-rich minisatellites was that these loci represented recombination hotspots (Jeffreys *et al.*, 1988). This seemed feasible because the repeats that make up the minisatellite arrays possess an 11 - 18 bp "core" sequence, which shows homology with the *chi* recombination signal of *E. coli.* Additionally, minisatellites are located in the highly recombinogenic subterminal regions of the genome. However, this theory became less likely with the discovery of AT-rich minisatellites (Huang & Breslow, 1987; Vergnaud *et al.*, 1991), and GC-rich minisatellites which showed little homology to the core sequence (Jeffreys, 1987). Detailed examination of the basis of minisatellite instability revealed that minisatellite mutation occurred by a gene conversion-like process which copied information from one allele into another. This left the donor allele unchanged while the recipient was often quite dramatically altered. Mutation was also highly polarised towards one end of the repeat array. This provided some evidence of recombination at minisatellites, although evidence of crossing over (reciprocal recombination) remained elusive (Wolff *et al.*, 1988, 1989). Only recently has crossing over been associated with minisatellites, MS32 (Jeffreys *et al.*, 1998a & b), MS31a (Jeffreys *et al.*, 1998a) and CEB1 (J. Buard, pers. comm.). At MS32 there is a crossover hotspot located within the flanking DNA adjacent to the unstable end of the minisatellite array, reviving speculation that minisatellites may be involved in chromosome homologue recognition and synapsis (Jeffreys *et al.*, 1998a).

To investigate the basis of germline instability

This work examines the molecular basis of unstable sequences, with a particular interest in recombination in the male germline. Studies at MS32 have demonstrated that crossover activity is concentrated in the flanking DNA at the end of the minisatellite array associated with the gene conversion hotspot. Pilot studies at MS31a have also shown that crossovers can occur within the end of the array showing the highest concentration of conversion activity (Jeffreys *et al.*, 1998a).

This suggests that MS31a may be an ideal candidate for additional recombination studies. The polarity of gene conversion towards one end of the minisatellite has meant that MS31a has been well characterised at this end of the locus. However, very little is known about mutation at the opposite, or more stable end of the locus. This was therefore investigated by mapping minisatellite allele structures in individuals from different populations from the 3′ end of the locus (Chapter 3). Polymorphic markers flanking both ends of the minisatellite were identified by detailed sequencing, these were genotyped and used to determine the relationship between minisatellite repeat array structure and the flanking DNA. The flanking markers were also examined by comparative analysis to investigate linkage disequilibrium, and thus recombination within the flanking DNA (Chapter 4). These markers were then used to isolate and characterise *de novo* crossover mutants in sperm DNA (Chapter 5). Finally, because minisatellites are not a representative sample of mutable loci in the genome, these studies were extended to investigate recombination between Alu repeats in the male germline (Chapter 6). Each chapter is introduced with a discussion of the background and, where appropriate, the methods used; the results are then detailed and discussed in depth, along with any appropriate further work. A separate chapter is included with details of general materials and methodologies used (Chapter 2). The final chapter summarises the important findings and suggests future directions that may be expanded from this work (Chapter 7).

It must be noted that the designation of upstream (5′) and downstream (3′) ends of a minisatellite are purely arbitrary and serve only to orientate features of a minisatellite within that locus with no relationship to minisatellites at different loci.

Chapter 2

Most methods used during this work are adequately described in standard laboratory manuals (e.g. Sambrook *et al.*, 1989). Below are a brief description of protocols, more detailed descriptions of work carried out are found within the relevant results chapters.

Materials

Consumables and hardware. All chemicals, reagents and plasticware used were standard and purchased from established suppliers of molecular biology reagents (Applied Biotechnologies Ltd, Boehringer Mannheim, Sigma Biochemical Co. *etc.*). Custom built gel tanks were constructed in-house.

Oligonucleotides. Oligonucleotides for PCR amplification were synthesised by the Protein and Nucleic Acid Chemistry Laboratory, University of Leicester. They were prepared for use either by ethanol precipitation using 3 M sodium acetate (pH 7.0) or by N-butanol precipitation (Sawadogo & Vandyke, 1991), and dissolved in PCR-clean water.

DNA. DNAs from lymphoblastoid cell lines derived from 40 large Caucasian families were supplied by H. Cann and J. Dausset of the Centre d'Etude du Polymorphisme Humain (CEPH, Paris, France). Japanese blood samples were donated by Y. Katsumata (Nagoya University, Japan). West African DNAs either originated from Nigeria, or were Zimbabwean sperm samples donated by A .D. Nakomo and S. B. Kanoyangwa (Forensic Science Laboratory, Causeway, Zimbabwe). Caucasian sperm DNAs were donated by J. Blower (Leicester Royal Infirmary) and from members of the Department of Genetics (University of Leicester).

Methods

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

Measuring DNA concentration. DNA concentrations were estimated by visual comparisons of signal intensity against DNA of known concentration following gel electrophoresis. Oligonucleotide primer concentrations were measured more accurately by measuring the optical density at a wavelength of 260 nm (OD_{260}) of 3 dilutions (D) of the oligonucleotide (1 in 1000, 2 in 1000, 3 in 1000) in a Cecil Instruments CE 202 Ultraviolet Spectrophotometer. The average of these readings was used to calculate the concentration (C),

based on the approximation that 1 OD unit is equivalent to 33 μg/ml oligonucleotide ($C = OD_{260} \times D \times 24$). The dilution factor (DF) needed to make 10 μM primer stocks was calculated using the approximation that the mass of a single deoxyribonucleotide is 330 kD (DF = C/[primer length x 330 x 0.01]). Genomic DNA concentrations were also measured by spectrophotometry, or more accurately using a DNA Fluorimeter TK0 100 (Hoefer Scientific Instruments). The fluorimeter was calibrated with 500 ng fragmented herring sperm DNA in 1x TNE (100 mM Tris-HCl pH 7.4, 10 mM EDTA pH 8.0, 1 M NaCl), 0.1 μg/ml Hoechst dye 33258 (Sigma Biochemical Co.). Three different aliquots (1, 2 and 3 μl) of digested genomic DNA were added to 1 ml of the above solution and the DNA concentration was directly measured. For more accurate calculations of the number of amplifiable molecules within a genomic DNA sample, a Poisson analysis was used. DNA was sequentially diluted to the single molecule level (calculated using the previous methods) and amplified under optimal conditions for the primers used with high cycle number (28 - 30). Gel electrophoresis and Southern hybridisation of the amplified samples allowed the number of negative and positive reactions to be counted as a binary code. The true, or Poisson corrected number of amplifiable molecules was then calculated using the χ^2 test on the number of negative reactions. This was also used to correct for the possibility of more than one identical length change mutant being present in the same PCR reaction. So if N PCR reactions yield R reactions positive for a length change mutant (or progenitor molecule), the Poisson corrected number of length change mutants (or progenitor bands) is given by -Nln [(N-R)/N].

DNA manipulation. General methods for handling DNA, gel electrophoresis, Southern blotting *etc.* were performed as described previously (Sambrook *et al.*, 1989). Enzymatic manipulations were used according to the manufacturer's instructions with the supplied buffers.

Gel extraction. DNA to be extracted was electrophoresed under normal conditions (Sambrook *et al.*, 1989) and the fragment of interest was excised from the gel in an agarose block. DNA was then extracted by three different methods.

1. If the agarose block was small and the DNA concentration low, the agarose block was crushed with a disposable pipette tip, frozen at -80°C, crushed again and spun for 30 min at 13,000 rpm. The supernate was removed and ethanol precipitated.

2. DNA of molecular weight less than 2 kb was extracted using spin columns. These consisted of a 0.5 ml microcentrifuge tube 1/4 filled with "Supa" Filter Wool which was pre-washed in 150 μl TE pH 8.0. A hole was made in the bottom of this tube and the agarose block placed into the tube, centrifuged at 13,000 rpm for 30 min and collected in a 1.5 ml microcentrifuge tube. The eluate was then ethanol precipitated.

3. DNA of any molecular weight was extracted by electroelution onto dialysis membrane. The dialysis membrane was cut about 4 mm larger than the agarose block and boiled in PCR clean water for 5 min. A rectangular slot was cut in a high percentage agarose gel to fit the dimensions of the agarose block. The block was placed into this and the DNA was electrophoresed onto the dialysis membrane at 10-15 volts/cm for 1-10 min, depending on the size of the DNA fragment. With the current still running, the membrane was swiftly removed

using a pair of tweezers. This was placed into a 1.5 ml microcentrifuge tube with one corner trapped under the lid allowing the DNA solution to be collected at the bottom of the tube by centrifugation for 13,000 rpm for 1 min. The membrane was then removed and the DNA purified by ethanol precipitation.

PCR Reactions. Tables 2.2 to 2.5 give all the details necessary to carry all PCR reactions referred to in each chapter; primer sequences are listed in Table 2.6. Amplifications were carried out in 7 µl reactions containing 1x PCR buffer (45 mM Tris-HCl (pH 8.8), 11 mM $(NH_4)_2SO_4$, 4.5 mM $MgCl_2$, 6.7 mM β-mercaptoethanol, 4.4 µM EDTA (pH 8.0), 1 mM dATP, 1 mM dCTP, 1 mM dGTP, 1 mM dTTP, 113 µg/ml BSA). PCR machines used include the Geneamp 9700™ and 9600™ thermal cyclers (Perkin Elmer) or the PT-200 DNA Engine (MJ Research). 0.05 units/µl of thermostable DNA polymerase was used under normal amplification conditions. Long-range PCR (Barnes, 1994; Cheng *et al.*, 1994; Michalatos-Beloin *et al.*, 1996) conditions were used for amplification of regions over approximately 2 kb, particularly highly repetitive regions. This protocol used a 20:1 mixture of 0.7 units of Ampli*Taq* (Perkin Elmer) and 0.035 units of *Pfu* (Stratagene) polymerases. *Pfu* has a 3′ to 5′ exonuclease activity which can edit regions of depurinated DNA that would otherwise cause the Ampli*Taq* to stall, resulting in incomplete products and allowing jumping PCR to occur. Where indicated, herring sperm DNA (5 µg/ml) was added to the reaction to preferentially coat the side of PCR tubes and prevent the target DNA being sequestered in this way. This was particularly important when using low concentrations of target DNA.

Automated sequencing of the 3′ flanking region of MS31a. The flanking region to be sequenced was pre-amplified using primer pairs 31HHR and 31O, 31IIR and 31V (Table 2.2). Following agarose gel electrophoresis the amplified DNA from each individual was recovered by electroelution on to dialysis membrane followed by ethanol precipitation, and dissolved in 10 µl water. DNA was then sequenced following the manufacturer's recommendations using *Taq*FS with the Big Dye Terminator Kit (Perkin Elmer). Half the recommended reaction mix was used and reactions were scaled down accordingly. Sequencing products were purified by a modified ethanol precipitation protocol: 0.1 volume 3M sodium acetate pH 4.6 and 2.5 volumes 95% ethanol were added to the finished reaction, followed by a 10 min incubation on ice. The mixture was centrifuged for 20 min at 13000 rpm at room temperature, the pellet was washed with 70% ethanol and vacuum dried.

RFLP typing of all flanking variant sites. Amplification and restriction enzyme digestion conditions are shown in Table 2.6. PCR conditions were appropriate for use in either a Geneamp 9600™ thermal cycler (Perkin Elmer) or a PT-200 DNA Engine (MJ Research). Unless otherwise stated, 7 µl PCR reactions contained ~100 ng genomic DNA, 0.63 µl 11x PCR buffer, 0.25 µM primers and 0.25 units of Taq polymerase. PCR products were digested by adding 3 units of the appropriate enzyme with 1 µl of the manufacturer's recommended 10x reaction buffer and 1 µl 10 mM spermidine trichloride, unless otherwise indicated. The

digestion products were resolved by electrophoresis through a 3% MetaPhor™ (Sigma Biochemical Co.) agarose gel in 0.5x TBE buffer and visualised by ethidium bromide staining.

Haplotyping of MS31a flanking sequence. The minisatellite array was amplified using allele-specific primers 31-1663C/T and 3′ universal primer 31V, this was diluted 100 fold and used to seed subsequent PCR reactions. Sites -221 and -109 were typed in a multiplex reaction using 31ER and allele-specific primers (31Psp+/- and 31Hga+/-) for these sites; sites -759 and -457 were typed simultaneously by RFLP analysis; 3′ sites were typed initially by allele-specific PCR between primers 31+1165G/T and 31LR followed by 10x dilution and reamplification between 31+744C/G and 31LR; the -4 site was typed by RFLP analysis; and site -1172 was typed directly from genomic DNA by RFLP analysis. Individuals homozygous at the -1663 site were typed by nested allele specific PCR with the primers 31-759A/C and 31N followed by subsequent typing as described above. Any sites that could not be assigned a phase were reamplified by step-down PCR using allele-specific primers of known phase followed by RFLP analysis of the ambiguous site.

DNA hybridisation. DNA was depurinated, alkali-denatured and transferred from agarose gels to Hybond-N FP (Amersham) membrane by Southern blotting for 1 hour. The DNA was then crosslinked to the membrane by exposure to UV radiation at 10,000 microjoules/cm^2 in a UV crosslinker RPN 2500/2501 (Amersham). 10 ng probe DNA was labelled overnight at room temperature, or for 2 - 3 hr at 37°C by the random hexamer priming method (Feinberg & Vogelstein, 1983) incorporating α-^{32}P-ATP. The labelled probe was recovered by ethanol precipitation using 100 µg high molecular weight herring sperm DNA as carrier, and boiled for 3 min immediately prior to use. Filters were pre-hybridised at 65°C in 0.5 M sodium phosphate (pH 7.2), 7% SDS, 1 mM EDTA (modified Church and Gilbert solution, Wong *et al.*, 1987) in a Hybaid rotating bottle oven. Hybridisations were carried out under the same conditions for between 4 - 18 hr. Filters were washed at high stringency in 0.1 x SSC, 0.01% SDS and dried prior to autoradiography. Filters were exposed to Fuji RX100 X-ray film for between 1 hr to 1 week depending on the signal strength, as described (Sambrook *et al.*, 1989). The MS31a probe was synthesised as described previously (Neil, 1994). Probes for the C1 inhibitor gene and MS32 were amplified as described in Table 2.4, and purified by gel electroelution Membranes were stripped of radiolabelled probes by repeated immersion in 1 mg/ml boiling SDS solution for ~10 min, until the level of radioactivity was at background level.

Computing and bioinformatics. DNA sequences were analysed and processed using Factura™ and AutoAssembler software (Perkin Elmer). Further analysis was carried out using the Genetics Computer Group Sequence (GCG) Wisconsin Package Version 10.0 developed at the University of Wisconsin in Madison, Wisconsin through the IRIX system on the University of Leicester computer network.

Table 2.1 Primer sequences

Primer	Sequence (5′ - 3′)
MS31A primers (Chapters 3 & 4)	
31A	CCCTTTGCACGCTGGACGGTGGCG
31B	CCCACACGCCCATCCGGCCGGCAG
31C	GGCACAACCTAGGCAGGGGAAGCC
31DD	CCCAGCCAGCTGTTCCCATG
31ER	GGACAGCCAAGGCCAGGTCC
31EER	GTACCTGGCCACCGAGTGAG
31F	CCACTCGGAACCACCTGCAG
31FF	CTGAGGTTCAGCCACTTGCC
31GGR	GCCTCACACAGGACTTCAGC
31HHR	ATAGGGCTGGGCTTCAGGAG
31IIR	GAGGGCTTCAGGGAAACCTG
31LR	CAGACACTGCCGGCCGGATG
31M	AGGGATGGCCAGCACCATTC
31N	AGTGGCCCGGTCAACTGCAG
31O	AGCCAAACCCCATAGGCTCC
31QQ	GAGCCAAAGAGTTCGAGACC
31R	AAGTCCCGGGTGAGAGCGTG
31V	CGGCCTAATGGATCCGTCAG
31FspI	CTTTTCCAGATCCTTCTCAATGCGC
31Hga+/-	CCTCCCCCACTCAGCc/g
31Psp+/-	GGAAGCCGCATGCACAAt/c
31PstI	GAGGCAGGGCCCTGGTTGTCCTGC
31RsaI	GGCAGGAAACCCCAGCAGGCCGTA
31-1663T/C	CGCCATGGTGTCCCCACt/c
31-759C/A	GGCGCTCATGCCTGGAAc/a
31-759RG/RT	CCTCCCCAGTTCCTGGGg/t
31-457RA/RG	CAGATCCTTCTCAATGGGCa/g
31-221RC/RG	CCCCAGCAGGCCGGAc/g
31-109RA/RG	TGCAAAGGGAAGAGCCAACa/g
31-ΔH+/-	GGAGCCCCACTCAGCg/c
31+744C/G	TCAGCTGCTTCCAACACACACATc/g
31+1165G/T	GAGCAGGGGTGTCCGg/t
31TAG-A	tcatgcgtccatggtccggaAGTGTCTGTGGGAGGTGGA
31TAG-G	tcatgcgtccatggtccggaAGTGTCTGTGGGAGGTGGG
31TAG-AT	tcatgcgtccatggtccggaAGTGTCTGTGGGAGGTGGAT
31TAGnew-A	tcatgcgtccatggtccggaTCTGTGGGAGGTGGA
31TAGnew-G	tcatgcgtccatggtccggaTCTGTGGGAGGTGGG
31TAGnew-AT	tcatgcgtccatggtccggaTCTGTGGGAGGTGGAT
31rTAG-AT	tcatgcgtccatggtccggaCCACCTCCCACAGACACTAT
31rTAG-GC	tcatgcgtccatggtccggaCCACCTCCCACAGACACTGC
31rTAG-GT	tcatgcgtccatggtccggaCCACCTCCCACAGACACTGT
TAG	tcatgcgtccatggtccgga
MS32 Primers (Chapter 5)	
32-10F	ATGACTTAACCTAGGCCTATCAGTG
32-9.7F	GAGGGAAGTCATAGACAACAGCTG
32-9.2R	CCACCATGTGAGGACAAAGCG
32-6.8F	GACTCATACATGAGCCAAGT
32-6.5R	CACTTCAGCTAGACTACTTC
32-5F	CAGTGCTTGGCACATAATGAGCAC
32-5NF	CTCTTTCTAGAAGCCGTTAGAGGAG
32-5R	GAATCCTACATGTAGGCGAGCAGT
32-4.9R	CTGTGTACAGGAACTTAGGGAACG
32-1.6F	GCCAACAGTGTACTTTGAAGAGCA
32-1.3R	CCAGGTTCTGGGGTGACTAG
32-1.3NR	GCAGGTAGATAGTGGCCCAGAG
C1 inhibitor gene primers (Chapter 5)	
C1+2.4F	CCCAACAGATTCTCCTACCC
C1+2.4NF	CACTACTGGGTCCTTCTG
C1+2.4NFnew	GCCCACTACTGGGTCCTTCTGCCCAG
C1+2.4R	ATACATCCCTCTACCCCACC
C1+6.8F	TGGGTCAAAGGAGTCTTGG
C1+6.8NR	CCCACAATATAAAATGGACCCTG
C1+6.8NRnew	ATGGACCCTGGCTGAGGCTGGGTG
C1+6.9R	GCAGCAGAACTGCTTTCAAAC

Table 2.2 PCR conditions used in Chapter 3

PCR	Primers	Final conc. (μM)	Taq used	PCR mac-hine	Conditions of analysis
Pre-amplification for sequencing	31O 31HHR	0.25 0.25	Taq	MJ	(96°C,1')30[(96°C,30")(65°C,45")(70°C,1')] (70°C,2')
Pre-amplification for sequencing	31V 31IIR	0.25 0.25	Taq	MJ	as above
Reverse MVR-PCR from +744	31+744C/G rTAG-AT rTAG-GC rTAG-GT TAG	0.25 0.01 0.008 0.05 0.25	Taq	9600	(96°C,30") 11[(96°C,30")(70°C,3'30")] 13[(96°C,30")(70°C,3'30"+20")] (70°C,10') Electrophorese at 180v on 1% HGT agarose (Sigma Biochemical Co.) for ~15 hr
Reverse MVR-PCR from +1165	31C 31+1165G/T 31B rTAG-AT rTAG-GC rTAG-GT TAG	0.25 0.25 0.25 0.01 0.008 0.05 0.25	Taq	MJ	Pre-PCR 31C and 31+1165G+T (50 ng genomic DNA) 10[(96°C,30")(66.5°C,30")(70°C,1'30")] MVR PCR (remaining primers) dilute PCR product from above 1+100 (96°C,30") 28[(96°C,30")(70°C,3')](70°C,4') 9600 Electrophorese on 1% HGT agarose (Sigma Biochemical Co.) at 120v for ~15.5 hr.
Allele-specific PCR from +744 to type MS31b	31+744C/G 31LR	0.25 0.25	Taq	MJ	(96°C,1') 28[(96°C,30")(70°C,45')] (70°C,1') Size of PCR product determines the allelic state of MS31b
Allele-specific PCR from +1165 to type MS31b	31+1165G/T 31LR	0.25 0.25	Taq	MJ	(96°C,1') 28[(96°C,30")(66°C,45')] (70°C,1') Size of PCR product determines the allelic state of MS31b

PCR machines: MJ - PT-200 DNA Engine (MJ Research) ; 9600 - Geneamp 9600™ thermal cycler (Perkin Elmer).
DNA polymerases: Taq - Ampli*Taq* (Perkin Elmer); Taq+pfu - 20:1 mixture of Ampli*Taq* (Perkin Elmer) and *Pfu* (Stratagene).

Table 2.3 PCR conditions used in Chapter 4

PCR	Primers	Final conc. (μM)	Taq used	PCR mac-hine	Conditions of analysis
Pre-amplification for haplotyping	31-1663C/T 31V	0.25 0.25	Taq+pfu	MJ	(96°C,1') 25[(96°C,30")(68°C,30")(70°C,5')) (70°C,6') Input 100 ng genomic DNA
Pre-amplification for haplotyping	31-759A/C 31N	0.25 0.25	Taq+pfu	MJ	(96°C,1') 24[(96°C,30")(70°C,5')] (70°C,6') Input 100 ng genomic DNA
Multiplex PCR to type sites -220 and -109	31ER 31Psp+/- 31Hga+/-	0.25 0.3 0.3	Taq	MJ	(96°C,1')20[(96°C,30")(65°C,45"-0.5° per cycle)] 10[(96°C,30")(55°C,45" + 1" per cycle)] Seperate on 1.5% LE agarose (Sigma Biochemical Co.). Pool 1 = 31 ER with 31Psp+ and 31Hga+. Pool 2 = 31 ER with Psp- and Hga-
Stepdown PCR for haplotyping	Any primer Any primer	0.25 0.25	Taq	MJ	(96°C,1')20[(96°C,30")(65°C,45"-0.5° per cycle)] 10[(96°C,30")(55°C,45" + 1" per cycle)]

PCR machines: MJ - PT-200 DNA Engine (MJ Research).
DNA polymerases: Taq - Ampli*Taq* (Perkin Elmer); Taq+pfu - 20:1 mixture of Ampli*Taq* (Perkin Elmer) and *Pfu* (Stratagene).

Table 2.4 PCR conditions used in Chapter 5

PCR	Primers	Final conc. (μM)	Taq used	PCR mac-hine	Conditions of analysis
SP-PCR on genomic DNA	31C 31N	1.0 1.0	Taq+pfu	9600	(96°C,1')22[(94°C,30")(65°C,30")(70°C,5')] (65°C,1')(70°C,10') Electrophorese overnight on a 40 cm, 0.7% LE agarose (Sigma Biochemical Co.) and blot.
Nested PCR on recovered mutants	31A 31M	0.25 0.25	Taq	9600	(96°C,1')30[(94°C,30")(65°C,30")(70°C,5')](65°C,1')(70°C,10']
MVR-PCR of SP-PCR mutants	31A 31TAG-A 31TAG-G 31TAG-AT TAG	0.25 0.01 0.025 0.005 0.25	Taq	9600	(96°C,1') 20[(94°C,30")(65°C,30")(70°C,3'+10"] (65°C,1') (70°C,10') Electrophorese overnight on a 1.2% HGT agarose (Sigma Biochemical Co.) at 120v.
Crossover mutant isolation Step 1	31DD 31V	0.4 0.4	Taq+pfu	9700	(96°C,1')20[(94°C,30")(66°C,30")(70°C,5')](65°C,1')(70°C,10'] 28 cycles were used for Poisson analysis and assessing the degree of enrichment following size fractionation, 20 cycles were used for the crossover detection strategy, both included herring sperm DNA as carrier.
Step 2	31-1663C 31+1165G	2.5 2.5	Taq+pfu	9700	(96°C, 1') 8[(96°C,30")(70°C,4')] 12[(96°C,30")(68°C,4')](70°C,10') 1.5 μl S1 nuclease-treated DNA from step 1 and 1 μg final concentration of herring sperm DNA as carrier.
Step 3	31-759C 31N	0.25 0.25	Taq+pfu	9700	(96°C, 1') 19[(96°C,30")(70°C,4')](70°C,10') 1.5 μl S1 nuclease-treated DNA from step 2 and herring sperm DNA, as above.
Crossover analysis (step 4)	31QQ 31M	0.25 0.25	Taq+pfu	9700	(96°C, 1') 22[(96°C, 30")(67°C, 30")(70°C, 5')] (70°C, 6') Recombinants were amplifed from 0.5μl of S1 nuclease-treated step 3 PCR product.

Determination of MS31b allele size in recombinants	31LR 31M	0.25 0.25	Taq	MJ	(96^0C, 1') 30[(96^0C, 30")(68^0C, 30")(70^0C, 30")] (70^0C, 1') 1/1000 dilution S1 treated step 4 PCR product
MVR analysis of recombinants	31A 31TAGnew-A 31TAGnew-G 31TAGnew-AT TAG	0.25 1.0 0.5 0.1 0.25	Taq+pfu	MJ	(96^0C, 1') 8[(96^0C, 30")(57^0C, 30")(70^0C, 2'30")] 12[(96^0C, 30")(70^0C, 2'30")] (70^0C, 3') 1/100 dilution S1 treated step 4 PCR product was amplified with herring sperm DNA as carrier and electrophoresed overnight on a 1.2% HGT agarose (Sigma Biochemical Co.) at 120v.
Allele-specific MVR analysis of recombinants	Psp+/- 31TAG-A 31TAG-G 31TAG-AT TAG	0.25 0.01 0.025 0.005 0.25	Taq	9600	25 [(94^0C, 30")(66^0C, 30")(70^0C, 3'10")] (66^0C, 1') (70^0C, 10') 2 µl of 1/100 dilution S1 treated PCR product from step 4 was amplified and electrophoresed overnight on a 1.2% HGT agarose (Sigma Biochemical Co.) at 120v.

PCR machines: MJ - PT-200 DNA Engine (MJ Research); 9600 - Geneamp 9600™ thermal cycler (Perkin Elmer); 9700 - Geneamp 9700™ thermal cycler (Perkin Elmer).
DNA polymerases: Taq - Ampli*Taq* (Perkin Elmer); Taq+pfu - 20:1 mixture of Ampli*Taq* (Perkin Elmer) and *Pfu* (Stratagene).

Table 2.4

Table 2.5 PCR conditions used in Chapter 6

PCR	Primers	Final conc. (μM)	Taq used	PCR machine	Conditions of analysis
To locate problematic region MS32	32-10F 32-9.2R	0.1 0.1	Taq+pfu	MJ	(94°C,1') 20[(94°C, 15")(65°C,1')(70°C,4')] (67°C,1')(70°C,10') 0.5 ng cosmid DNA
To locate problematic region MS32	32-9.7F 32-6.5R	0.1 0.1	Taq+pfu	MJ	as above
To locate problematic region MS32	32-6.8F 32-4.9R	0.1 0.1	Taq+pfu	MJ	as above
Amplification across entire target region	32-9.7F 32-1.3R	0.1 0.1	Taq+pfu	MJ	(94°C,1') 30[(94°C, 15")(67°C,30")(70°C,6')] (70°C,6') 200ng genomic DNA. Poor amplification. Herring sperm DNA included as carrier.
MS32 outside PCR	32-5F 32-1.3R	0.4 0.4	Taq+pfu	MJ	(94°,1') 26 [(94°,30")(68°,30")(70°,6')](70°,5') Herring sperm DNA used as carrier during screening.
MS32 semi-nested PCR	32-5F 32-1.3NR	0.4 0.4	Taq+pfu	MJ	(94°C,30) 26[(94°C,15')(68°C,30)(70°C,6')](70°,5') Amplifies to visible level with 100 ng genomic DNA, 18 cycles for 0.1 ng cosmid DNA.
Amplification outside region of interest - MS32	32-6.8F 32-5R	0.4 0.4	Taq+pfu	MJ	30 [(94°C,1')(62°C,5')](70°C,2') Input 200 ng CEPH DNA or 400 ng gorilla or chimpanzee DNA
MS32 probe 1 synthesis	32-5NF 32-5R	0.4 0.4	Taq	9600	20 [(96°C,1')(65°C,1')(70°C,1')](67°C,1)(70°C,10') Input 1 ng cosmid DNA. Products were verified by restriction digestion, following electroelution.
MS32 probe 2 synthesis	32-1.6F 32-1.3NR	0.4 0.4	Taq	9600	As above
C1 outside PCR	C1+2.4F C1+6.9R	0.4 0.4	Taq+pfu	MJ	(94°C,30)32 [(94°C,15")(68°C,5')](70°,5') Herring sperm DNA used as carrier during screening.
C1 nested PCR	C1+2.4NFnew C1+6.8NRnew	0.25 0.25	Taq+pfu	MJ	(94°C,30")28[(94°,15")(70°C,6')[(70°,5')
C1 probe 1 synthesis	C1+2.4NF C1+2.4R	0.25 0.25	Taq	MJ	(96°,1') 30[(96°,1')(62°,1')(70°,1')](65°,1')(70°C,10') Probes were amplified from genomic DNA verified by restriction digestion following electroelution.
C1 probe 2 synthesis	C1+6.8F C1+6.8NR	0.25 0.25	Taq	MJ	as above
C1 SP-PCR (semi-nested)	C1+2.4F C1+6.8NR	0.25 0.25	Taq+pfu	MJ	(94°C,30") 31[(94°C,15")(66°C,30")(70°,6')](70°,5')

Table 2.5

PCR machines: MJ - PT-200 DNA Engine (MJ Research); 9600 - Geneamp 9600™ thermal cycler (Perkin Elmer).
DNA polymerases: Taq - Ampli*Taq* (Perkin Elmer); Taq+pfu - 20:1 mixture of Ampli*Taq* (Perkin Elmer) and *Pfu* (Stratagene).

Table 2.6 RFLP analysis of flanking variants

Site typed	PCR conditions	Digest conditions	Visible band size (bp)	Phenotype	Genotype
-1752T/C	31DD + 31EER 30[(96°C,30")(66°C,30")(70°C,1')]	*Bgl* II	757 757 + 490 + 267 490 +267	*Bgl* II -/- *Bgl* II +/- *Bgl* II +/+	-1752C/C -1752C/T -1752T/T
-1663C/T	31DD + 31EER 20[(96°C,30")(68°C,30")(70°C,30")] dilute 1/10 31DD + 31PstI 28[(96°C,30")(66°C,30")(70°C,30")]	*Pst* I + *Mse* I in NE buffer 2 with 0.1 mg/ml BSA	264 + 119 264 + 119 + 98 264 +98 +21	*Pst* I -/- *Pst* I +/- *Pst* I +/+	-1663C/C -1663C/T -1663T/T
-759A/C and-457T/C	31GGR + 31R 20[(96°C,30")(66°C,30")(70°C,1')] dilute 1/10 31FspI + 31F 30[(96°C,30")(66°C,30")(70°C,30")]	*Nla* IV + *Fsp* I in NE buffer 4	243 + 193 + 90 243 + 193 + 170 + 90 243 + 170 +90 243 + 193 +137 +106 + 90 243 + 193 + 170 +137 + 106 +90 243 + 170 +137 + 106 + 90 193 + 137 + 106 + 90 193 + 170 +137 + 106 + 90 170 + 137 + 106 + 90	*Nla* IV -/-, *Fsp* I -/- *Nla* IV -/-, *Fsp* I +/- *Nla* IV -/-, *Fsp* I +/+ *Nla* IV +/-, *Fsp* I -/- *Nla* IV +/-, *Fsp* I +/- *Nla* IV +/-, *Fsp* I +/+ *Nla* IV +/+, *Fsp* I -/- *Nla* IV +/+, *Fsp* I +/- *Nla* IV +/+, *Fsp* I +/+	-759A/A, -457C/C -759A/A, -457C/T -759A/A, -457T/T -759C/A, -457C/C -759C/A, -457C/T -759C/A, -457T/T -759C/C, -457C/C -759C/C, -457C/T -759C/C, -457T/T
-457T/C in recombinants	31FspI + 31QQ 30[(96°C,30")(66°C,30")(70°C,30")]	*Fsp* I	290 267	*Fsp* I - *Fsp* I +	-457C -457T
-220G/C	31F + 31RsaI 30[(96°C,30")(68°C,30")(70°C,30")]	*Rsa* I	206 206 + 183 183	*Rsa* I -/- *Rsa* I +/- *Rsa* I +/+	-221C/C -221G/C -221G/G
ΔH deletion assay	as above	no digestion	206 206 + 197 197	Δ- Δ+/Δ- Δ+/Δ+	full length alleles heterozygous deleted alleles
-109C/T	31ER + 31F 30[(96°C,30")(66°C,30")(70°C,30")]	*Acl* I (was *Psp* 1406I) with 0.1 mg/ml BSA	406 406 + 293 + 113 293 + 113	*Acl* I -/- *Acl* I +/- *Acl* I +/+	-109T/T -109C/T -109C/C
-4A/G	31TAG-A + 31A (1 μM final concentration) 35[(96°C,30")(70°C,30")]	*Alu* I	137 137 + 96 96	*Alu* I -/- *Alu* I +/- *Alu* I +/+	-4G/G -4A/G -4A/A

+744G/C	31O + 31HHR 30[(96°C,30")(66°C,30")(70°C,1')]	*Nla* III	310 + 217 +171 + 138 + 123 + 118 310 + 217 + 171 + 169 + 138 + 123 + 118 310 + 171 +169 + 138 + 123 + 118	*Nla* III -/- *Nla* III -/+ *Nla* III +/+	+744G/G +744G/C +744C/C
+1165C/A	31O + 31 HHR 30[(96°C,30")(66°C,30")(70°C,1')]	*Msp* I	207 + 182 + 153 + 115 + 112 + 62 207 + 182 + 153 + 115 + 112 + 96 + 62 207 + 182 + 115 + 112 + 96 + 62	*Msp* I -/- *Msp* I +/- *Msp* I +/+	+1165A/A +1165C/A +1165C/C

Table 2.6

Chapter 3

Reverse MVR mapping and allele-length analysis of the human minisatellite, MS31a

Summary

Previous analysis of repeat variation in the hypervariable minisatellite MS31a was limited to the 5′ end of the tandem repeat array (Neil, 1994). Here, this analysis has been extended by the analysis of minisatellite allele-length, and the examination of variation using minisatellite variant repeat mapping by PCR (MVR-PCR) at the 3′ end of the repeat array. Analysis of minisatellite allele length revealed two distinct groups. Short alleles of between 50 and 200 repeats appeared to predominate in the African and Afro-Caribbean populations. Conversely, longer alleles of between 200 and 300 repeats were more prevalent in the Caucasian and Japanese populations. This distribution may reflect an ancient founder effect in the non-African populations. Allele structures were then examined in more detail by MVR-PCR at the 3′ end of the minisatellite array. Single-allele MVR codes were derived using allele-specific primers complementary to two single nucleotide polymorphisms (SNPs) identified in the 3′ flanking DNA of MS31a. Allele-specific MVR-PCR was applied to a number of unrelated individuals from Caucasian, Japanese, African and Afro-Caribbean individuals to compile a single-allele databank of MVR codes. All alleles mapped had different structures, demonstrating that high levels of instability must exist at this minisatellite. Out of the 72 alleles mapped, 45 were assigned to 10 groups on the basis of shared structures. These alignments revealed that there is no polarisation of variation within the repeat array, which contradicts mutation studies that indicate mutation is generally clustered at the 5′ end of the repeat array. It was also notable that all but one short allele could be aligned, generally with other alleles of similar size from the same population. Further investigation into these features will be required to understand more about the mutation processes that operate at this minisatellite. These, as indicated by the heterogeneity of these alleles, are likely to be complex.

Introduction

Polymorphism at human minisatellites was first observed in the form of allele-length variation between individuals. This can be can be detected within pedigrees at the most variable loci (Wong *et al.*, 1987). This allelic variation results from changes in tandem repeat copy number due to high rates of spontaneous germline mutation. These have been estimated to be as high as 13% per gamete in humans (Vergnaud *et al.*, 1991). These allele-length changes have facilitated the isolation of minisatellite mutants from sperm DNA, using techniques such as SP-PCR, which have allowed germline instability at these loci to be characterised in great detail (Chapter 5; Jeffreys *et al.*, 1990; May *et al.*, 1996; Buard *et al.*, 1998). These studies have shown that

during germline mutation, minisatellites have a tendency to gain repeats. However, this expansion is not indefinite, suggesting that minisatellite length is regulated, and that this regulation is likely to be intrinsic to the process of mutation. Exactly how this level of control is exerted is unclear but its discovery may be important for a better understanding of tandem repeat biology, and may also be relevant with respect to the gross expansions that cause triplet repeat disorders, such as Fragile X syndrome (Nelson, 1993).

All human minisatellites examined to date also exhibit variation in the interspersion pattern of different repeat units along the tandem array. The order of these repeats can be mapped using a technique called MVR-PCR which has revealed levels of allelic diversity far greater than could be resolved by allele-length measurement (Jeffreys *et al.*, 1991a). MVR-PCR uses a fixed primer in the DNA flanking the minisatellite coupled with primers specific to the different minisatellite variant repeats (Figure 3.4A). This generates a ladder of different size fragments from which the order of the different repeats along the array can be read in a similar way to a sequencing gel. This technique can be further refined by using an allele-specific primer in the minisatellite flanking DNA to map single alleles in individuals heterozygous for particular flanking sites. By analysing the structure of multiple alleles within human populations in this way, inferences about mutation events at these loci can be made. MVR-PCR can also be used for the detailed dissection of mutation processes by structural comparison of progenitor and mutant alleles (Chapter 5; Jeffreys *et al.*, 1991a; May *et al.*, 1996; Buard *et al.*, 1998). These studies have revealed that population variability and mutation of the minisatellite is generally polarised towards one end of the repeat array. This, along with the fact that most minisatellites are too long to characterise in their entirety by MVR-PCR has meant that most of this analysis has been carried out on the most variable end of minisatellites. Little is therefore known about the more stable end of these tandem repeat arrays.

MS31a

MS31a is a "typical" hypervariable minisatellite which has been examined in detail by pedigree analysis. It has been localised to 7p22-pter which, like most other human hypervariable minisatellites is in the subterminal regions of the human genome (Royle *et al.*, 1988). The extreme level of allele-length variation observed at this minisatellite following pedigree analysis (Wong *et al.*, 1987; Armour *et al.*, 1989b) is due to a 1% germline mutation rate to new length alleles (Jeffreys *et al.*, 1988). MS31a allele sizes can range between 42 and 580 repeats with an average size of 220 repeats, similar to the range of array length at MS32 (Neil, 1994). In addition, there also appears to be evidence for a subset of short alleles, of ~50 repeats present in African and Japanese populations which share very similar structures. This is suggestive of increased stability in these alleles (Huang *et al.*, 1996). This may be due to suppression of mutation by flanking variants, as demonstrated with the O1C flanking variant at MS32 (Monckton *et al.*, 1994). Alternatively, the short length of these arrays may be below the threshold of instability, as observed some minisatellites such as CEB1 (Buard *et al.*, 1998) and B6.7 (Tamaki *et al.*, 1999). Details such as these can only be resolved by in depth examination of the effect of minisatellite allele length on mutation rate at MS31a.

Upstream of MS31a, three SNPs have been identified at -220, -109 and -4 bp from the start of the repeat array. Primers complementary to these sites have been routinely used for single-allele analysis of the 5′ end of MS31a. Fourteen nucleotides downstream of MS31a is the relatively stable bi-allelic minisatellite MS31b (Armour *et al.*, 1989b). This minisatellite is heterogeneous in both length and structure in human, chimpanzee and orang-utan (Clarke, 1997). In humans the smaller of the two alleles is made up of three different repeat types, one of 19 bp repeat units, one of 64 bp repeats and a third of 26 bp repeats.

MVR-PCR at MS31a

MS31a is particularly amenable to MVR-PCR because the repeat units are all 20 bp in length (Wong *et al.*, 1987). In addition, the internal heterogeneity conveniently consists of two adjacent polymorphic sites, G or A followed by C or T (Figure 3.4A). Initially, an MVR-PCR system was established at the 5' end of the array because of the potential difficulties of a 3′, or reverse mapping system due to the proximity of MS31a and MS31b. The original system used two MVR-specific primers, 31-TAG-A and 31-TAG-G designed to MVR map T and C variant repeats at the 5′ end of MS31a alleles (Neil & Jeffreys, 1993). This was soon extended to a four-state mapping system (Tamaki *et al.*, 1993) by using the internal G/A site and increasing the informativeness of the system. However, because of the relatively low frequency of AC variant repeats within the array a three-state mapping system is more often used (Neil, 1994). Single-allele MVR-PCR uses three-state mapping in conjunction with allele-specific primers complementary to three SNPs (-220, -109, -4) identified upstream of the minisatellite (Neil & Jeffreys, 1993; Huang *et al.*, 1996). This allele-specific MVR-PCR has been used to construct a single-allele databank from population diversity studies and pedigree analysis. Some alleles could be grouped according to shared structures at this end of the array, although the majority of alleles showed insufficient similarity to be aligned (Neil, 1994). This suggests that high levels of instability may be present throughout the entire length of the MS31a repeat array. MVR-PCR has also been used to analyse mutant and progenitor alleles identified in pedigrees on the basis of length differences (Neil, 1994; Jeffreys *et al.*, 1998b). This has demonstrated that mutation is polarised towards one end (the 5′ end) of the repeat array, and that it is dominated by gene conversion events. These features are typical of minisatellite mutation processes.

MVR-mapping allows alleles to be mapped a maximum of 70 repeats (1.4 kb) into the array, and because most MS31a alleles are larger than this, they cannot be mapped in their entirety. This means that very little is known about the 3′ end of the minisatellite repeat array. In addition, the polarised nature of minisatellite mutation has meant that most of the analysis has been concentrated at the 5′ end of the repeat array. The disparity observed between population studies and mutant analysis concerning the polarity of mutation and/or variation in the repeat array has also not been resolved. However, one pedigree mutant has been isolated where the mutation breakpoint could not be identified by MVR-mapping the 5′ end of the minisatellite array. This implies that although mutation is concentrated at the 5′ end of the array, the 3′ end of the array may also show some degree of instability. It is not clear whether mutational events that

occur much deeper within the minisatellite array are the same as those identified at the 5′ end. Alternatively, they may be more reminiscent of the non-polar, simple deletion and duplication events observed in somatic cells (Jeffreys & Neumann, 1997; see Chapter 5).

This work

MVR mapping and allele-length analysis can be used to study variation at and around minisatellites and may offer insights into the evolution and the mutation processes that contribute to this variation. This work attempts to assess population variability in detail at the human minisatellite, MS31a and provide background knowledge for in-depth mutation studies at this locus. This work examines the distribution of different allele lengths in the Caucasian, Japanese, African and Afro-Caribbean populations. The structure of the 3′ end of MS31a has so far been inferred from MVR maps that do not extend beyond about 70 repeats into the array (Neil, 1994). For completeness these inferences should be confirmed by direct analysis of this end of the minisatellite. Allele structure can then be compared between and within four populations (Caucasian, Japanese, African, and Afro-Caribbean) to make inferences about mutation at this end of the repeat array.

Results

MS31a allele size

Instability at minisatellites was initially identified by changes in allele length. However, little is actually known about the range of allele length in different populations. In addition, previous work has demonstrated the presence of short alleles at high frequency within the Japanese population (Huang *et al.*, 1996). This work has now been extended to include Caucasian, African and Afro-Caribbean populations.

The allele size of MS31a was determined in 44 African, 17 Afro-Caribbean, 37 Japanese and 241 Caucasian individuals. This was achieved by MVR-analysis of short (< 2 kb) alleles, by Southern blot length analysis directly on genomic DNA, and by allele-specific amplification across the minisatellite array using primers 31-1663C/T and 31V. Allele sizes were grouped into bins of fifty repeats and the frequencies of each bin in all four populations were compared (Figure 3.1). The African and Afro-Caribbean alleles show a similar size distribution, mainly consisting of short alleles of between 100 - 200 repeats. Likewise, the Japanese and Caucasian alleles also share a similar size distribution (Kolmogorov-Smirnov test, $p > 0.01$), consisting of alleles which are generally about hundred repeats longer (between 200 and 300 repeats) than the majority of African and Afro-Caribbean alleles (Kolmogorov-Smirnov test, $p > 0.01$). However, the differences between these two population sets are significant (Kolmogorov-Smirnov test, $p < 0.01$). The Japanese population also shows an additional peak at between 50 and 100 repeats, a pattern which has been observed before in this population (K. Tamaki, pers. comm.), and may consist of the subgroup of short alleles discussed by Huang *et al.* (1996). There are very few alleles over about 400 repeats, suggesting that there is a maximum threshold

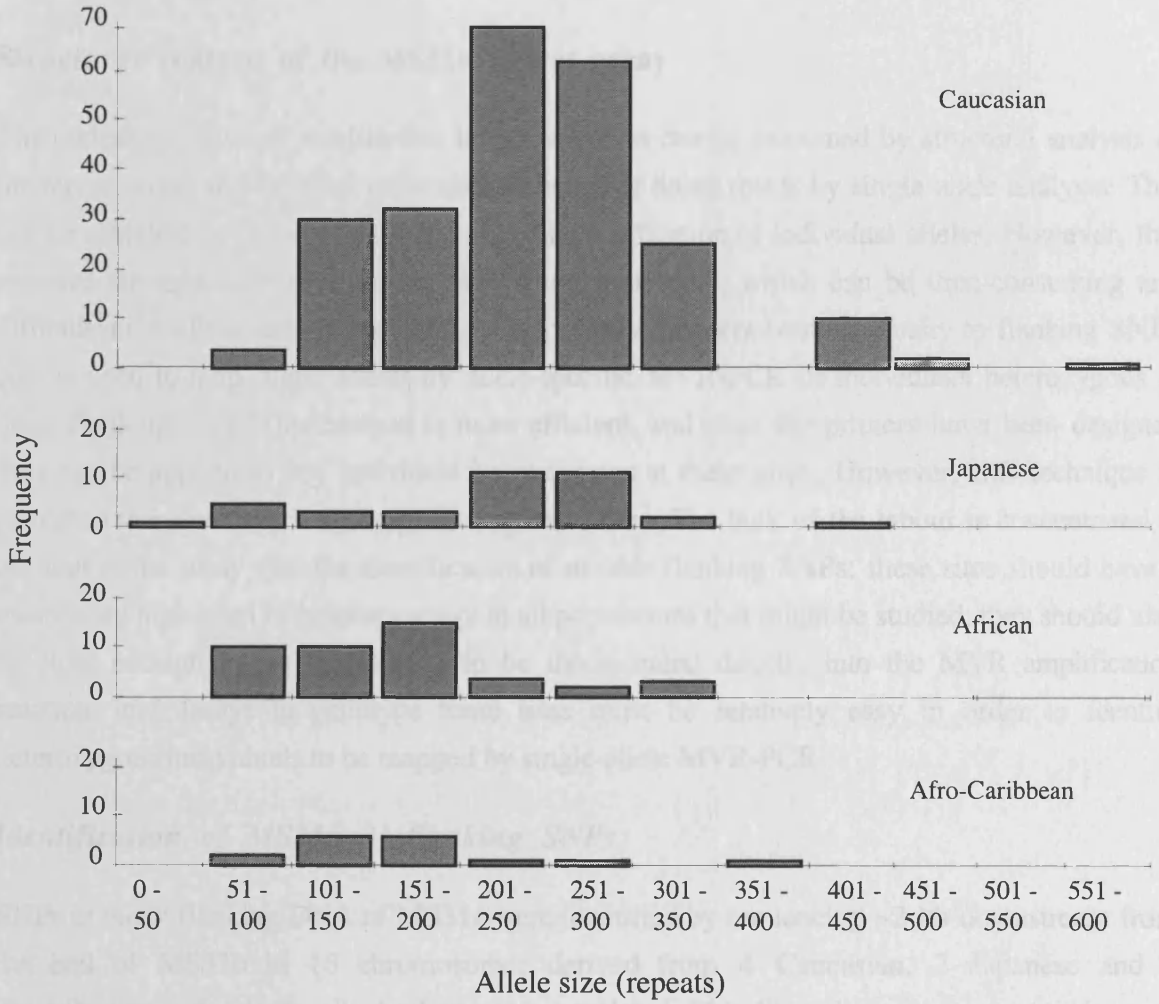


Figure 3.1. Distribution of MS31a allele sizes.

Frequency of allele length of 241 Caucasian, 37 Japanese, 44 African and 17 Afro-Caribbean alleles. Allele size (to the nearest 5 repeats) was determined following MVR analysis of alleles < 2 kb, by Southern blot length analysis directly on genomic DNA, or from PCR amplification across the minisatellite array. The variable size of MS31b was taken into account when determining MS31a array length, where the size of MS31b was unknown the smallest MS31b allele size was assumed. Allele sizes were categorised into 50 repeat bins for this analysis.

of minisatellite size which may be maintained by deletions of large numbers of repeat units (Gray & Jeffreys, 1991). The explanation for these population-specific differences is far from clear, although it may be closely linked to mutation processes.

Structural analysis of the MS31a repeat array

The underlying basis of minisatellite length variation can be examined by structural analysis of the repeat array, and the most informative method of doing this is by single allele analysis. This can be achieved by physical size separation and purification of individual alleles. However, this requires the separation of alleles in each study individual, which can be time-consuming and difficult if the alleles are similar sizes. Alternatively, primers complementary to flanking SNPs can be used to map single alleles by allele-specific MVR-PCR in individuals heterozygous at these flanking sites. This method is more efficient, and once the primers have been designed they can be applied to any individual heterozygous at these sites. However, this technique is restricted to individuals heterozygous at specific sites. The bulk of the labour is concentrated at the start of the study with the identification of suitable flanking SNPs: these sites should have a reasonably high level of heterozygosity in all populations that might be studied; they should also be close enough to the minisatellite to be incorporated directly into the MVR amplification reaction; and assays to genotype these sites must be relatively easy in order to identify heterozygous individuals to be mapped by single-allele MVR-PCR.

Identification of MS31a 3' flanking SNPs

SNPs in the 3′ flanking DNA of MS31a were identified by sequencing ~2 kb downstream from the end of MS31b in 16 chromosomes derived from 4 Caucasian, 2 Japanese and 2 Zimbabwean individuals, all of whom were unrelated. This limited screen was carried out to identify polymorphic sites present at reasonably high frequency in some, if not all populations. Comparison of this sequence data revealed two sites of base substitutional polymorphism. One was a C/A transition, located 1165 bp 3′ from the end of MS31b (+1165C/A) which gives rise to a *Msp* I RFLP. The other was a C/G transversion located 744 bp from the end of MS31b (+744C/G) which generates a polymorphic site for the enzyme *Nla* III. Including the previously characterised dimorphic minisatellite MS31b, a total of three polymorphic sites have been identified in the 3′ flanking DNA of MS31a.

Determining the genotype of the flanking polymorphisms

A simple RFLP assay was developed to genotype both the +744 and +1165 polymorphic positions (Table 2.6, Materials and Methods). In both assays a region of flanking DNA spanning the site to be genotyped was first amplified directly from genomic DNA by PCR, then digested with the appropriate diagnostic restriction enzyme (Figure 3.2). Unfortunately, at position +744 this assay was incapable of distinguishing heterozygous individuals (+744C/G) from those homozygous for the +744G variant (+744G/G). A PCR-based approach was therefore developed which used allele-specific primers complementary to this site (31+744C or

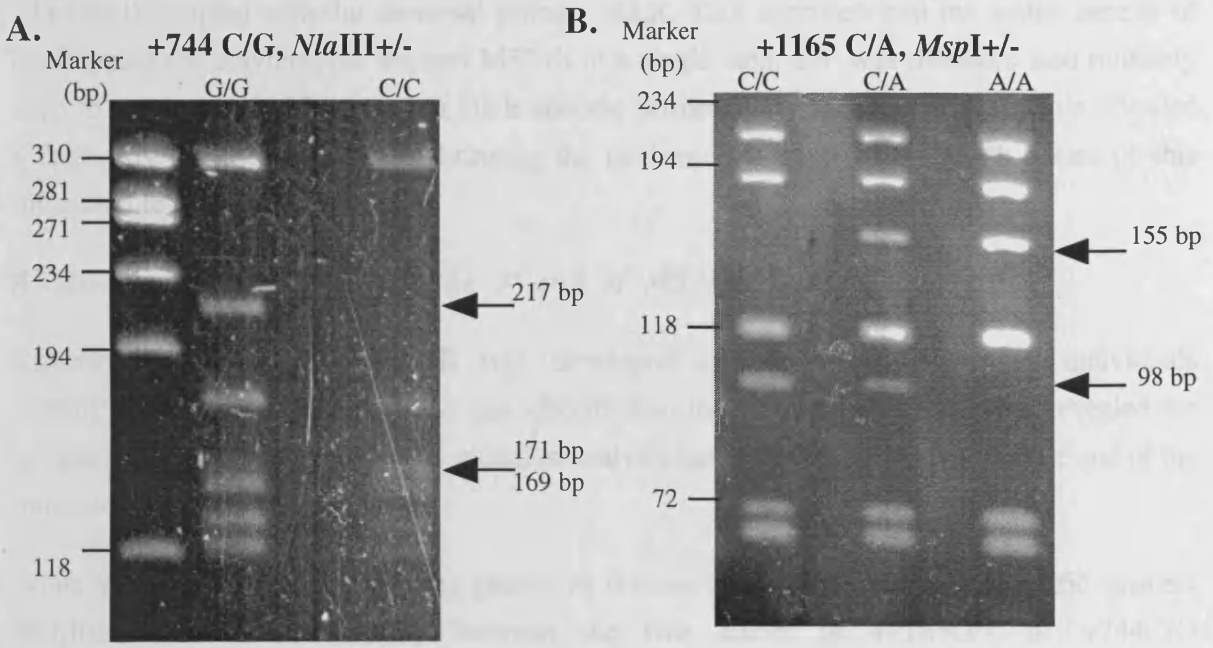


Figure 3.2. Assays for the two MS31a 5´ flanking polymorphisms.

All assays use PCR amplification of a DNA fragment spanning the site to be genotyped, followed by digestion with a diagnostic restriction enzyme (Materials and Methods).

A. +744 C/G, *Nla* III +/- polymorphism assay. The 218 bp PCR product is cleaved into two fragments of 169 and 48 bp (not seen) by *Nla* III if the +744C variant is present. The 169 bp fragment cannot be distinguished from the 171 bp fragment normally present following electrophoresis so +744C/G heterozygotes cannot be distinguished from +744C/C homozygotes.

B. +1165 C/A, *Msp* I +/- polymorphism assay. The 155 bp PCR product is cleaved into two fragments of 98 bp and 57 bp (not seen) by *Msp* I if the +1165C variant is present. +1165C/A heterozygotes will have both fragments of 155 bp and 98 bp in addition to the original 155 bp fragment, and can be distinguished from +1165C/C and +1165A/A homozygotes.

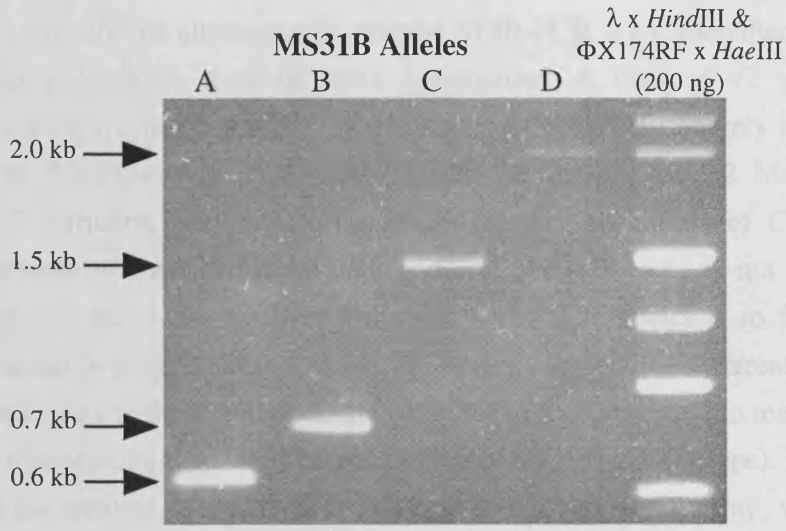


Figure 3.3. The four allelic states of MS31b.

Amplification of genomic DNA from different individuals using primers 31LR and 31M revealed four different sizes of MS31b. The larger of the four alleles, C and D are only found in Japanese and African populations. Changes in allele size are thought to be due to changes in repeat number of the minisatellite.

31+744G) coupled with the universal primer, 31LR. This approach had the added benefit of haplotyping the polymorphic site and MS31b in a single step, and was therefore also routinely used to assay the +1165 site, using allele-specific primers 31+1165C/T. This analysis revealed a further two alleles of MS31b, bringing the total number of different length alleles of this minisatellite to four (Figure 3.3).

Allele-specific MVR-PCR at the 3′ end of MS31a

Reverse allele-specific MVR-PCR was developed to map single alleles in individuals heterozygous at either polymorphic site identified in the 3′ flanking DNA. This revealed the structure of the repeat array at what mutation analysis has identified as the more stable end of the minisatellite.

When employed as the 3' flanking primer in reverse MVR-PCR, the allele-specific primers discriminate almost completely between the two alleles in +1165G/T or +744C/G heterozygotes. This allows selective mapping of either MS31a allele from total genomic DNA (Figure 3.4 B & C). Individuals heterozygous at position +744 were analysed directly by three-state reverse MVR mapping (Table 2.2, Materials & Methods). However, the 31+1165G/T allele-specific primers exhibited sequence incompatibilities with the MVR-primers. Mapping alleles from this site therefore required a pre-amplification step using the 31+1165G/T allele specific primer and primer 31C. Alleles were then mapped using the 3-state system with the universal flanking primer 31B (Figure 3.4C). Individuals heterozygous at both polymorphic sites were mapped from both sites to confirm the fidelity of the process. Occasionally, mapping using the allele-specific primers +744G and +1165T was not completely discriminatory but maps could be derived by subtraction from codes obtained for the opposite allele (Figure 3.4B). If the other site was also heterozygous, maps derived in this way were confirmed by allele-specific MVR-PCR from both sites.

Individuals suitable for allele-specific reverse MVR-PCR were identified by genotyping large numbers of individuals from different populations. A total of 72 alleles from unrelated individuals were mapped from the 3′ end of the array or in their entirety if shorter. This dataset consisted of 50 Caucasian alleles (22 French, 16 British and 12 Mormon individuals), 4 Japanese, 2 Africans, and 5 Afro-Caribbean alleles. An excess of Caucasian alleles were mapped because of their availability. In addition, the +744 site is not polymorphic in Afro-Caribbeans, and has a low heterozygosity (Ht = 0.05) in Africans so few alleles from these populations can be mapped from this site. All alleles mapped had different structures, indicating extreme variability at the 3′ end of the minisatellite array. Compared to maps derived from the 5′ end of the minisatellite there are a larger number of null repeats (**O**-type). This may be due to an increase in the number of AC repeats (Y-type) at this end of the array, which are not typed in this three state system. The absolute frequency of each of the four different repeat types (**E**, **e**, **y** and **O**) in all populations is the same ($\chi^2 = 0.006$, 3 d.f. , $p > 0.995$).

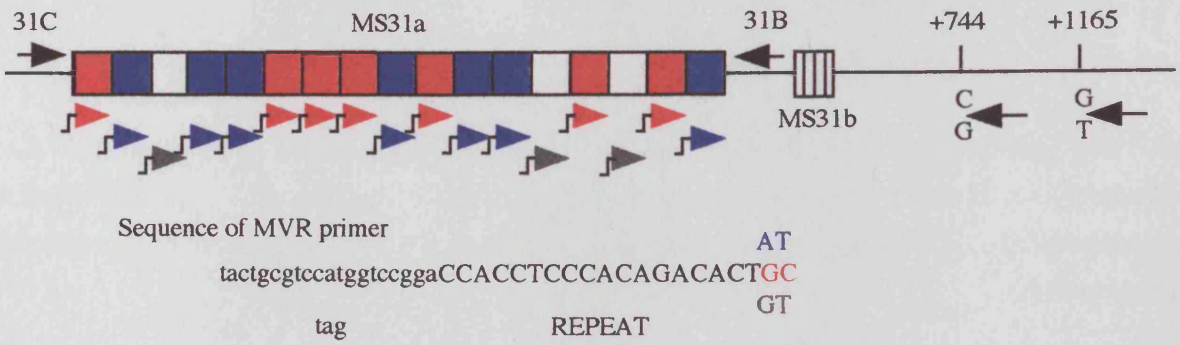
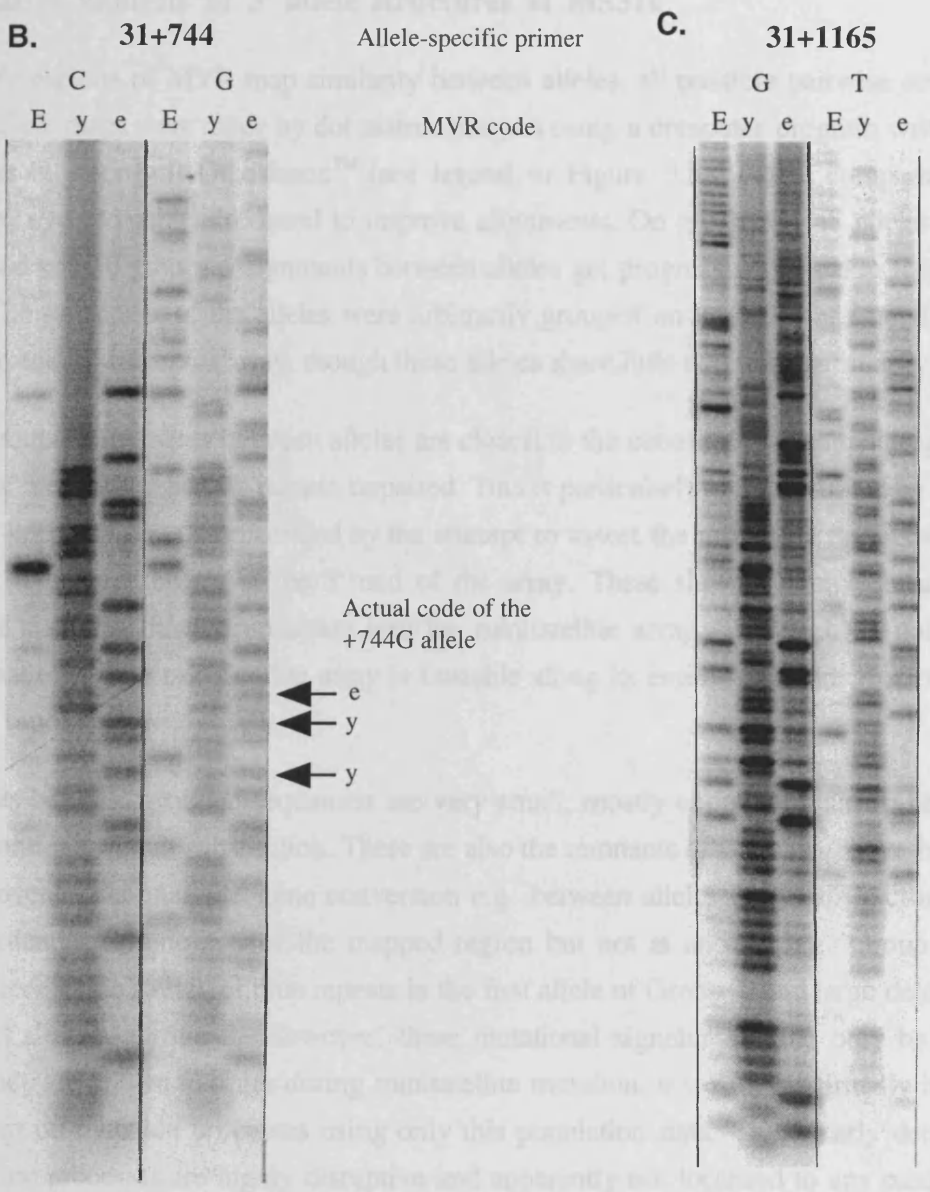


Figure 3.4. Three-state reverse allele-specific MVR-PCR at MS31a.

A. Schematic diagram to illustrate the principle of this technique. MS31b is shown as a striped box (diagram not to scale). A fixed primer (black arrows) in the flanking DNA of the minisatellite is coupled with MVR primers (coloured arrows) specific to the different variant repeat types of the minisatellite array (shown in blue, red and white). The sequence of each MVR repeat is shown with the diagnostic bases highlighted in colour. In the +744 system the fixed primer, included at high concentration consists of an allele-specific primer complementary to one or other allelic state of the +744C/G flanking polymorphic site. The +1165 system required pre-amplification of the array with primers 31C and 31+1165G/T, followed by mapping with universal primer 31B. For mapping, these fixed primers were coupled with a low concentration of each of three MVR primers 31rTAGAT (blue arrow), 31rTAGGC (red arrow) and 31rTAGGT (grey arrow). Each of these MVR primers has a TAG sequence ("tail" of arrow, sequence shown in lower case), a primer complementary to which was also included in the PCR reaction at high concentration. This allows the uncoupling of the PCR reaction from the MVR primers so that the longer PCR products are not lost in subsequent rounds of amplification. Each minisatellite variant repeat is given a code so those complementary to 31rTAGAT were coded as **E**, 31rTAGGC as **y**, and 31rTAGGT as **e**. The products from each PCR reaction are separated in adjacent lanes by electrophoresis and detected by Southern hybridisation.



B. Three-state allele-specific MVR-PCR on an individual heterozygous at +744 using flanking primers 31+744C and 31+744G as indicated. The distinguishing code of the MVR primers used (E, y or e) are indicated. The map derived from the +744C allele was very short, terminating after ~50 repeats. The code obtained using the 31+744G primer occasionally exhibited some "bleed through" from the other allele (arrows). The code at these ambiguous positions is obtained by subtracting the code of the +744C allele, the true code is shown next to the arrows.

C. Three-state allele-specific MVR-PCR on an individual heterozygous at +1165 by the pre-amplification and mapping steps described above. The distinguishing code of the MVR primers used (E, y or e) are indicated.

Comparative analysis of 3′ allele structures at MS31a

To identify regions of MVR map similarity between alleles, all possible pairwise comparisons between allele maps were made by dot matrix analysis using a computer program written by A. J. Jeffreys in Microsoft Quickbasic™ (see legend to Figure 3.5). Allele comparisons were checked by eye and gaps introduced to improve alignments. On this basis, 45 alleles out of 72 were placed into 10 groups. Alignments between alleles get progressively weaker from Groups 1 to 10. The remainder of the alleles were arbitrarily grouped on the basis of shared repeats at the very 3′ end of the repeat array, though these alleles share little additional similarity.

In most groups, alignments between alleles are closest in the centre of the mapped region while the ends of the maps generally remain unpaired. This is particularly well illustrated by the alleles of Group 4. This is also demonstrated by the attempt to assort the remaining unalignable alleles according to shared repeats at the 3′ end of the array. These show a complete lack of any significant similarity further upstream into the minisatellite array. This lack of polarity may suggest that either the minisatellite array is unstable along its entire length, or that mutation at MS31a is bipolar.

Differences between grouped sequences are very small, mostly comprising small insertions or deletions and repeat unit substitution. There are also the remnants of what may have been larger mutation events. For example, gene conversion e.g. between alleles in Group 3; crossing over in alleles identical at one end of the mapped region but not at another e.g. Group 1 alleles; duplications e.g. the stretch of blue repeats in the first allele of Group 7; and large deletions e.g. the second allele of Group 8. However, these mutational signatures could only be identified because they are known to occur during minisatellite mutation, it would be virtually impossible to comment on mutation processes using only this population data. This clearly demonstrates that mutation processes are highly disruptive and apparently not localised to any particular part of the array, although mutation studies will be required to confirm this.

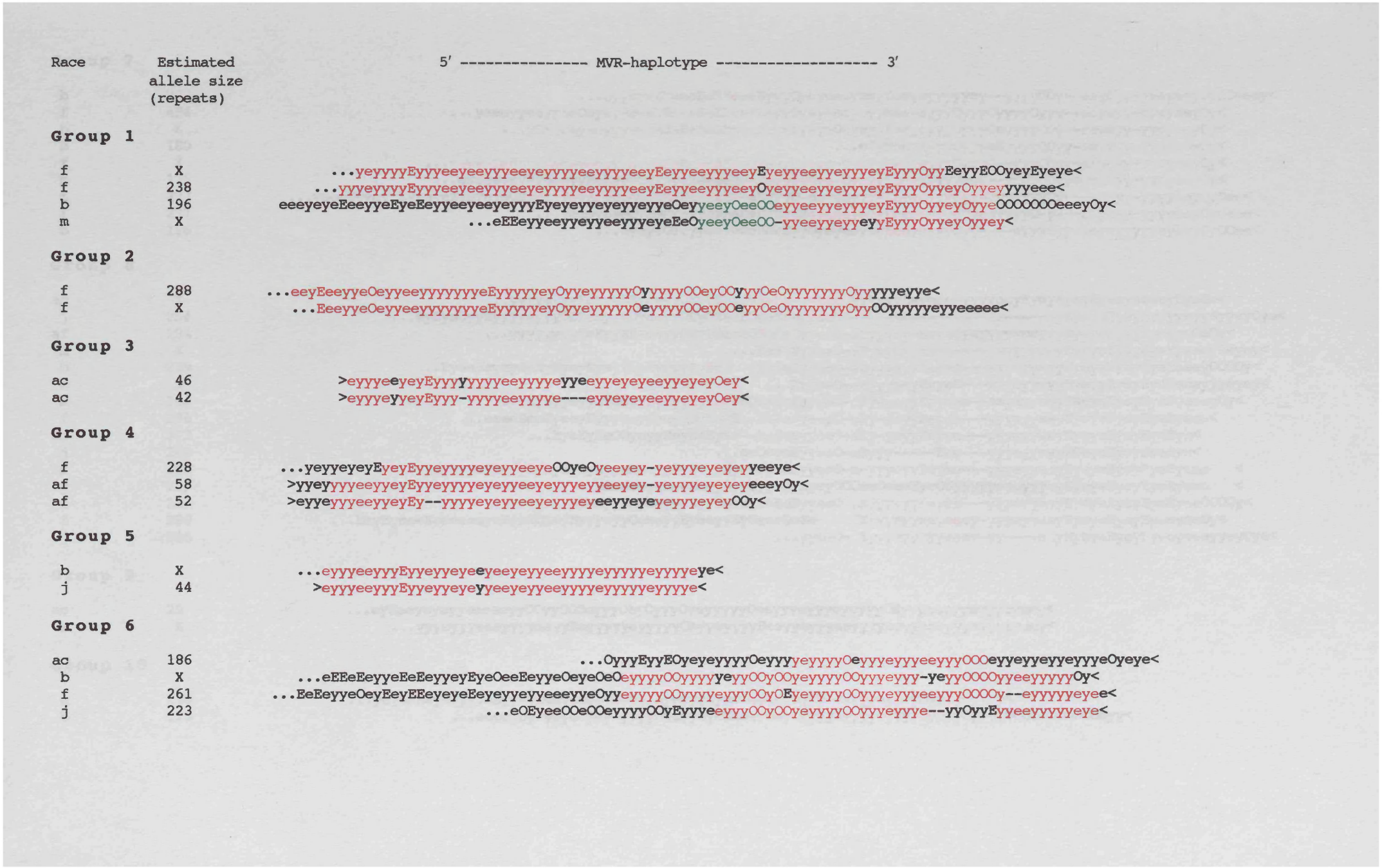
These population data may also be useful to identify common motifs, for example the 5′ yEyeyeEyeyEyeEye 3′ motif shared by alleles from different populations in Group 8. It may be informative to follow the inheritance of these motifs to determine why they are maintained in the human population at such high frequency. This motif could represent a more stable part of the minisatellite array, that for some reason, perhaps because of its structure, cannot be disrupted by minisatellite mutation processes. Alternatively it may have escaped mutation by chance. This would be useful to determine in terms of the dynamics of minisatellite mutation and the effect of minisatellite structure on instability.

Minisatellite structure and allele size

There appears to be very little association between allele length and minisatellite structure. However, it may be significant that of a total of six short alleles (< 70 repeats) only one could not be grouped, and those that were aligned fall within the same groups (Groups 3, 4 and 5). From the analysis of allele length it was expected that alleles from the African and

Figure 3.5 Databank of MS31a reverse MVR maps.

Groups of MS31a alleles aligned by dot matrix analysis. Alleles showing extensive regions of map similarity were identified in three searches of all possible pairwise comparisons of all 72 allele codes. The first pass searched for diagonals with perfect nine repeat matches in a window of ten; the second searched for perfect six repeat matches in a window of 6; and the third for perfect 10 repeat matches in a window of 12. Pairs of alleles showing at least twenty matching positions over the best two diagonals were selected and checked by eye for relatedness. Gaps (-) have been introduced to improve alignments. For each allele its ethnic origin; British (b), Mormon (m), French (f), Japanese (j), Afro-Caribbean (ac), African (af): and MVR haplotype are shown. **E** = AT type repeat; **y** = GC type repeat; **e** = GT type repeat; **O** = null repeat; ..... = allele continues beyond mapped region; > = 5′ end of short allele; < = 3′ end of allele. Array sizes are given in repeats, x = allele length not determined. Predominant regions of identity within a group are shown in red; additional regions of identity shared between different subgroups are shown in blue, green and yellow; positions of divergence are shown in black.


Race
Estimated allele size (repeats)
5′ ---------- MVR-haplotype ---------- 3′
Group 1
f X ...yeyyyyEyyyeeyeeyyyeeyeyyyyeeyyyyeeyEeyyeeyyyeeyEyeyyeeyyeyyyeyEyyyOyyEeyyEOOyeyEyeye<
f 238 ...yyyeyyyyEyyyeeyeeyyyeeyeyyyyeeyyyyeeyEeyyeeyyyeeyOyeyyeeyyeyyyeyEyyyOyyeyOyyeyyyyeee<
b 196 eeeyeyeEeeyyeEyeEeyyeeyeeyeyyyEyeyeyyyeyeyyeyyeOeyyeeyOeeOOeyyeeyyeyyyeyEyyyOyyeyOyyeOOOOOOOeeeyOy<
m X ...eEEeyyeeyyeyyeeyyyeyeEeOyeeyOeeOO-yyeeyyeyyeyEyyyOyyeyOyyey<
Group 2
f 288 ...eeyEeeyyeOeyyeeyyyyyyyeEyyyyyeyOyyeyyyyyOyyyyyOOeyOOyyyOeOyyyyyyyOyyyyyeyye<
f X ...EeeyyeOeyyeeyyyyyyyeEyyyyyeyOyyeyyyyyOeyyyyOOeyOOeyyOeOyyyyyyyOyyOOyyyyyeyyeeeee<
Group 3
ac 46 >eyyyeeyeyEyyyyyyyyyeeyyyyeyyeeyyeyeyeeyyeyeyOey<
ac 42 >eyyyeyyeyEyyy-yyyyeeyyyye---eyyeyeyeeyyeyeyOey<
Group 4
f 228 ...yeyyeyeyEyeyEyyeyyyyeyeyyeeyeOOyeOyeeyey-yeyyyeyeyeyyeeye<
af 58 >yyeyyyyeeyyeyEyyeyyyyeyeyyeeyeyyyyeyyeeyey-yeyyyeyeyeyeeeyOy<
af 52 >eyyeyyyeeyyeyEy--yyyyeyeyyeeyeyyyeyeeyyeyeyeyyyeyeyOOy<
Group 5
b X ...eyyyeeyyyEyyeyyeyeeyeeyeyyeeyyyyeyyyyyeyyyye<
j 44 >eyyyeeyyyEyyeyyeyeyyeeyeyyeeyyyyeyyyyyeyyyye<
Group 6
ac 186 ...OyyyEyyEOyeyeyyyyOeyyyyeyyyyOeyyyeyyyeeyyyOOOeyyeyyeyyeyyyeOyeye<
b X ...eEEeEeyyeEeEeyyeyEyeOeeEeyyeOeyeOeOeyyyyOOyyyyyeyyOOyOOyeyyyyOOyyyeyyy-yeyyOOOOyyeeyyyyyOy<
f 261 ...EeEeyyeOeyEeyEEeyeyeEeyeyyyeyyeeeyyeOyyeyyyyOOyyyyeyyyOOyOEyeyyyyOOyyyeyyyeeyyyOOOOy--eyyyyyeyee<
j 223 ...eOEyeeOOeOOeyyyyOOyEyyyeeyyyOOyOOyeyyyyOOyyyeyyye--yyOyyEyyeeyyyyyeye<


Group 7
b 203
f 496
b X
b 180
f X
m 176
f 227
f 237
b 126
Group 8
ac X
j 254
af 194
m X
b 229
b X
m 282
f 290
ac 373
j 242
f 280
j 209
f 284
f 280
j 216
Group 9
ac 29
m X
Group 10
m 238
b 128

Unalignable alleles:

```
ac      182  ...EyOEeeEOEeEeEyyeeyyEyyyeyyyeyeeyyyyeyEyyeyyeyyyyeyyeeeyyyyeyEyyeyEyeOyeOOOOy<
 f       X                                                ...eyeEyeEeyOyeeeeeeyeEEyyeOyyyyyeyyEyeeyyyeeeyyeEyeEyyeyOOOy<
 b      248                                 ...eEeEeeeEyEeeeyeEeyEeeEEeyOyyeyeyEyeyeyeeyEyeyeyEyEyeeyyeEyyyyeyyyyyeEEyyyeeeeyeyeeEyeyEyeEyeeeOOy<
ac      234                                                ...eOeeyeyeOeyyEeEeyeeyyeeeyyeEEeEeeeEEeeeeyOyeyeeeyyyOyyyyyOyyOyyeyyeyyyyeeyOy<
af       X                                                 ...yyeyeOOyyeOyyyeOyyyOyeeyyyeyOOyyyeOyyeyyeyyEyyOyyeyOyyeyOyyyOOyyyeeeyOy<
 j      253                                                                ...EEeeyyeeeyEeeEeyyyyyeyyyEeyyyyOOOyyeyyyyyyeeeyOy<
 f      169                                                          ...EeeEeyOyOOyyyOeyeyyeyyyyyOEyOyyyyEyeOyyyOyyyeeyyEOyyyeyyyeyeyOy<
 m      175                                             ...EeyEeyeyeEeyEeyEeyeyeeEeeEEeyyeeyyyeyyyeeyyyyeyeeOyeeyOeeOyeeyeyeyEyyOyeyyeyOOOOOOOOOOOOeyOy<

 f      319                                                                ...yeEeeEeEeeEEyyeeEEyyyeeyEeyeeyeyeyyeyyEyeyeEEyeeEyeyeyeeyyeyeyOe<
 m      234                                                    ...eyEyEyEyEyeyeeyyeyeyyeeyeyyeyeeeyeyeyeEeyEeyyeyeeeeyeeeyeyyeyeeyOyyeyyeyyOOyyeeyeyyeyeeyyOe<

 m      304                ...yeEEyyeeEeeeEeeEyyeEyyeEeyEeeeeeyEeeyyyyyeyyyyyEyeyOOyyyyyOyyyyyyyyyyyOyEyyyeyyeyyeeeyyyeyyeeeyyOOOyyyeyyyyyeyyyeeyey<
 j       43                                                                                                                >eyyyeeyyeyEyyeyyyeyeyyeyyyyyyyyyeyeyeyyyeyey<
 f      253                                                                                                                ...eeyyyyyeyeEeyeeeeyyyeeyeeeEyeyeEyeyyOeOyey<
 m       X                                                                                                                 ...yyyyyeeeeyeyeeyyyeeeOyeEeeyyeeyyyyeyyeeyeeOeeOeyeyyeyyeeOeyeeeeyyeeyyyeyey<
 b       81                                                                ...OOOOOEeeEEeyeyEeyeyeyEyEeyyeEyEEyyyyOyOyyyOyyyeyeeyyyOyyOyyeyyyyOOyyyyOOOyyyyyeyeyOeey<

 b       75                                                                                ...yeeeeyeyeyeyeeyeeeeyeyyeyyeyyOyeeeOeyeeyyyyeeyyeeyyeyeeyOyee<
 f      194                                                                          ...EeEeyyeyeyeEeyEeyeeEeyeyeyeOeOeyyOyyyeyyOyOyyOOOyyyyyyyeyyyOOOyyyyyyyOyee<
 f       X                                                                     ...yEOyeyeOyyyOeyOOyyyyyOeyyyeyyeeyyOOyyOyyyeyyeyyyyOyyyyeeyyyOOOyyyyyeyeeyOyee<
 b       X                                                                              ...eeyEyEyeyOOyyeeyeEyEyeyeyeyeyyeeyEyeyyEEyyyeyyOEyyyyyEeEyyEEEyeyEyeEyee<
 f       X                                                                                                  ...yeeOyeyeyeeyEeyeeeyyyeeyyeeyEyeyeeeeeEyeyeeee<
 m       X                                                                     ...yEyEyyyeyyyyeeyyyyyyeeOOyOOeeyyOyeyyyeeyyyyyyyOyyOeyeyyeyyyyyyyyeyyOyyyeyyyeyyyeyeeyOyOee<

ac       X                                                                ...eeyyEyeeyEyyyyEyeyyyeyyyOOEOeyOOOOOyEyyyEyyeEyyyyyyeyeyOyyyeeOyeyeyyeyeyyyeyey<
 f      187                                                          ...eeyyyeyyyyyeEyyyyyyOOyyyyyyOyyyyyOOyyOOyyyyOyOyyyyyyyyyOOOyyyyyye<
 b       X                                                                                                  ...yeOeEyyeyyEyeeeyyyeeyyeeyEyeyeOEyeyEye<
 b      313                                                                                                ...EyeOOEyyyeyyEyeeyyyyeyyyeeyEyeyeEyyyEyyEye<

ac       75                                                          ...OOOOOyyyyyeyyyyeyOOyyeyeyyyyeyyyyyeyyyyyyyOOOOeyyOOOOeyyyeyyeeyyyOOOOOOOyyyeyeyOe<

 m       X                                  ...yeyyyeyeyeeeyeOeyyeyEEyyeOyeyyyEyyyeyyeeyyyeyyeeyeyyyyeyyeeyeEyyyeeeeyeyyeyEyyyOyyeyyyyeyyyyyyyOOEyyyyyyyy<
```

Afro-Caribbean populations would generally be much shorter. What is striking however, is that short alleles from the same population are very similar yet they cannot be aligned with short alleles from other populations. The fact that these short alleles share similar structures suggests these two features may be linked. This may be either because the size of the allele prevents any major alterations in allele structure; because allele length expansion is suppressed by a flanking variant shared between these alleles; or because the allele structure can suppress allele size changes. Allele structure is unlikely to be responsible for suppressing changes in allele length because alleles from different populations show little sequence similarity and have been placed in three mutually exclusive groups. In addition these groups also contain longer alleles which share similar structures. It is possible that the flanking variant which confers this stability does exist but has not yet been identified, perhaps because it is rare, as in the case of the O1C variant at MS32 (Monckton *et al.*, 1994). The other possibility is that because these alleles are short they may represent alleles which are below the threshold for instability. The fact that short alleles from the same population are very similar is consistent with the idea that short alleles are stable and can spread to a high frequency in a population. The remaining groups appear to consist of a mixture of alleles of over about 120 repeats in length, suggesting that larger allele sizes are equally variable. This putative link between allele size and stability can only be determined by mutation analysis of short alleles.

Population-specific alignments

Some of these groups are population specific, Group 1 (Caucasians), Group 2 (French), Group 3 (Afro-Caribbean), Group 7 (Caucasians) and Group 10 (Caucasians). The significance of these population specific groups is difficult to interpret due to the predominance of Caucasian alleles typed. There appear to be a greater number of population specific groups identified at this end of the minisatellite than by previous mapping of the opposite end of the minisatellite (Neil, 1994). However, it is difficult to make comparisons because a less informative typing system was used in conjunction with a larger sample size in the previous study.

Discussion

The alleles investigated here do not represent equal numbers of all the populations sampled, the main reason for this was the scarcity of non-Caucasian DNAs. The dataset used was also rather small, particularly for the analysis of minisatellite structure, and this was exacerbated by the fact that much of the allele length data for these mapped alleles is incomplete. In addition, the use of specific flanking SNPs to map these alleles imposes another constraint which may also bias the data. However, it is still informative to examine these data and make some inferences about the mutational behaviour of MS31a.

Analysing allelic variation within the minisatellite

Comparison of this work with previous mutation studies at MS31a has demonstrated that it is virtually impossible to predict mutation processes from minisatellite allele structure. This has

reaffirmed the status of MS31a as one of the most unstable loci in the genome.

Analysis of array structure at MS31a has demonstrated that there is no polarity of variability at this minisatellite. Variability, and hence probably also mutation is distributed along the entire mapped length of the repeat array. This disagrees with previous studies of progenitor and mutant alleles in pedigrees. These have demonstrated that mutation is mainly restricted to the 5′ end of the minisatellite array (Neil, 1994); although there is one case where the origin of a length-change mutation could not be identified by mapping the 5′ end of the repeat array (Neil, 1994; Jeffreys *et al.*, 1998b; Chapter 5). The lack of any obvious polarity found at either end of the minisatellite during population studies suggests that mutation at MS31a may be bipolar. However, mutation studies suggest that this polarity is biased towards the 5′ end of the array, while mutation at the 3′ end must be of high enough frequency to maintain the high levels of variation observed here. The explanation for this distribution of mutation may be found by examining the DNA flanking the minisatellite. It has long been accepted that minisatellite mutation is determined by *cis*-acting factors located in the flanking DNA adjacent to the most unstable end of the minisatellite, and is not intrinsic to the array itself. The alternative explanation, that mutation events are randomly scattered along the entire length repeat array is not consistent with this idea that mutation is controlled by *cis*-acting factors, nor does it correlate with the mutation data so far obtained.

Population analysis

Obvious differences between the populations were observed on examination of minisatellite allele-length frequencies. This showed that the alleles fell into two groups, the short alleles of the Afro-Caribbean and African populations; and the longer alleles of the Caucasian and Japanese individuals (Figure 3.1). This difference in allele size distribution of the different populations may be due to an ancient founder effect in the non-African populations. These distributions may be interesting in terms of minisatellite instability because it has been shown at some minisatellites that mutation increases with the size of the allele. This phenomenon has been observed at CEB1 (Buard *et al.*, 1998), B6.7 (Jeffreys *et al.*, 1997; Tamaki *et al.*, 1999 in press), and in pedigree mutant data from minisatellite pλg3 (Andreassen *et al.*, 1996). Conversely, no relationship between mutation and minisatellite size has been observed at either MS32 (Jeffreys *et al.*, 1994), or MS205 (May *et al.*, 1996). If this is true of MS31a this evidence suggests that the mutation rate may be higher in the Caucasian and Japanese populations due to the larger allele size in these populations. This may also explain the high similarity between short alleles from the same population, observed on mapping allele structure.

There were no strong links found between allele structure and population group, although this may be mainly due to the small number of alleles used in this study. Most of the population specific groups consist of Caucasian alleles, which is not surprising considering the number of these alleles typed. What is perhaps more surprising is the degree of similarity between some alleles, particularly the French alleles in Groups 2 and 3. This suggests that it may be possible to examine the ancestry of certain MS31a alleles to trace mutation through particular lineages. It

is not clear why the similarity between these alleles should be so great, perhaps it relates to mutation processes specific to these minisatellites.

The short alleles are potentially the most interesting because they can be typed in their entirety. It is notable that all the short alleles typed except one could be aligned, generally with short alleles from the same population group. Previous population studies have also identified short population specific minisatellite alleles in the Japanese (Neil, 1994). Within this group there were examples of alleles which showed large deletions or insertions which were postulated to represent intra-allelic mutation events typical of somatic cells (Jeffreys & Neumann, 1997). No such events were observed within this dataset which may be due to the small number of alleles typed. Unfortunately, comparisons between alleles mapped in this study and those previously mapped from the 5′ end could not be made because the three-state MVR system was used in this work, and a two-state system used in the previous work.

Further work

This work has demonstrated that mutation at MS31a may be more complicated than first thought, and there is much work to be done for an improved understanding of these processes. Firstly the population data obtained in this study could be completed by the determination of allele length of all mapped alleles. This dataset can then be extended to include more alleles to determine whether the present alignments are maintained, particularly between the short alleles. It would also be informative to re-analyse those alleles in which the mutation could not be identified by mapping from the 5′ end of the array. This would determine where in the array mutation has occurred, and whether mutation involves the gene conversion events which dominate the 5′ end of the array. These studies can also be extended into the flanking DNA to examine recombination between flanking markers at both ends of the minisatellite array and perhaps identify the *cis*-acting factors which drive minisatellite mutation.

Chapter 4

Flanking DNA analysis at MS31a

Summary

This work extends the analysis of variation to include the flanking DNA of the hypervariable minisatellite, MS31a in the Japanese, African, Afro-Caribbean and Caucasian populations. A total of eight polymorphic positions in the 5′ flanking DNA, and three in the 3′ flanking DNA have been identified. Assays have now been developed for genotyping, and for haplotype analysis of these sites. In all populations, analysis of linkage disequilibrium between polymorphic sites in the flanking DNA of MS31a indicated high levels of recombination existed throughout the locus. These patterns were complicated, particularly in the African and Afro-Caribbean populations, but the general trend suggested a breakdown of linkage disequilibrium in the 5′ flanking DNA adjacent to the minisatellite. This concurs with patterns of linkage disequilibrium at the MS32 locus in Caucasians, which were subsequently found to correspond to a meiotic recombination hotspot in the 5′ DNA flanking MS32 (Jeffreys *et al.*, 1998b). Comparison of flanking haplotypes between related MS31a repeat array structures provided further evidence of recombination in the flanking DNA at each end of MS31a. This work indicates that instability of the repeat array may be bipolar, i.e. driven by elevated levels of recombination in the flanking DNA at both ends of MS31a.

Introduction

Variation in the DNA flanking minisatellites can be found in the form of single nucleotide polymorphisms (SNPs), and other repeat types, such as simple tandem repeats and dispersed repeats. Flanking SNPs have been routinely used for allele-specific amplification during MVR-PCR (see Chapter 3), and for the isolation of single minisatellite alleles. They can also be used to place minisatellite alleles into haplogroups according to shared patterns of flanking polymorphic sites. This can permit the definition of allele lineages in the human population. Studies have also implied that particular flanking polymorphic sites can influence minisatellite instability (Monckton *et al.*, 1994). This is probably mediated by changes in chromatin conformation effected by this site, because neither the primary nucleotide sequence, nor the DNA secondary structure are conserved between the flanking DNA of different minisatellites (Murray *et al.*, 1999). For a long time evidence has suggested that minisatellite instability is due to *cis*-acting factors outside the repeat array. This has been confirmed by recent work at MS32 which identified a recombination hotspot immediately flanking the most unstable part of the array that appears to drive instability of the

minisatellite. This hotspot is located in a region which showed a breakdown of linkage disequilibrium in the Caucasian population (Jeffreys *et al.*, 1998a & b). It is not known whether this hotspot is common to all minisatellites but comparative studies of flanking polymorphic sites should be able to assess the probability that a recombination hotspot exists in the flanking DNA of MS31a.

MS31a Flanking DNA

Like the majority of hypervariable minisatellites, MS31a is embedded in a repeat-rich region. This shows no significant primary or secondary sequence similarity with the DNA flanking other minisatellites (Armour *et al.*, 1989b; Murray *et al.*, 1999). This region is highly GC-rich (~60% GC), which is consistent with its location in a terminal isochore of a human chromosome. Approximately 15 kb of the downstream DNA, and ~14 kb of the upstream DNA flanking MS31a has been sequenced. This revealed the presence of multiple dispersed repeats, including L1 and Alu elements, simple tandem repeats, many of which are associated with the polyA tails of the dispersed repeats, and three additional minisatellites, MS31b, MS31c and MS31d (Murray *et al.*, 1999). More detailed sequencing of 2 kb of the 3′ flanking DNA immediately downstream of MS31b has identified the two SNPs, +744 and +1165, as described in Chapter 3. Extensive sequencing has also identified eight polymorphic sites in the 2 kb immediately flanking the 5′ end of the minisatellite array (D. L. Neil, pers. comm.). Seven of these are SNPs, and one is a 12 bp insertion/deletion (indel) site adjacent to the -220 site. Preliminary linkage disequilibrium studies between sites -220, -109 and -4 in a limited number of Caucasian alleles have shown low levels of linkage disequilibrium in the 5′ flanking DNA of MS31a (Neil, 1994). This suggests there may be high levels of recombination in this region and justifies the need for further studies on the flanking DNA.

This work

Studies have suggested that mutation at MS31a is polarised towards the 5′ end of the minisatellite array. The majority of mutation events identified have been located within ten repeats of the upstream end of this minisatellite (Neil, 1994). However, MVR codes obtained during population studies by Neil (1994), and in Chapter 3 failed to confirm this polarity of variability. This chapter investigates further the flanking regions of the minisatellite in an attempt to determine the stability of the flanking regions of MS31a. This may also define features within these regions which affect the stability of the tandem repeat array. Haplotypes of 5′ and 3′ flanking markers were constructed in alleles from different populations (Caucasian, Japanese, African and Afro-Caribbean). These haplotypes were then examined for evidence of recombination and comparison of linkage disequilibrium patterns between these sites was used to make inferences concerning the distribution of recombination in the flanking DNA. Finally, the relationship between flanking haplotype and MS31a allele structure was explored

Results

Population surveys of flanking polymorphic positions

The distribution of flanking polymorphisms in different populations was determined by genotyping unrelated African, Afro-Caribbean, Caucasian and Japanese individuals using the conditions determined in Chapter 3 (see also Table 2.6, Materials & Methods) for sites in the 3′ flanking DNA, and conditions detailed in the Materials & Methods (Table 2.6) for sites in the 5′ flanking DNA. The majority of the sites in the 5′ flanking DNA were genotyped by David Neil. The heterozygosities of these polymorphic sites were defined (see Table 4.1 for allele frequencies and Table 4.2 for genotype frequencies), demonstrating that polymorphic sites in the four populations surveyed are at Hardy-Weinberg equilibrium (maximum $\chi^2 = 5.57$, 1 d.f., $p > 0.062$).

These polymorphic markers are distributed in a small area adjacent to the minisatellite (Figure 4.1). Within a region of 2.0 kb upstream eight polymorphic sites have been identified, and three polymorphic sites are located within 1.9 kb downstream of MS31b. The frequency of SNPs in the 5′ flanking DNA is significantly higher than the genomic average of one SNP per kilobase ($\chi^2 = 12.3$, 1 d.f., $p = 0.0005$), although this is not true of the 3′ flanking DNA ($\chi^2 = 0.000239$, 1 d.f. , $p = 0.988$). The markers in the 5′ flanking DNA also show a non-random distribution ($\chi^2 = 43$, 10 d.f., $p > 0.001$) compared to the random distribution of markers in the 3′ flanking DNA ($\chi^2 = 7.8$, 8 d.f., $p = 0.45$). This suggests that the 5′ flanking DNA may be associated with an increase in mutation rate. Alternatively, high rates of recombination in the flanking DNA can result in rapid shuffling of the SNPs, so new variants are spread quickly through the population. The plasticity of this region also means that it is virtually impossible to use these data for evolutionary analysis; for example, to predict rates of turnover of these sites in the flanking DNA, and to determine the evolutionary ages of the alleles. However, the relationship of these sites to one another can be examined by determining the combination of alleles at each site on each chromosome in different individuals (haplotyping). Patterns of linkage disequilibrium within these haplotypes can then be examined to make inferences about recombination in the DNA flanking the minisatellite.

Haplotype variation and recombination

It has long been suggested that minisatellites mark regions of recombinational activity (Jeffreys *et al.*, 1985). This has recently been proven following the identification of a recombination hotspot in the 5′ flanking DNA immediately adjacent to MS32 (Jeffreys *et al.*, 1998b). The first step in accumulating such evidence at MS31a involves examining patterns of haplotype variation and linkage disequilibrium both between and within the different populations. Complete haplotypes were obtained for 63 Caucasian, 21 Japanese, 17 Afro-Caribbean and 29 African alleles (Figures 4.5A, 4.6A, 4.7A and 4.8A, respectively) as described in Materials & Methods. Two simple tests,

Table 4.1. Heterozygosities of each flanking polymorphic site in 44 Caucasian, 15 Japanese, 9 Afro-Carribean and 20 African individuals

		Caucasian			Japanese			Afro-Carribean			African		
Site	Allele	No.	Frequency	Heterozygosity	No.	Frequency	Heterozygosity	No.	Frequency	Heterozygosity	No.	Frequency	Heterozygosity
-1752	C	88	1	0	26	0.87	0.23	16	0.89	0.20	30	0.75	0.38
	T	0	0		4	0.13		2	0.11		10	0.25	
-1663	C	45	0.51	0.5	29	0.97	0.06	5	0.28	0.40	20	0.50	0.50
	T	43	0.49		1	0.03		13	0.72		20	0.50	
-759	A	49	0.56	0.5	16	0.53	0.5	7	0.39	0.48	24	0.60	0.48
	C	39	0.44		14	0.47		11	0.61		16	0.40	
-457	C	38	0.43	0.5	14	0.47	0.5	12	0.67	0.44	9	0.23	0.35
	T	50	0.57		16	0.53		6	0.33		31	0.78	
ΔH	+	0	0	0	2	0.07	0.13	1	0.06	0.10	15	0.38	0.47
	-	88	1		28	0.93		17	0.94		25	0.63	
-220	C	36	0.41	0.5	9	0.3	0.42	4	0.22	0.35	9	0.23	0.35
	G	52	0.59		21	0.7		14	0.78		31	0.78	
-109	C	25	0.28	0.4	15	0.5	0.5	3	0.17	0.28	16	0.40	0.48
	T	63	0.72		15	0.5		15	0.83		24	0.60	
-4	A	16	0.18	0.29	6	0.2	0.3	5	0.28	0.40	7	0.18	0.29
	G	72	0.82		24	0.8		13	0.72		33	0.83	
MS31B	A	66	0.75	0.38	20	0.67	0.51	12	0.67	0.48	28	0.70	0.46
	B	22	0.25		3	0.1		5	0.28		5	0.13	
	C	0	0		5	0.17		1	0.06		7	0.18	
	D	0	0		2	0.06		0	0.00		0	0.00	
+744	C	16	0.18	0.29	1	0.03	0.06	0	0.000	0.00	1	0.03	0.05
	G	72	0.82		29	0.97		18	1.000		39	0.98	
+1165	G	39	0.44	0.5	8	0.27	0.4	6	0.33	0.44	18	0.45	0.50
	T	49	0.56		22	0.73		12	0.67		22	0.55	

Table 4.2 Hardy-Weinberg calculations for each polymorphic site in 44 Caucasian, 15 Japanese, 9 Afro-Caribbean and 10 African individuals.

	Caucasian		Japanese		Afro-Carribean		African	
Genotype	O.	E.	O.	E.	O.	E.	O.	E.
-1752C/C	44		12	11.35	7	7.13	11	10.95
-1752C/T	0		2	3.39	2	1.76	8	7.70
-1752T/T	0		1	0.25	0	0.11	1	1.35
χ^2			2.81	p~0.25	0.14	p~0.93	0.10	p~0.95
-1663C/C	12	11.44	14	12.97	4	4.16	5	8.98
-1663C/T	22	21.99	1	1.95	4	3.92	10	8.84
-1663T/T	10	10.56	0	0.07	1	0.92	5	2.18
χ^2	0.06	p=0.97	0.62	p=0.73	0.01	p=0.99	5.57	p=0.062
-759A/A	17	13.70	5	4.21	2	1.23	8	6.96
-759A/C	15	21.70	6	7.47	3	4.53	8	9.68
-759C/C	12	8.60	4	3.31	4	4.16	4	3.36
χ^2	4.21	p=0.12	0.58	p=0.75	1.00	p=0.61	0.57	p=0.75
457T/T	17	14.30	5	4.21	1	1.23	13	11.86
-457C/T	15	21.57	6	7.47	4	4.20	5	7.08
-457C/C	12	8.14	4	3.31	4	3.57	2	1.06
χ^2	4.35	p=0.11	0.58	p=0.75	0.10	p=0.95	1.56	p=0.46
ΔH-/-	44		13	12.97	8	8.12	8	7.69
ΔH-/+	0		2	1.95	1	0.86	9	9.42
ΔH+/+	0		0	0.07	0	0.02	3	2.89
χ^2			0.075	p=0.96	0.049	p=0.98	0.036	p=0.98
-220G/G	16	15.32	6	7.35	6	4.93	12	12.48
-220G/C	19	21.29	9	6.30	2	3.46	7	6.64
-220C/C	9	7.40	0	1.35	1	0.61	1	0.88
χ^2	0.624	p=0.73	2.755	p=0.25	1.103	p=0.58	0.054	p=0.97
-109T/T	21	22.81	2	3.75	6	6.35	9	6.96
-109T/C	21	17.74	11	7.50	3	2.42	6	9.68
-109C/C	2	3.45	2	3.75	0	0.23	5	3.36
χ^2	1.35	p=0.51	3.27	p=0.2	0.39	p=0.82	2.79	p=0.25
-4G/G	30	29.59	10	9.60	4	4.93	13	13.45
-4G/A	12	12.99	4	4.80	5	3.46	7	5.90
-4A/A	2	1.43	1	0.60	0	0.61	0	0.65
χ^2	0.31	p=0.86	0.42	p=0.81	1.47	p=0.48	0.87	p=0.65
MS31BA/A	27	24.75	8	6.73	4	4.28	9	9.52
MS31BA/A'	13	16.50	4	6.63	4	3.85	10	8.56
MS31BA'/A'	4	2.75	3	1.63	1	0.86	1	1.92
χ^2	1.52	p=0.47	2.43	p=0.3	0.05	p=0.98	0.71	p=0.7
+744G/G	29	29.59	14	14.11	9		20	18.82
+744G/C	15	12.99	1	0.87	0		0	1.16
+744C/C	0	1.43	0	0.01	0		0	0.02
χ^2	1.75	p=0.42	0.03	p=0.99			1.26	p=0.53
+1165T/T	13	13.80	8	7.99	4	4.16	5	5.83
+1165T/G	21	21.68	6	5.91	4	3.92	12	9.94
+1165G/G	10	8.52	1	1.09	1	0.92	3	4.23
χ^2	0.33	p=0.85	0.01	p=0.99	0.01	p=0.99	0.91	p=0.64

O. - Observed
E. - Expected

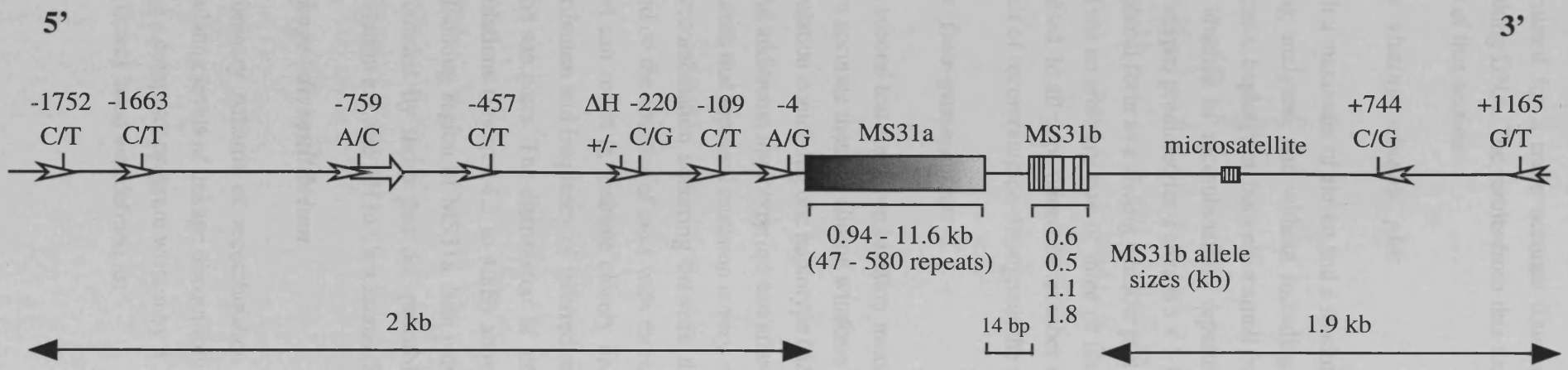


Figure 4.1. Schematic diagram of MS31a plus flanking DNA.

Shading within MS31a represents the distribution of gene-conversion activity, with the darkest region representing the hotspot of activity at the 5′ end of the array. The three different repeat types of MS31b (19 bp, 64 bp and 29 bp respectively) are represented by vertical lines within the minisatellite. Allele specific primers are represented by arrow heads and the discriminatory base of the primer along with the site reference are indicated above. A single Alu repeat is represented by an arrow in the 5′ flanking DNA and a microsatellite is shown in the 3′ flanking DNA. The allele sizes for both MS31a and b are indicated. Diagram not to scale.

the sliding window plot and the four-gamete test, have been used to assess these haplotypes directly for evidence of recombination. Levels of linkage disequilibrium between sites were then calculated for a more accurate illustration of the distribution of recombination throughout the flanking DNA. The results from this analysis will be examined in detail for each population at the end of this section.

The sliding window plot

With a maximum of eleven and a minimum of nine segregating sites (depending on the population being analysed, and without including the alleles of MS31a itself) there are a huge number of potential haplotypes, but only a small fraction of these would be seen in a finite population. So, in the absence of recombination, repeated mutation or back mutation, the maximum number of haplotypes predicted for *s* sites is *s* + 1 haplotypes (Clark *et al.*, 1998). This can be represented in graphical form as a sliding window plot for each population. In each population the sliding window is given an arbitrary size of three or four sites, and the number of haplotypes for each window is counted. In all populations the number of haplotypes is greater than expected, suggesting elevated levels of recombination throughout the flanking DNA (Figures 4.5C to 4.8C).

The four-gamete test

The second test for recombination makes a direct comparison between two sites and is potentially more accurate than the sliding windows plot (Clark *et al.*, 1998). This test is based on the fact that a population containing one haplotype (*AB*) may mutate to give a further two haplotypes, *A-b* and *a-B*. An additional haplotype *a-b* can arise only through recombination or repeated mutation. The test assumes that repeated mutation is very rare, so all site pairs having four haplotypes can only arise by recombination occurring between the sites. The assumption that repeated mutation is rare is based on the absence of sites with more than two segregating nucleotides. A single recombination event can result in multiple closely linked site pairs having all four haplotypes present, so the distribution and frequency of inferred recombination events requires consideration of the locations of the site pairs. The distribution of site pairs for which all four haplotypes were present in all populations (Figures 4.5 to 4.8B) show that recombination is likely to have occurred throughout the flanking region of MS31a. Site pairs that do not have four haplotypes present despite being surrounded by those that do, probably indicate sites which have low heterozygosity in that population e.g. site ΔH in Caucasians (Figure 4.5B).

Linkage disequilibrium

Preliminary patterns of recombination within the flanking DNA of MS31a were obtained by calculating levels of linkage disequilibrium between polymorphic sites (χ^2 statistics were calculated using a computer program written by Y. E. Dubrova according to Weir, 1990). The disequilibrium coefficient used was defined as:

$$D_{uv} = p_{uv} - p_u p_v$$

where p_{uv} is the frequency of haplotype uv formed from alleles u and v at two loci with frequencies p_u and p_v, respectively. This coefficient has a maximum value of 0.24 when loci are in complete linkage disequilibrium. The correlation (r_{uv}) between these coefficients was determined to correct for allele frequencies:

$$r_{uv} = \frac{D}{\sqrt{p_u(1-p_u)p_v(1-p_v)}}$$

The chi-squared value (degrees of freedom = 1) for each pair of loci was then determined by:

$$\chi^2 = n(r_{uv})^2$$

It must be noted that for simplicity, the graphical representation of these data consists of a pairwise comparison between sites, so that only values for adjacent sites can be directly compared (Figure 4.2). No significant linkage disequilibrium was found between non-adjacent sites (data not shown). This suggests that recombination at this locus is dominated by equal crossing over and not gene-conversion events. This analysis is complicated by the small numbers of alleles examined in the non-Caucasian populations, such that the level of linkage disequilibrium between most sites are not significant in these populations. However, it does indicate that there are likely to be high levels of recombination in these flanking regions, and that these patterns of linkage disequilibrium can be used to infer patterns of recombination. The levels of linkage disequilibrium in all four populations appear to be highest between sites -1663, -759 and -457 in the 5′ flanking DNA and, to a lesser extent between the three sites in the 3′ flanking DNA (Figure 4.2).

Determination of MS31a flanking haplogroups

Applying these linkage disequilibrium data, the haplotypes were classified into eight haplogroups according to the three 5′ flanking sites. These were then subdivided (a-i) according to the three 3′ polymorphic sites, MS31b, +744 and +1165 (Figure 4.3). Twenty population specific haplogroups were defined, although many of these comprise single alleles, and most are Caucasian specific. The largest haplogroup, which also includes individuals from all populations is Group 1a. This is interesting because the alleles present at sites -759 and -457 are not the most common in the human population (black circles). The three most common alleles (white circles) at all three 5′ sites, -1663, -759 and -457 are actually those which define Group 3. This group contains the largest number of different subgroups, which is logical if, as expected these common 5′ alleles represent the ancestral state of these sites. Many (6 out of 21) of the Japanese alleles typed fall within Group 5a which suggests that there is a strong founder effect in this population. Other population specific groups include: haplogroup 6a which is made up entirely of alleles of African or Afro-Caribbean origin; and haplogroup 7 which is divided neatly into subgroup 7a consisting of Caucasian alleles, and subgroup 7b consisting of Afro-Caribbean alleles. It is also notable that within certain groups there are a limited number of haplotypes. For example Group 2h contains eight alleles but only two

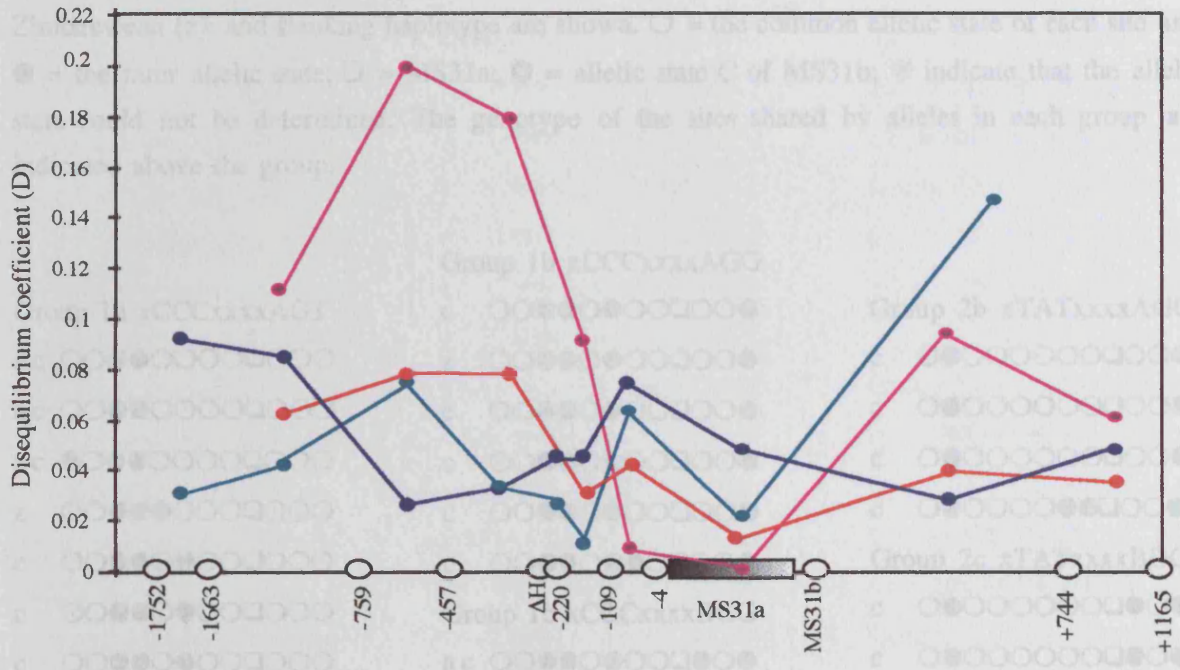


Figure 4.2. Patterns of linkage disequilibrium in all populations studied.

Patterns of linkage disequilibrium observed in 63 Caucasian (magenta), 21 Japanese (red), 17 Afro-Caribbean (light blue) and 29 African alleles (dark blue). A schematic diagram of MS31a showing the minisatellite (black shade represents the more unstable end of the array), and the flanking markers is drawn to scale underneath the graph. The linkage disequilibrium coefficient (D) was calculated as described in the text for each adjacent pair of markers shown on the representative MS31a allele shown below.

Figure 4.3. MS31a Flanking Haplotype Database.

Alleles were assigned into eight different haplogroups (1-8) according to the 5′ sites -1663, -759 and -457, these were then subdivided according to the 3′ sites MS31b, +744 and +1165 (a-i). For each allele its ethnic origin; Caucasian (c), Japanese (j), Afro-Caribbean (ac), West African (af), Zimbabwean (z): and flanking haplotype are shown. ○ = the common allelic state of each site and ● = the rarer allelic state; ❑ = MS31a; ❂ = allelic state C of MS31b; ✳ indicate that the allelic state could not be determined. The genotype of the sites shared by alleles in each group are indicated above the group.

Group 1a xCCCxxxxAGT

a c ○○●●○○○○❑○○○
a c ○○●●○○○○❑○○○
a c ●○●●○○○○❑○○○
z ○○●●●○○○❑○○○
c ○○●●○●○○❑○○○
c ○○●●○●○○❑○○○
c ○○●●○●○○❑○○○
c ○○●●○●○○❑○○○
c ○○●●○●○○❑○○○
z ○○●●○●○○❑○○○
c ○○●●○●○○❑○○○
c ○○●●○●○○❑○○○
c ○○●●○●○○❑○○○
c ○○●●○●○○❑○○○
a c ✳○●●○●○○❑○○○
j ○○●●○○●○❑○○○
c ○○●●○○●○❑○○○
c ○○●●○○●○❑○○○
a c ○○●●○○○●❑○○○
j ○○●●○○○●❑○○○
c ○○●●○●●○❑○○○
j ○○●●○●●○❑○○○
j ○○●●○●●○❑○○○
a c ○○●●○○●●❑○○○

Group 1b xCCCxxxxAGG

c ○○●●○●○○❑○○●
z ○○●●○●○○❑○○●
c ○○●●○●○○❑○○●
c ○○●●○●○○❑○○●
c ○○●●○●○○❑○○●
c ○○●●○●●○❑○○●

Group 1c xCCCxxxxBGG

a c ○○●●○●○○❑●○●
z ○○●●○●○○❑●○●

Group 1e xCCCxxxxBGT

c ○○●●○●○○❑●○○

Group 1g xCCCxxxxACG

c ○○●●○●○○❑○●●
c ○○●●○○●○❑○●●

Group 2a xTATxxxxAGT

a c ○●○○○○○○❑○○○
z ○●○○○○○○❑○○○
z ○●○○○○○○❑○○○
a c ○●○○●○○○❑○○○
z ○●○○●○○○❑○○○
z ○●○○●●○○❑○○○
z ○●○○○○●○❑○○○
c ○●○○○○●○❑○○○
c ○●○○○○●○❑○○○
c ○●○○○○●○❑○○○
c ○●○○○○●○❑○○○
c ○●○○○○●●❑○○○
c ○●○○○○●●❑○○○

Group 2b xTATxxxxAGG

c ○●○○○○○○❑○○●
c ○●○○○○○○❑○○●
c ○●○○○○○○❑○○●
c ○●○○○○●●❑○○●

Group 2c xTATxxxxBGG

c ○●○○○○○○❑●○●
c ○●○○○○○○❑●○●
c ○●○○○○○○❑●○●
c ○●○○○○○○❑●○●
c ○●○○○○●●❑●○●

Group 2f xTATxxxxCGT

z ●●○○●○●○❑❂○●

Group 2h xTATxxxxBCG

c ○●○○○○○○❑●●●
c ○●○○○○○○❑●●●
c ○●○○○○○○❑●●●
c ○●○○○○○○❑●●●
c ○●○○○○●○❑●●●
c ○●○○○○●○❑●●●
c ○●○○○○●○❑●●●

Group 3a xCATxxxxAGT

j ○○○○○○○○❑○○○

j ○○○○○○○○❑○○○

c ○○○○○○○○❑○○○

j ○○○○○○○○❑○○○

a c ○○○○○○○○❑○○○

a f ●○○○○○○○❑○○○

z ○○○○●○○○❑○○○

z ○○○○●○○○❑○○○

j ○○○○○●○○❑○○○

j ○○○○○●○○❑○○○

z ○○○○○○●○❑○○○

c ○○○○○○●○❑○○○

c ○○○○○○○●❑○○○

j ○○○○○●●●❑○○○

a f ○○○○●●●●❑○○○

Group 3b xCATxxxxAGG

z ○○○○○○○○❑○○●

j ○○○○○●○○❑○○●

c ○○○○○○○●❑○○●

c ○○○○○○○●❑○○●

Group 3c xCATxxxxBGG

a c ○○○○○○○○❑●○●

Group 3d xCATxxxxCGG

j ●○○○○○○○❑❂○●

Group 3e xCATxxxxBGT

c ○○○○○○●●❑●○○

Group 3f xCATxxxxCGT

z ●○○○○○○○❑❂○○

Group 3g xCATxxxxACT

c ○○○○○○○○❑○●○

c ○○○○○○○○❑○●○

Group 3h xCATxxxxBCG

j ○○○○○●○●❑●●●

Group 3i xCATxxxxBCT

z ●○○○●○○●❑●●○

Group 4a xTCTxxxxAGT

z ●●●○○○○○❑○○○

z ○●●○○○○○❑○○○

z ○●●○○○●●○❑○○○

c ○●●○○○●○○❑○○○

z ○●●○○○●○○❑○○○

c ○●●○○○●○○❑○○○

Group 4b xTCTxxxxAGG

c ○●●○○○●○○❑○○●

z ○●●○○●●○○❑○○●

Group 4d xTCTxxxxCGG

j ●●●○○○○○❑❂○●

a c ○●●○○○○○❑❂○●

z ○●●○○●○○○❑❂○●

j ●●●○○○●○○❑❂○●

Group 5a xCACxxxxAGT

j ○○○●○○○○❑○○○

j ○○○●○○●○○❑○○○

j ○○○●○○○●○❑○○○

a c ○○○●○○○●○❑○○○

j ○○○●○○○●○❑○○○

c ○○○●○○○●○❑○○○

j ○○○●○○●●○❑○○○

j ○○○●○○●●●❑○○○

Group 5c xCACxxxxBGG

a c ○○○●○○○●○❑●○●

Group 6a xCCTxxxxAGT

a f ●○●○✳○○○❑○○○

a c ○○●○○○●○○❑○○○

z ○○●○○○●●○❑○○○

a c ○○●○○○●●○❑○○○

z ○○●○○●○●○❑○○○

a f ●○●○✳○○●●❑○○○

Group 6b xCCTxxxxAGG

z ○○●○○○○●○❑○○○●

c ○○●○○○●○○❑○○○●

Group 7a xTCCxxxxAGT

c ○●●●○●○●○❑○○○

c ○●●●○●○●○❑○○○

c ○●●●○●○●○❑○○○

c ○●●●○●○●○❑○○○

Group 7c xTCCxxxxBGG

a c ○●●●○○○○○❑●○●

a c ✳●●●○●○○○❑●○●

Group 8a xTACxxxxAGT

a c ○●○●○○○●●❑○○○

Group 8b xTACxxxxAGG

c ○●○●○○○●○❑○○○●

Group 8c xTACxxxxBGG

z ○●○●○○○○○❑●○●

z ○●○●○○○●○❑●○●

Group 8d xTACxxxxCGG

z ○●○●●●○○○❑❂○●

haplotypes, whereas Group 5c which also contains eight alleles, contains six different haplotypes. This may be due to the greater population diversity in Group 5c which suggests the group is more ancient. The distribution of African and Afro-Caribbean alleles throughout all groups indicates the high level of diversity within these populations. This may reflect more ancient allele lineages in these populations, which is in support of the out-of-Africa theory of evolution.

The distribution of haplogroups between populations is shown more clearly in Figure 4.4. Haplogroups (5′ sites) are distinguished by colour, and subgroups (3′ sites) are distinguished by pattern. Afro-Caribbean alleles show more variability of 5′ flanking haplotype (colour) than the Japanese. This is interesting because a similar number of alleles from both populations were typed. The lack of variability in the 5′ flanking DNA of the Japanese may be due either to a founder effect within this population, or because particular alleles which share flanking haplotypes have spread to high population frequency. The reduced distribution of the 3′ haplotype (pattern) in Afro-Caribbeans is due to the fact that the +744 site is not polymorphic in this population. The diversity of alleles in the Caucasian population is mainly due to the fact that twice as many alleles from this population were typed. Most Caucasian alleles appear to be concentrated in groups 1 and 2, whereas there is a more even spread of all haplogroups and subgroups in the African population. This provides strong evidence that Africans have the highest diversity of all human populations. It may also indicate that there is a founder effect in the Caucasian population, though this is less pronounced than in the Japanese.

The significance of these haplogroups and their frequencies in the different populations is difficult to determine, particularly with the variable sample sizes. It is likely that the more alleles that are typed the more ethnically diverse these groups will become. The large number of haplotypes (77 out of a possible 121 haplotypes) identified in this limited sample size (135 alleles) suggests that haplotypes within these regions undergo frequent disruption via recombination or repeated mutation.

Population specific analysis

Caucasian

As can be seen from Figure 4.5A the majority of Caucasian alleles have very similar flanking haplotypes, which fall into two major haplogroups (Figure 4.4). Further analysis shows significant linkage disequilibrium in the 5′ flanking DNA between sites -1663 to -109 (Figure 4.5D, χ^2 = 40.33, 1 d.f., $p < 0.0001$). There is also a region of strong linkage disequilibrium in the 3′ flanking DNA, between sites MS31b to +1165 (Figure 4.5D, $\chi^2 = 24.03$, 1 d.f., $p < 0.0001$). Between these regions there is a rapid decline in linkage disequilibrium starting at site -220 immediately upstream of the minisatellite which reaches its lowest between site -4 and MS31b. This pattern is mirrored in the sliding window analysis which shows a larger number of haplotypes, indicating increased levels of recombination, immediately adjacent to the minisatellite. This is

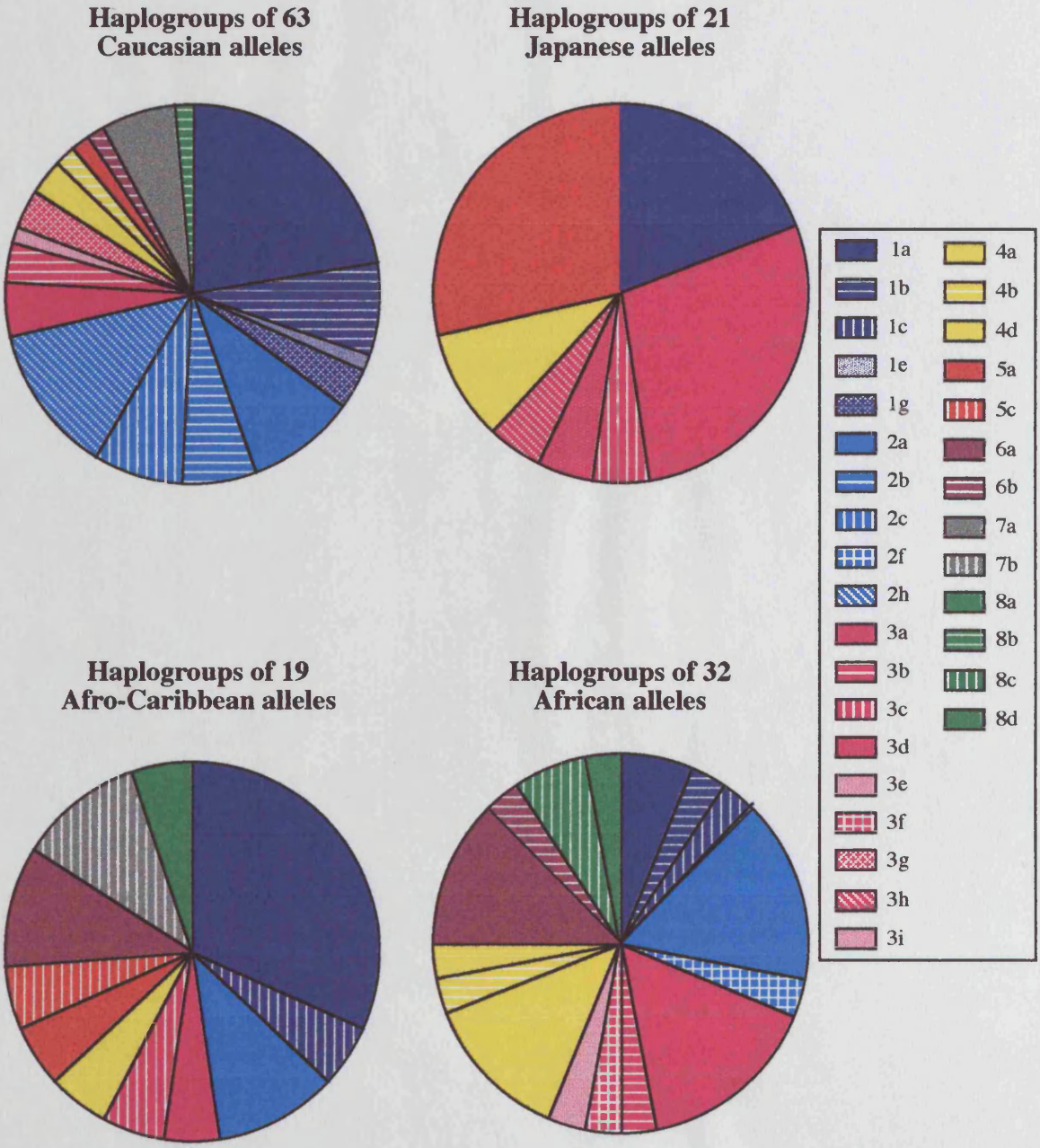
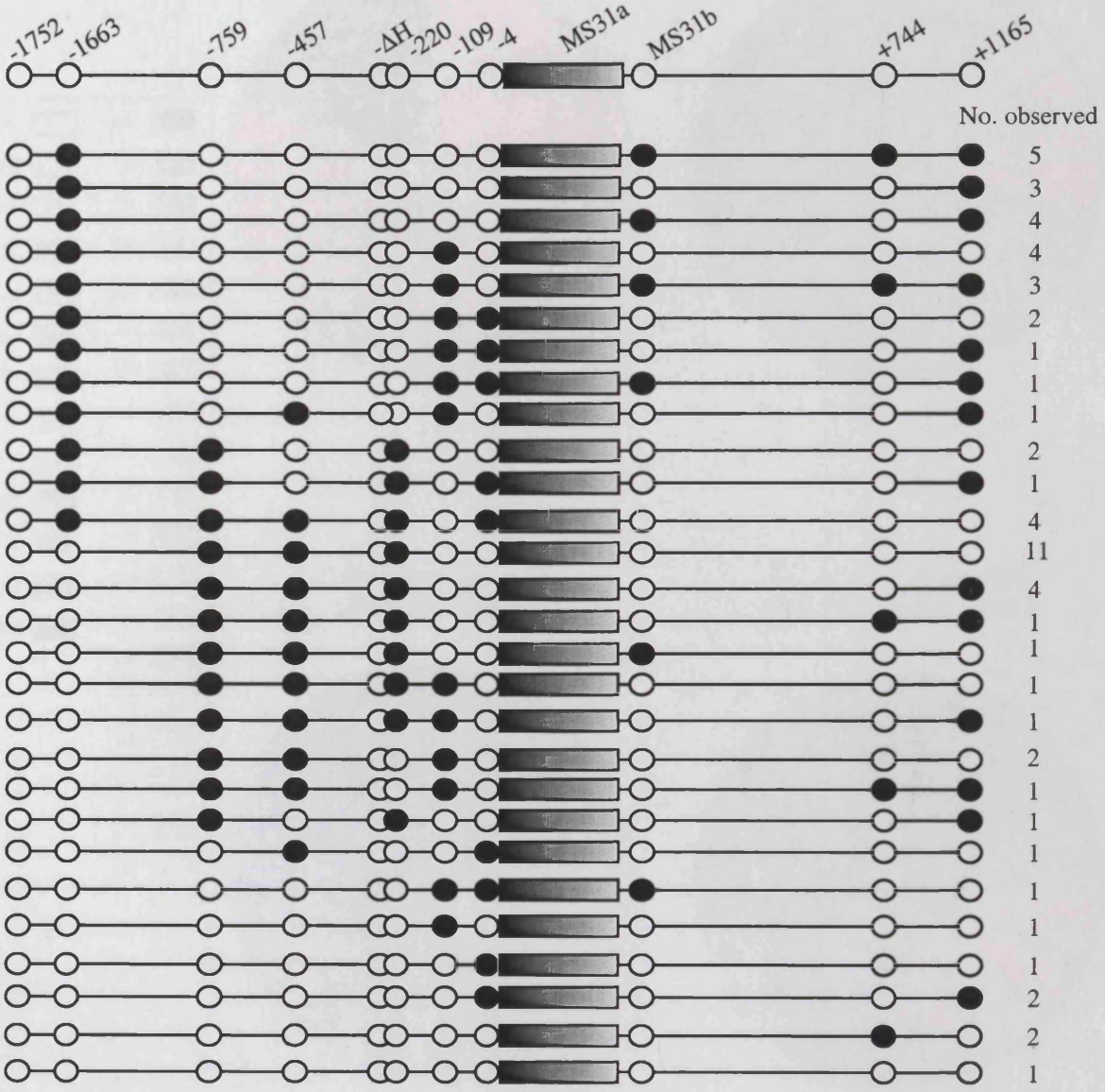


Figure 4.4. Comparison of the MS31a flanking haplogroups.

Each pie chart shows the frequency of all flanking haplogroups observed in the number of alleles indicated from the population shown. Slices are coloured according to 5′ haplotypes (-1663, -759 and -457) labelled 1-8, and patterned according to 3′ haplotype sites labelled a-h (MS31b, +744 and +1165) as shown in the key.

Figure 4.5. Haplotype structure and mutational inferences from nucleotide sequence variation around MS31a observed in sixty-three Caucasian alleles.

A.

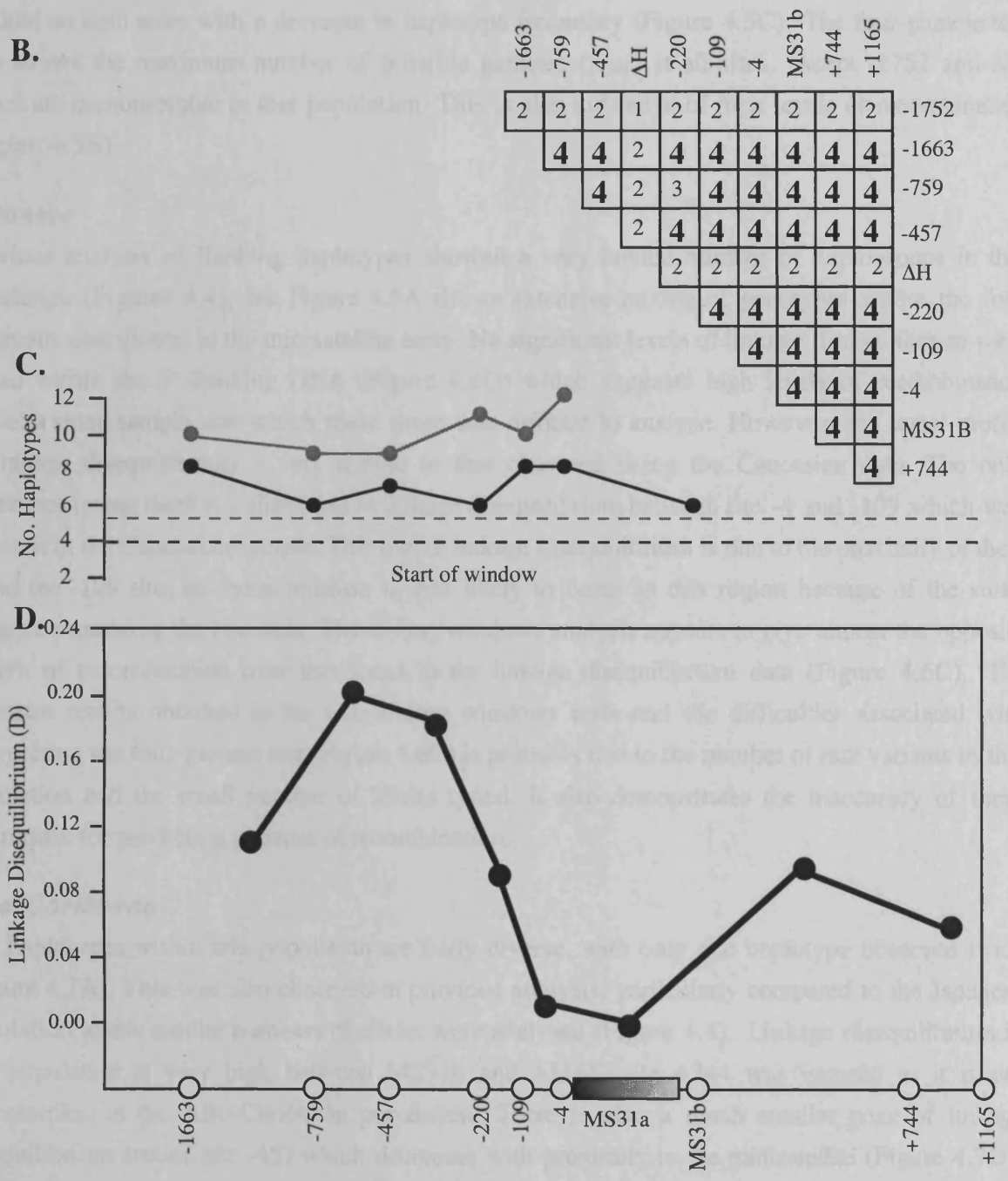


A. Different flanking haplotypes present in sixty-three Caucasian alleles.

The structure of each haplotype and the number observed in this population is indicated. The MS31a allele is represented as a rectangle with the unstable end in black, and the white end representing the more stable end. Polymorphic sites in the flanking DNA are drawn to scale; white circles represent the most common allelic state of each site, and black circles the least common allelic state. Also shown is a representative MS31a allele to indicate the position of each site within the flanking DNA.

B. Plot of the four-gamete test between each polymorphic site.

The number of haplotypes at each site are shown. The maximum possible haplotypes is four, including those involving MS31b in this population. The presence of all four haplotypes in site pairs suggest recombination between each site, these sites are shown in bold type.



C. Observed number of distinct haplotypes in sliding windows of four or three polymorphic sites.

The number of haplotypes seen in sliding windows of both four (grey line) and three (black line) polymorphic sites are shown. The number of haplotypes expected for each window size are shown by dotted lines in the corresponding shade. Observation of regions with a number of haplotypes greater than $s + 1$ (where s is the size of the window) indicates numerous recombination events in the ancestral history of that sequence. The graph is plotted from the first site of each window.

D. Distribution of linkage disequilibrium between flanking polymorphic sites .

The linkage disequilibrium coefficient (D) was calculated as described in the text for each adjacent pair of markers shown on the representative MS31a allele shown below.

flanked on both sides with a decrease in haplotype frequency (Figure 4.5C). The four-gamete test also shows the maximum number of possible gametes (four) at all sites, except -1752 and ΔH which are monomorphic in this population. This is also indicative of high levels of recombination (Figure 4.5B).

Japanese

Previous analysis of flanking haplotypes showed a very limited number of haplogroups in this population (Figures 4.4), but Figure 4.6A shows extensive mixing of genotypes within the four upstream sites closest to the minisatellite array. No significant levels of linkage disequilibrium were found within the 5′ flanking DNA (Figure 4.6D) which suggests high levels of recombination and/or a small sample size which make these data difficult to analyse. However, the actual profile of linkage disequilibrium is very similar to that observed using the Caucasian data. The only difference is that there is a slight rise in linkage disequilibrium between site -4 and -109 which was not seen in the Caucasian sample. This rise in linkage disequilibrium is due to the proximity of the -4 and the -109 site; so recombination is less likely to occur in this region because of the small distance separating the two sites. The sliding windows analysis appears to give almost the opposite pattern of recombination over this locus to the linkage disequilibrium data (Figure 4.6C). The disparate results obtained in the two sliding windows tests and the difficulties associated with interpreting the four-gamete test (Figure 4.6D) is probably due to the number of rare variants in this population and the small number of alleles typed. It also demonstrates the inaccuracy of these techniques for predicting patterns of recombination.

Afro-Caribbean

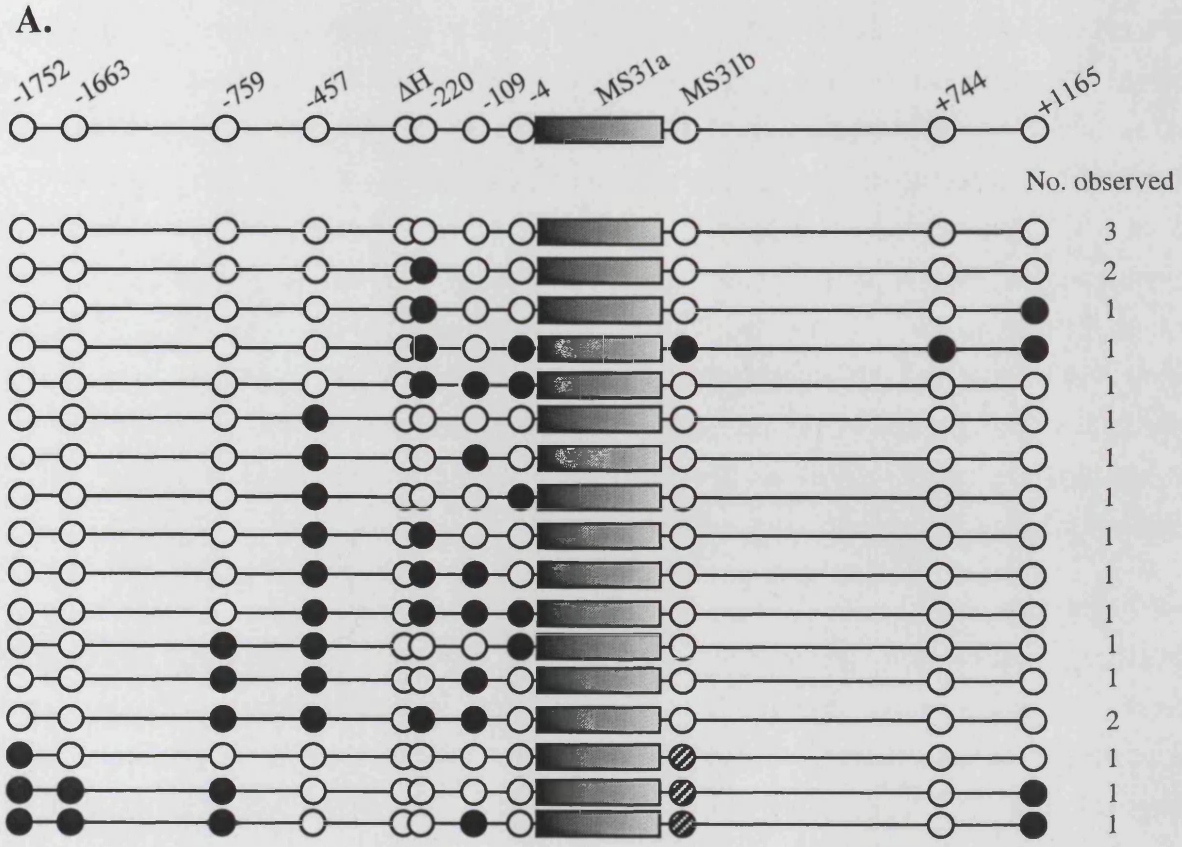
The haplotypes within this population are fairly diverse, with only one haplotype observed twice (Figure 4.7A). This was also observed in previous analysis, particularly compared to the Japanese population where similar numbers of alleles were analysed (Figure 4.4). Linkage disequilibrium in this population is very high between MS31b and +1165; site +744 was ignored as it is not polymorphic in the Afro-Caribbean population. There is also a much smaller peak of linkage disequilibrium around site -457 which decreases with proximity to the minisatellite (Figure 4.7D). However, this trend is again disrupted by an increase in linkage disequilibrium between sites -109 and -4, which then decreases again. This apparent fall in recombination between -109 and -4 is also shown in the four-gamete test for site -4 (Figure 4.7B), and in the sliding windows analysis (Figure 4.7C), but the surrounding levels of recombination remain high. The -4A allele is also relatively rare in this population and this, coupled with the proximity of the -4 and -109 sites, would result in the apparent increase in linkage disequilibrium between these sites. The less obvious pattern of linkage disequilibrium overall is possibly due to the higher number of more ancient and diverse haplotypes in this population.

African

Haplotypes within the African population are very diverse, particularly in the 5′ flanking DNA.

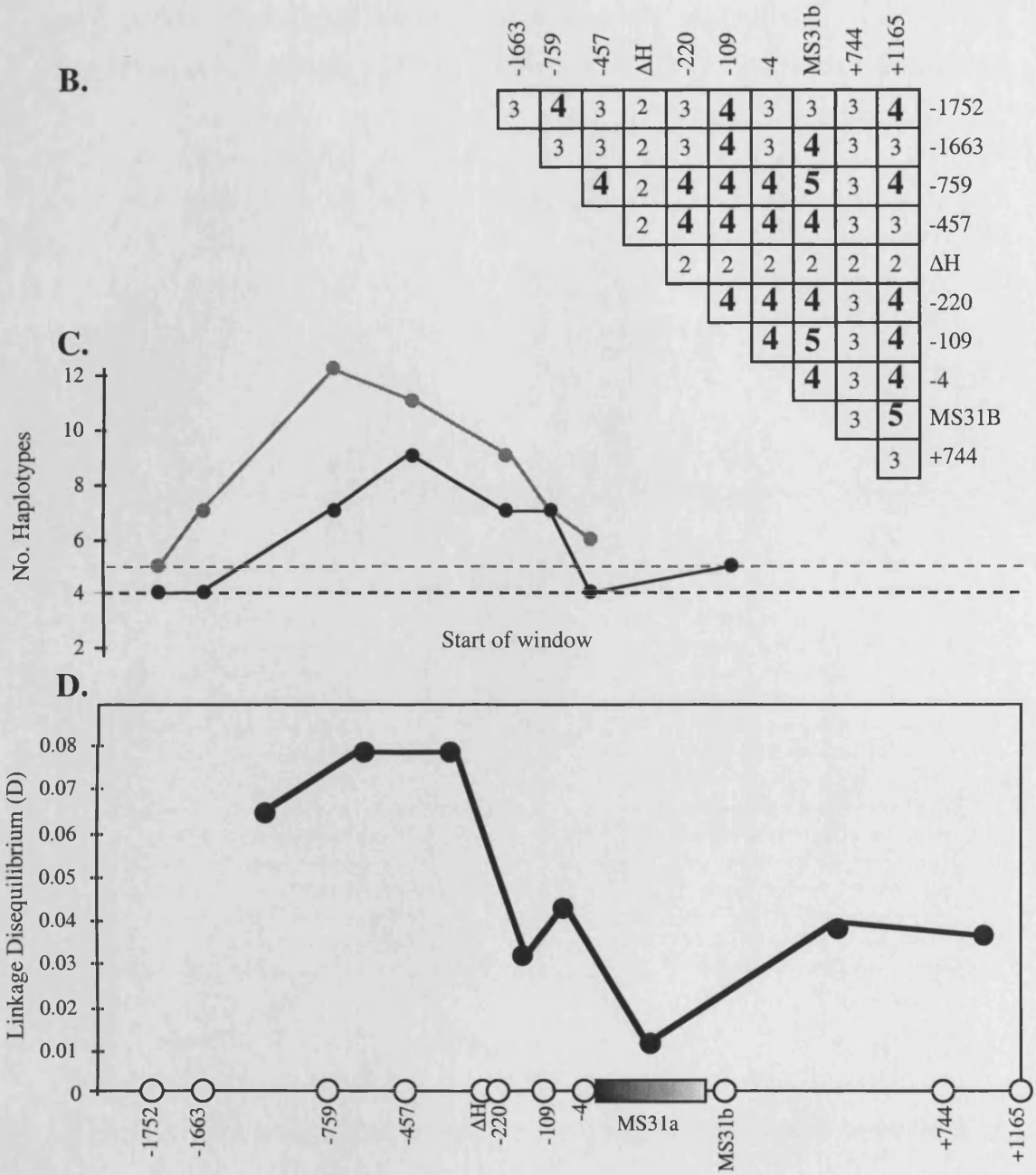
Figure 4.6. Haplotype structure and mutational inferences from nucleotide sequence variation around MS31a observed in twenty-one Japanese alleles.

Figure legends are identical to those for the Caucasian analysis (Figure 4.5) except where indicated.



A. Different flanking haplotypes present in twenty-one Japanese alleles.

Hatched circles indicate MS31b C alleles.



B. Plot of the four-gamete test between each polymorphic site.

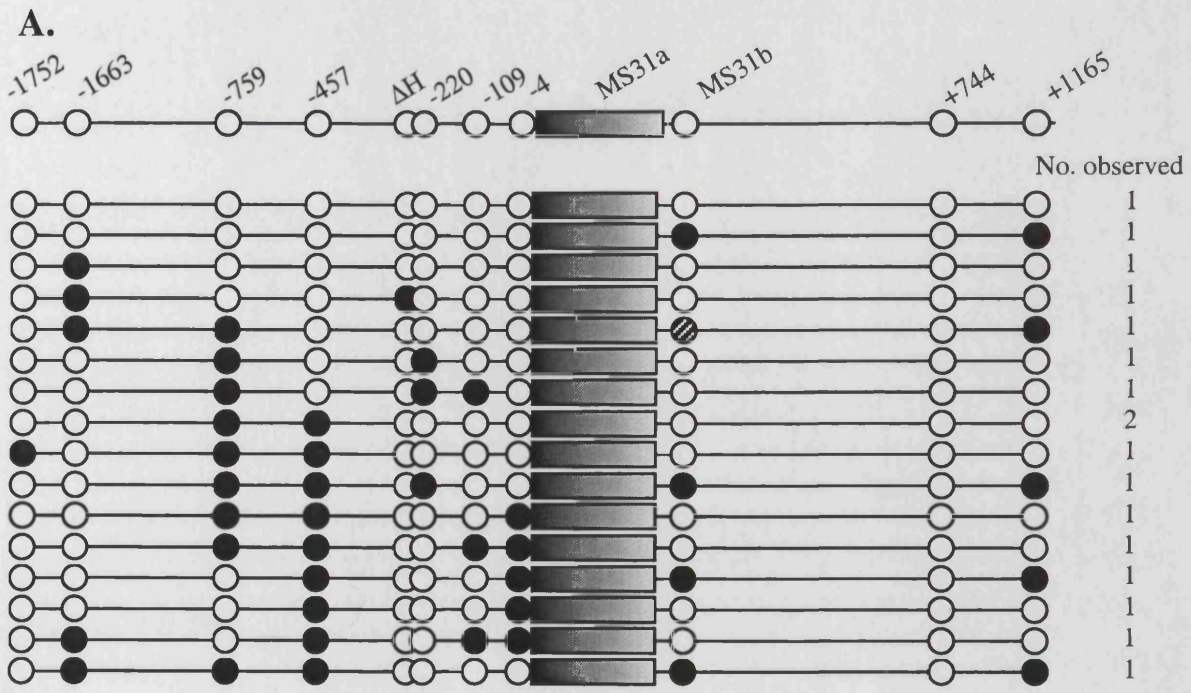
The maximum possible haplotypes is four, except those involving MS31b which has eight possible haplotypes in this population.

C. Observed number of distinct haplotypes in sliding windows of four or three polymorphic sites.

D. Distribution of linkage disequilibrium between flanking polymorphic sites around MS31a.

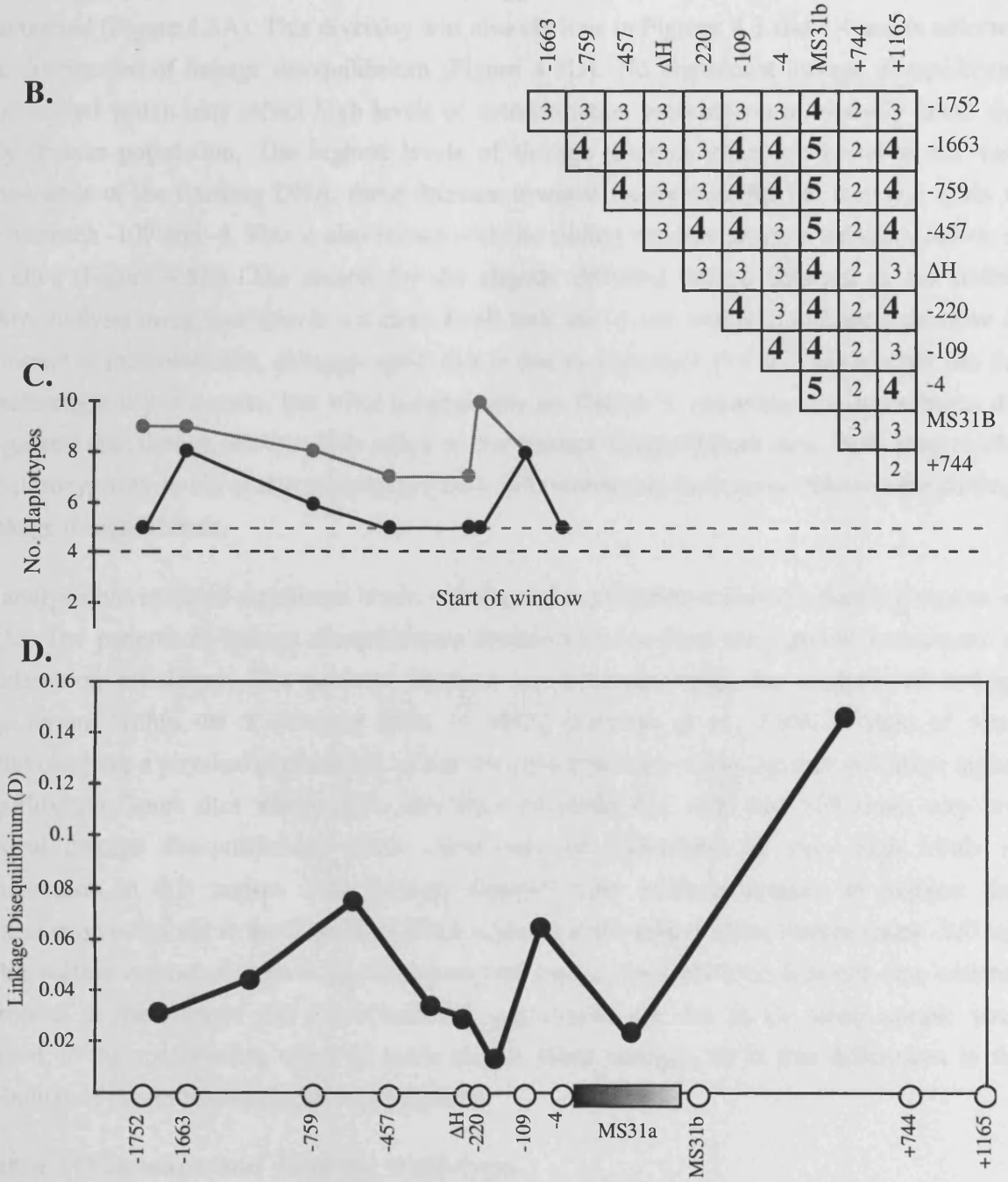
Figure 4.7. Haplotype structure and mutational inferences from nucleotide sequence variation around MS31a observed in seventeen Afro-Caribbean alleles.

Figure legends are identical to those for the Caucasian analysis (Figure 4.5) except where indicated.



A. Different flanking haplotypes present in seventeen Afro-Caribbean alleles.

Hatched circles indicate MS31b C alleles.



B. Plot of the four-gamete test between each polymorphic site.

The maximum possible haplotypes is four, except those involving MS31b which has six possible haplotypes in this population.

C. Observed number of distinct haplotypes in sliding windows of four or three polymorphic sites.

D. Distribution of linkage disequilibrium between flanking polymorphic sites around MS31a.

There are only two examples of the same haplotype being seen more than once in 29 alleles characterised (Figure 4.8A). This diversity was also obvious in Figures 4.3 and 4.4 and is reflected in the distribution of linkage disequilibrium (Figure 4.8D). No significant linkage disequilbrium was observed which may reflect high levels of recombination working on a relatively small and highly diverse population. The highest levels of linkage disequilibrium are found at the very extreme ends of the flanking DNA, these decrease towards the minisatellite but then rise again to peak between -109 and -4. This is also shown with the sliding window analysis using a window of three sites (Figure 4.8B). The reason for the slightly different picture obtained in the sliding window analysis using four sites is not clear. In all tests the -4 site seems to indicate a decrease in the amount of recombination, although again this is due to a proximity of this site to -109 and the low heterozygosity of the site. Site +744 is apparently unaffected by recombination according to the four-gamete test, though this has little effect on the linkage disequilibrium data. Both sites exhibit low heterozygosity levels in this population (Table 4.1) which may have some effect on the patterns of linkage disequilibrium.

This analysis has revealed significant levels of linkage disequilibrium within the flanking regions of MS31a. The patterns of linkage disequilibrium demonstrate that there are regional fluctuations in recombination efficiency. The patterns obtained are consistent with the analysis of linkage disequilbrium within the 5′ flanking DNA of MS32 (Jeffreys *et al.*, 1998b). Most of these fluctuations have a physical explanation, so that the sites which are closer together will show higher disequilibrium. Some sites which are in very close proximity e.g. -220 and -109 retain very low levels of linkage disequilibrium, which could only be maintained by very high levels of recombination in this region. This linkage disequilibrium analysis appears to suggest that recombination is highest in the 5′ flanking DNA adjacent to the minisatellite, between sites -220 and -4. This pattern is most obvious in the Caucasian and Japanese populations. It is not clear whether differences in the African and Afro-Caribbean populations are due to the small sample sizes analysed, to the confounding effect of more ancient allele lineages, or to true differences in the distribution of recombination in these populations.

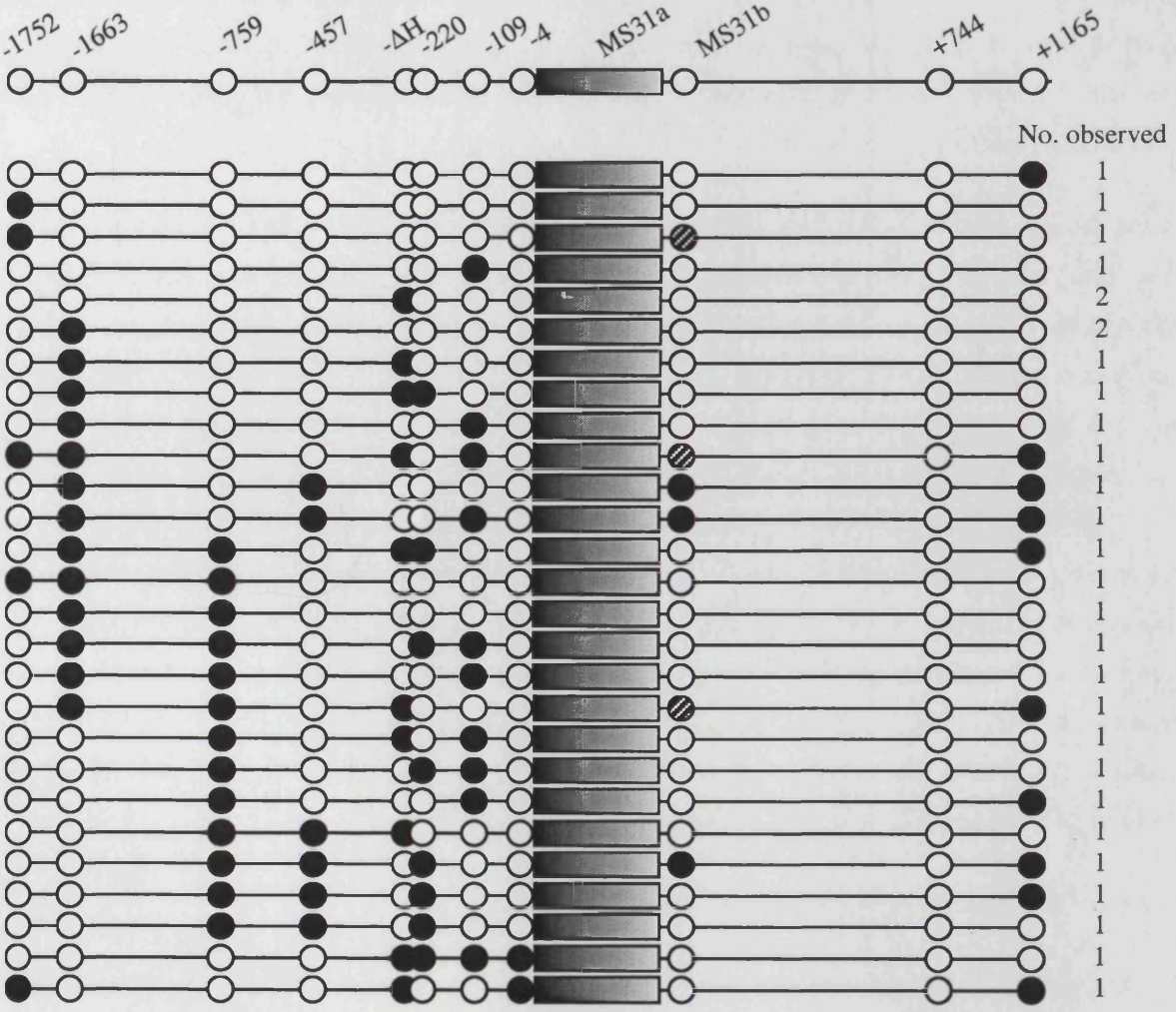
Reverse MVR maps and flanking haplotypes

The haplotyping data can now be combined with the analysis of minisatellite allele structure carried out in Chapter 3. This will identify putative associations between variation within and external to the minisatellite array (Figure 4.9). Unfortunately, this analysis is complicated by the fact that most of these flanking data is incomplete. Some groups do share flanking genotypes, but there is no overall pattern of shared haplotypes within groups. Alleles in Group 1 which have been typed for flanking markers share flanking haplotypes at both ends of the minisatellite. In addition, the 3′ flanking haplotype 5′ AGT 3′ predominates in Group 8; while the 3′ haplotype 5′ BCG 3′ predominates in Group 7. More interesting is the haplotype switching observed in some alleles which share identical array structures e.g. between individuals in Group 7 and Group 3. This is

Figure 4.8. Haplotype structure and mutational inferences from nucleotide sequence variation around MS31a observed in twenty-nine African alleles.

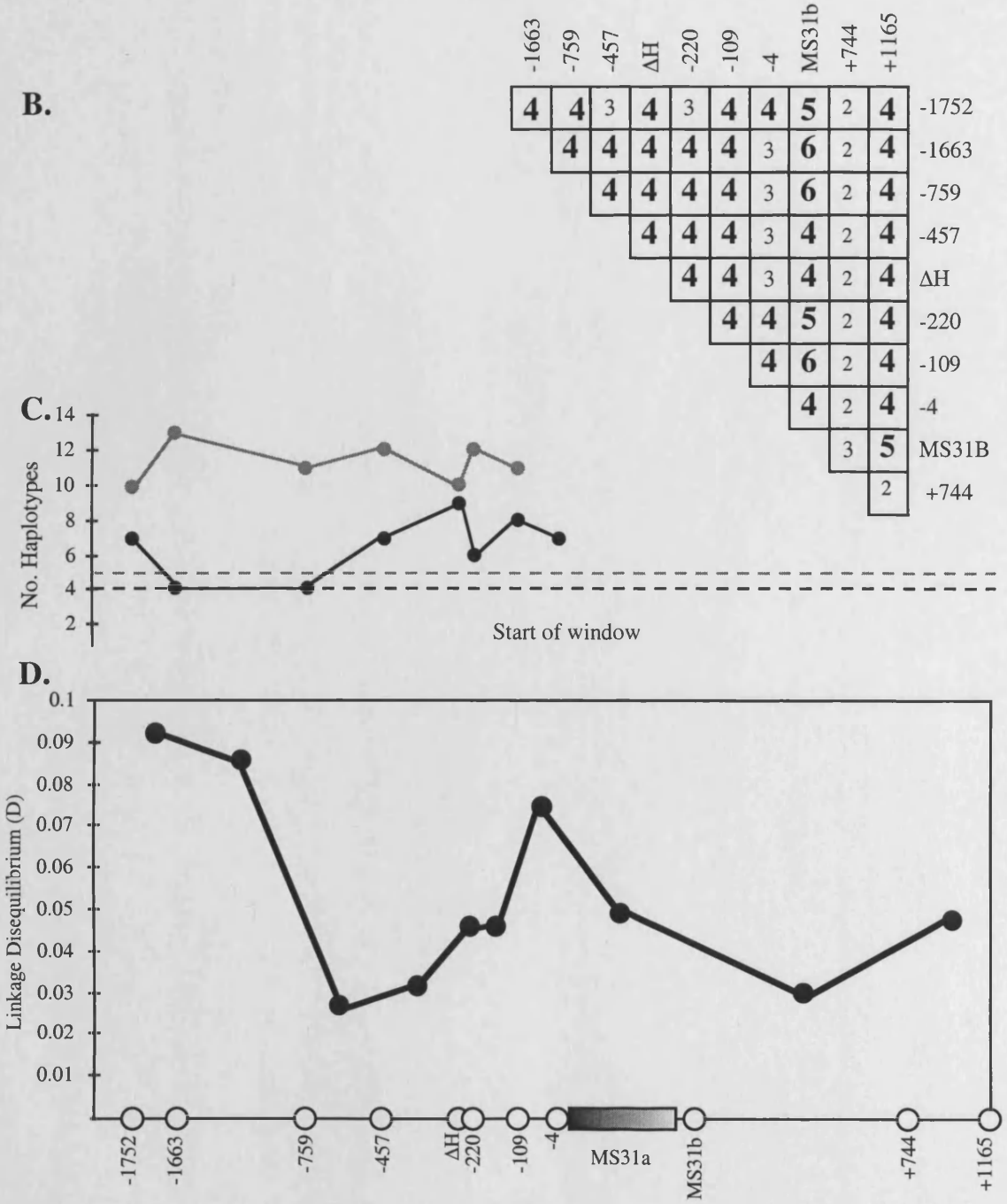
Figure legends are identical to those for the Caucasian analysis (Figure 4.5) except where indicated.

A.



A. Different flanking haplotypes present in twenty-nine African alleles.

Hatched circles indicate MS31b C alleles.



B. Plot of the four-gamete test between each polymorphic site.

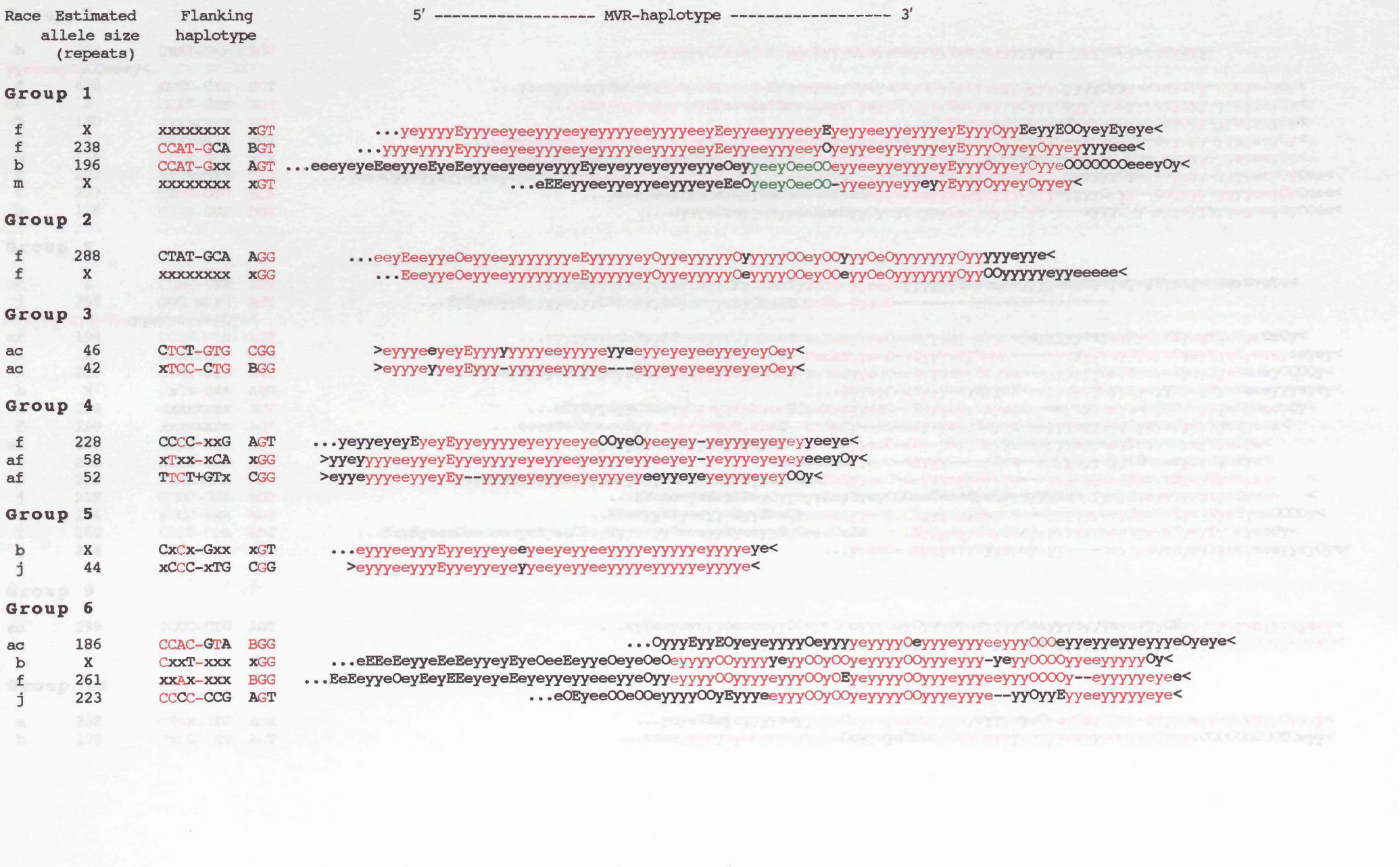
The maximum possible haplotypes is four, except those involving MS31b which has six possible haplotypes in this population.

C. Observed number of distinct haplotypes in sliding windows of four or three polymorphic sites.

D. Distribution of linkage disequilibrium between flanking polymorphic sites around MS31a.

Figure 4.9 Databank of MS31a reverse MVR maps including flanking haplotypes.

Groups of MS31a alleles aligned by dot matrix analysis, as described in Chapter 3 and in the legend for Figure 5. Gaps (-) have been introduced to improve alignments. For each allele its ethnic origin; British (b), Mormon (m), French (f), Japanese (j), Afro-Caribbean (ac), African (af): and MVR haplotype are shown. **E** = AT type repeat; **y** = GC type repeat; **e** = GT type repeat; **O** = null repeat; ..... = allele continues beyond mapped region; > = 5′ end of short allele; < = 3′ end of allele. Array sizes are given in repeats. Flanking haplotypes, in the order: 5′ flanking DNA -1752C/T, -1663C/T, -759A/C, -457C/T, ΔH+/-, -220C/G, -109C/T, -4A/G; 3′ flanking DNA MS31bA/B/C/D, +744C/G, +1165G/T are shown; x = position not typed. Predominant regions of identity within a group are shown in red; additional regions of identity shared between different subgroups are shown in blue, green and yellow; positions of divergence are shown in black.


Race
Estimated allele size (repeats)
Flanking haplotype
5′ ---------- MVR-haplotype ---------- 3′
Group 1
f X xxxxxxxx xGT ...yeyyyyEyyyeeyeeyyyeeyeyyyyeeyyyyeeyEeyyeeyyyeeyEyeyyeeyyeyyyeyEyyyOyyEeyyEOOyeyEyeye<
f 238 CCAT-GCA BGT ...yyyeyyyyEyyyeeyeeyyyeeyeyyyyeeyyyyeeyEeyyeeyyyeeyOyeyyeeyyeyyyeyEyyyOyyeyOyyeyyyyeee<
b 196 CCAT-Gxx AGT ...eeeyeyeEeeyyeEyeEeyyeeyeeyeyyyEyeyeyyeyeyyeyeOeyyeeyOeeOOeyyeeyyeyyyeyEyyyOyyeyOyyeOOOOOOOeeeyOy<
m X xxxxxxxx xGT ...eEEeyyeeyyeyyeeyyyeyeEeOyeeyOeeOO-yyeeyyeyyeyyEyyyOyyeyOyyey<
Group 2
f 288 CTAT-GCA AGG ...eeyEeeyyeOeyyeeyyyyyyyeEyyyyyeyOyyeyyyyyOyyyyyOOeyOOyyyOeOyyyyyyyOyyyyyeyye<
f X xxxxxxxx xGG ...EeeyyeOeyyeeyyyyyyyeEyyyyyeyOyyeyyyyyOeyyyyOOeyOOeyyOeOyyyyyyyOyyOOyyyyyeyyeeeee<
Group 3
ac 46 CTCT-GTG CGG >eyyyeeyeyEyyyyyyyyeeyyyyeyyeeyyeyeyeeyyeyeyOey<
ac 42 xTCC-CTG BGG >eyyyeyyeyEyyy-yyyyeeyyyye---eyyeyeyeeyyeyeyOey<
Group 4
f 228 CCCC-xxG AGT ...yeyyeyeyEyeyEyyeyyyyeyeyyeeyeOOyeOyeeyey-yeyyyeyeyeyyeeye<
af 58 xTxx-xCA xGG >yyeyyyyeeyyeyEyyeyyyyeyeyyeeyeyyyeyyeeyey-yeyyyeyeyeyeeeyOy<
af 52 TTCT+GTx CGG >eyyeyyyeeyyeyEy--yyyyeyeyyeeyeyyyeyeeyyeyeyeyyyeyeyOOy<
Group 5
b X CxCx-Gxx xGT ...eyyyeeyyyEyyeyyeyeeyeeyeyyeeyyyyeyyyyyeyyyye<
j 44 xCCC-xTG CGG >eyyyeeyyyEyyeyyeyeyyeeyeyyeeyyyyeyyyyyeyyyye<
Group 6
ac 186 CCAC-GTA BGG ...OyyyEyyEOyeyeyyyyOeyyyyeyyyyOeyyyeyyyeeyyyOOOeyyeyyeyyeyyyeOyeye<
b X CxxT-xxx xGG ...eEEeEeyyeEeEeyyeyEyeOeeEeyyeOeyeOeOeyyyyOOyyyyyeyyOOyOOyeyyyyOOyyyeyyy-yeyyOOOOyyeeyyyyyOy<
f 261 xxAx-xxx BGG ...EeEeyyeOeyEeyEEeyeyeEeyeyyeyyeeeyyeOyyeyyyyOOyyyyeyyyOOyOEyeyyyyOOyyyeyyyeeyyyOOOOy--eyyyyyyeye<
j 223 CCCC-CCG AGT ...eOEyeeOOeOOeyyyyOOyEyyyeeyyyOOyOOyeyyyyOOyyyeyyye--yyOyyEyyeeyyyyyyeye<


Group 7
b 203 CxAT-Gxx BCG
f 496 CCCC-CTG BCT
b X CCAC-Cxx xCG
b 180 xxxxxxxx BCG
f X xxxxxxxx xGx
m 176 CCAx-GCA xCx
f 227 xxxxxxxx BCG
f 237 CTAT-GTG BCG
b 126 CTAT-GCG BCG
Group 8
ac X CCAC-GTA AGT
j 254 CCCCxGxG AGT
af 194 CCAT-GTx AGT
m X xxxxxxxx xGx
b 229 CCCC-CTG AGT
b X CxCx-Gxx xGG
m 282 xxxxxxxx AGT
f 290 xxxxxxxx AGT
ac 373 CCCT-GTG AGT
j 242 xCAT-xTG AGT
f 280 xxxxxxxx AGT
j 209 CCCC-GCG AGT
f 284 xCCC-xxx AGT
f 280 CCAT-GTA AGG
j 216 CCAx+GCG BGT
Group 9
ac 299 xCCC-CTG AGT
m X xxxxxxxx xCx
Group 10
m 238 CTAx-GTG xGx
b 128 CxAC-Gxx AGT

Unalignable alleles:

```
ac      182     TCCC-GTGAGT     ...EyOEeeEOEeEeEyyeeyyEyyyeyyyeyeeyyyyeyEyyeyyeyyyeyyeeeyyyyeyEyyeyEyeOyeOOOOy<
 f       X      xxCx-xxxAGT     ...eyeEyeEeyOyeeeeeeyeEEyyeOyyyyeyyEyeeyyyeeeyyeEyeEyyeyOOOy<
 b     248      xxxxxxxxBGT     ...eEeEeeeEyEeeeyeEeyEeeEEeyOyyeyeyEyeyeyeeyEyeyeyEyEyeeyyeEyyyyeyyyyeEEyyyeeeeyeyeeEyeyEyeEyeeeOOy<
ac      234     CTAC-GxGBCG     ...eOeeyeyeOeyyEeEeyeeyyeeyyeEEeEeeeEEeeeyyOyeyeeeyyyOyyyyyOyyOyyeyyeyyyeeyOy<
af       X      xTxx-xCAxGT     ...yyeyeOOyyeOyyyeOyyyOyeeyyyeyOOyyyeOyyeyyeyyEyyOyyeyOyyeyOyyyOOyyyeeeyOy<
 j     253      CCAT-CTABCG     ...EEeeyyeeyyEeeEeyyyyyeyyyEeyyyyOOOyyeyyyyyyeeeyOy<
 f     169      xxxxxxxxBGG     ...EeeEeyOyOOyyyOeyeyyeyyyyOEyOyyyyEyeOyyyOyyyeeyyEOyyyeyyyeyeyOy<
 m     175      CTAT-GTAAGT     ...EeyEeyeyeEeyEeyEeyeyeeEeeEEeyyeeyyyeyyeeyyyyeyeeOyeeyOeeOyeeyeyeyEyyOyeyyeyOOOOOOOOOOOOeyOy<

 f     319      CTAx-GxABGG     ...yeEeeEeEeeeEEyyeeEEyyyeeyEeyeeyeyeyyeyyEyeyeEEyeeEyeyeyeeyyeyeyOe<
 m     234      CTAT-GCGBCG     ...eyEyEyEyEyeyeeyyeyeyyeeyeyyeyeeyeyeyeEeyEeyyeyeeeeyeeyeyyeyeeyOyyeyyeyyOOyyeeyeyyeyeeyyOe<

 m     304      xxxxxxxxBGG     ...yeEEyyeeEeeeeEeeEyyeEyyeEeyEeeeeeyEeeyyyyeyyyyyEyeyOOyyyyyOyyyyyyyyyyOyEyyyeyyeyyeeyyyeyyeeyyOOOyyyeyyyyyeyyyeeyey<
 j      43      TCAT-GTGCGG     >eyyyeeyyeyEyyeyyyeyeyyeyyyyyyyyyeyeyeyyyye<
 f     253      CCCC-CTGAGT     ...eeyyyyeyeEeyeeeeeyyyeeyeeeEyeyeEyeyyOeOyey<
 m      X       xxxxxxxxxxG     ...yyyyyeeeeyeyeeyyyeeeOyeEeeyyeeyyyeyyeeyeeOeeOeyeyyeyeeOeyeeeeyyeeyyyeyey<
 b      81      xxxxxxxxAGG     ...OOOOOEeeEEeyeyEeyeyeyEyEeyyeEyEEyyyyOyOyyyOyyyeyeeyyyOyyOyyeyyyyOOyyyyOOOyyyyyeyeyOeey<

 b      75      CCAT-GxxBCG     ...yeeeeyeyeyeyeeyeeeyeyyeyyeyyOyeeeOeyeeyyyyeeyyeeyyeyeeyOyee<
 f     194      CTAT-GTGBGG     ...EeEeyyeyeyeEeyEeyeeEeyeyeyeOeOeyyOyyyeyyOyOyyOOOyyyyyyyeyyyOOOyyyyyyyOyee<
 f      X       xTAx-xxxBGG     ...yEOyeyeOyyyOeyOOyyyyyOeyyyeyyeeyyOOyyOyyyeyyeyyyyOyyyyeeyyyOOOyyyyyeyeeyOyee<
 b      X       CxxT-xxxxGT     ...eeyEyEyeyOOyyeeyeEyEyeyeyeyeyyeeyEyeyyEEyyyeyyOEyyyyyEeEyyEEEyeyEyeEyeee<
 f      X       xxxxxxxxxxT     ...yeeOyeyeyeeyEeyeeeyyyeeyyeeyEyeyeeeeeEyeyeee<
 m      X       xxxxxxxxxGG     ...yEyEyyyeyyyyeeyyyyyyeeOOyOOeeyyOyeyyyeeyyyyyyyOyyOeyeyyeyyyyyyyyyeyyOyyyeyyyeyyyeyeeyOyOee<

ac      X       CCCC-CTGBGG     ...eeyyEyeeyEyyyyEyeyyyeyyyOOEOeyOOOOOyEyyyEyyeEyyyyyyeyeyOyyyeeOyeyeyyeyeyyeyeyye<
 f     187      CTxC-GTxBGG     ...eeyyyeyyyyyeEyyyyyyOOyyyyyyOyyyyyOOyyOOyyyyOyOyyyyyyyyyOOOyyyyyye<
 b      X       CCAC-CxxxGT     ...yeOeEyyeyyEyeeeyyyeeyyeeyEyeyeOEyeyEye<
 b     313      xxxxxxxxAGT     ...EyeOOEyyyeyyEyeeyyyyeyyyeeyEyeyeEyyyEyyEye<

ac       75     CCAT-GTGAGT     ...OOOOOyyyyyeyyyyeyOOyyeyeyyyyeyyyyyeyyyyyyOOOOeyyOOOOeyyyeyyeeyyyOOOOOOOyyyeyeyOe<

 m       X      xxxxxxxxxxT     ...yeyyyeyeyeeyeOeyyeyEEyyeOyeyyyEyyyeyyeeyyyyeyyeeyeyyyyeyyeeyeEyyyeeeyeyyeyEyyyOyyeyyyyeyyyyyyyOOEyyyyyyyyy<
```

suggestive of recombination processes operating in the flanking DNA to separate these haplotypes. Other allele pairs suggest that minisatellite mutation is restricted to the array, for example the last few alleles in Group 7, or can extend from the array and into the flanking DNA, for example Group 8. This switching has previously been observed in the 5′ flanking DNA of this minisatellite (Neil, 1994), but this is the first evidence to suggest that recombinogenic processes may also be found in the 3′ flanking DNA. This suggests that mutation at MS31a may be bipolar because it is widely believed that minisatellite mutation is driven by processes in the flanking DNA.

Attempts to group reverse MVR maps according to flanking haplogroup failed to find any relationship between flanking haplotype at either end of the minisatellite and minisatellite structure (data not shown). This suggests that instability at this locus is found at both ends of the minisatellite array.

Discussion

This analysis is subject to similar limitations found in the previous chapter, namely that the number of alleles sampled was small, particularly in the non-Caucasian populations. In addition, sequencing to identify these sites was performed on a limited number of predominantly Caucasian individuals (see Chapter 3; David Neil, pers. comm.). Consequently, all SNPs within this region in all different populations cannot be identified. Site -4 is one such site which, because of its low heterozygosity, would probably not be identified by the conventional sequencing methods used to identify other sites. It was only identified because it created a rare *Alu* I RFLP during preliminary analysis of MS31a (Armour *et al.*, 1989b).

Mutational inferences

Analysis of the haplotype data suggested that high levels of recombination may exist in the 5′ flanking DNA of MS31a in all populations. Detailed examination of the flanking DNA showed a general profile of linkage disequilibrium consistent with that seen at MS32; this was particularly obvious in the Caucasian population. However, it must be noted that patterns of linkage disequilibrium are poor indicators of recombination within a locus. They cannot be used to determine the nature or timing of a recombination event within a population (Clark *et al.*, 1998; Jeffreys *et al.*, 1998b). The only reasonable deduction which can be made is that high linkage disequilibrium is a marker of recombinatorial inertia, as proved at MS32 (Jeffreys *et al.*, 1998b). A true profile of recombination at this locus will therefore only emerge by direct analysis of germline crossover mutants. To date, analysis of length change mutants isolated from pedigrees has shown that there is a hotspot of gene conversion within the 5′ end of the minisatellite array (Neil, 1994). This work has also identified two pedigree mutants which exhibit exchange of 5′ flanking markers; one of which was strongly suggestive of a true crossover event, comprising 10 kb of one allele fused to 2 kb of the other (Jeffreys *et al.*, 1998a). Flanking crossover events have been examined

in great detail in sperm DNA at MS32, revealing the presence of a localised crossover hotspot within the 5′ flanking DNA of the minisatellite. This is immediately adjacent to the most unstable end of the tandem repeat array and appears to drive the instability of the array (Jeffreys *et al.*, 1998a and b). The analysis of linkage disequilibrium carried out during this work suggests that a similar profile of recombination may exist in the flanking DNA of MS31a.

If recombination in the flanking DNA is the mechanism which drives instability of minisatellite repeat arrays then there is evidence that instability at MS31a may be bipolar. Firstly, alignment of the MVR mapped alleles showed that there was no polarity of variability within the repeat array (Chapter 3; Neil, 1994). Secondly, flanking haplotype switching between related allele structures at both ends of MS31a suggests that recombination can occur in the DNA flanking both ends of the minisatellite. This is reinforced by evidence which suggests that recombination levels are elevated throughout the whole of the MS31a locus. Conversely, pedigree studies have shown that mutation is clustered at the 5′ end of the repeat array (Neil, 1994). This suggests that there must be a bias of mutation towards the 5′ end of the array; while mutation at the 3′ end of the array must be of high enough frequency to maintain the elevated levels of structural heterogeneity observed in Chapter 3.

On average, approximately one base substitution is expected every 1 kb along the human chromosome (Kwok *et al.*, 1996). These sites are referred to as polymorphic when the allele frequency of the most common variant is < 99% (Li & Graur, 1991). The fact that the frequency of SNPs in the 5′ flanking DNA of MS31a is significantly above average may be explained in two ways. First, that the level of mutation has risen concomitantly with predicted levels of recombination. Second, that the mutation rate is normal but variants can spread to high frequency by recombination and are less likely to be lost by stochastic processes, such as genetic drift. Recombination may also promote the spread of variants by meiotic drive, by preferentially using the rare variant as a template for mismatch repair of SNPs in heteroduplex DNA; though there is no evidence to support this. The elevated levels of recombination observed throughout the MS31a locus suggest that this is the explanation for the high frequency of SNPs. However, this explanation does not correlate with fact that the frequency of SNPs in the 3′ flanking DNA is normal. It is possible that the mutation rate in the 3′ flanking DNA is suppressed. Alternatively, the recombination events predicted at this end of the minisatellite may be ancient, therefore recombination no longer occurs to facilitate the spread of new SNPs.

Further work

Before carrying out additional population studies, improvements could be made in the methods used to haplotype individual alleles. This could be achieved using two different techniques. The first is using allele-specific multiplex PCR analysis to genotype and determine the phase of multiple sites in a single amplification reaction. The second is selective hybridisation, which uses short (~18 bp), labelled allele-specific oligonucleotides (ASO's) that bind specifically to one allele of a SNP

during hybridisation. This means that many different chromosomes can be analysed in a single hybridisation experiment. Multiple sites can also be analysed concurrently using labelled ASO's for different variant markers in different hybridisation reactions. These techniques allow the efficient haplotyping of large numbers of alleles, and are also flexible enough to include new variant sites, as and when they are discovered. The identification of new polymorphic sites, whether they lie within the regions already investigated or further from the minisatellite, is essential for more extensive analysis of the minisatellite flanking DNA. For example, this work suggests that linkage disequilibrium increases with distance from the minisatellite, so linkage disequilibrium between sites further from the minisatellite array could be assessed. The linkage disequilibrium studies in the African, Afro-Caribbean and Japanese individuals also need to be extended to include more individuals. This will determine whether the patterns of linkage disequilibrium are the same in all populations, or if the African and Afro-Caribbean populations show a different distribution of linkage disequilibrium within the minisatellite flanking DNA.

The most informative method of examining instability at MS31a would be to investigate germline mutation directly. This can be done by small-pool PCR analysis of diluted sperm DNA from selected sperm donors, to isolate and characterise large numbers of MS31a length mutants. Recombination at the MS31a locus can also be directly analysed using a method developed and tested at MS32 (Jeffreys *et al.*, 1998b). This isolates true crossover molecules in sperm DNA from a single individual using pairs of allele-specific primers. The upstream primer is complementary to one allele and the downstream primer is complementary to the other. The flanking SNPs which are essential for this analysis have been identified and characterised during this work. Studies of germline mutation at the MS31a locus will be described in Chapter 5.

Chapter 5

Mutation and recombination at the MS31a locus

Summary

Despite extensive pedigree studies carried out at MS31a, more informative techniques of analysing germline mutation directly at minisatellites have not been attempted. To this end, mutation at MS31a was examined in the sperm DNA of a single individual, concentrating in particular on the analysis of recombination at this locus. The sperm donor, LRIs137 was carefully chosen for these studies on the basis of MS31a minisatellite array length and heterozygosity of flanking polymorphic sites. Rearrangements were initially isolated by conventional methods of mutant detection: length-change mutants were isolated by small pool PCR and a selection of these were characterised in detail. The majority of mutants characterised showed polarised, complex gene conversion-like events typical of minisatellite length-change mutants, including those seen during pedigree analysis of MS31a. A more in-depth study was then carried out to identify and characterise true crossover events at this locus. Firstly, the smaller allele was enriched by physical size separation from the large allele. Crossover molecules of the same or similar size to the small allele were then isolated from this size-selected DNA by a series of nested allele-specific PCRs to specifically select for molecules consisting of the 5′ end of the large allele fused to the 3′ end of the small allele. The structure of each crossover molecule was then dissected by MVR mapping of the minisatellite repeat array and RFLP analysis of flanking markers. This revealed that the crossover points of recombination were densely clustered within the flanking DNA but tailed off into the beginning of the minisatellite array. Peak crossover activity appeared to be located between 100 and 220 bp upstream of MS31a in the flanking DNA, and was elevated 35 times the genomic average. This provided strong evidence for the existence of a recombination hotspot at this locus.

Introduction

Studies on minisatellite loci used in paternity testing first showed that these repeat elements could be highly unstable (Jeffreys *et al.*, 1988) with mutation rates up to 13% per gamete at some minisatellites (Vergnaud *et al.*, 1991). This posed problems for using minisatellites in pedigree analysis and parenthood testing, but revealed a useful tool for studying mutation at tandem repeat loci. This high rate of mutation to new length alleles allowed germline mutants of minisatellites to be easily detected and isolated. This was initially carried out by pedigree analysis using mother, father, child trios (Neil, 1994; Jeffreys *et al.*, 1988), although gathering sufficient information to make general comments on the nature of germline mutation by this method is labour intensive and rather unproductive. Germline mutants can also be isolated

directly from sperm DNA by PCR analysis of multiple aliquots containing single molecules of sperm. More effective alternatives have been developed for the detection and isolation of minisatellite length-change mutants directly from genomic DNA. These techniques are small pool PCR (SP-PCR), and size enrichment with small pool PCR (SESP-PCR). SP-PCR (Jeffreys *et al.*, 1994) involves the dilution of DNA into multiple pools of between 100-200 minisatellite molecules per pool. Amplification, gel electrophoresis and Southern hybridisation of these pools allows both progenitor and abnormal length mutants to be distinguished. This technique can be used for the detection and isolation of mutants occurring at a frequency higher than 10^{-3} per progenitor, and is therefore ideal for the detection of minisatellite length-change mutants in sperm (Jeffreys *et al.*, 1994; May *et al.*, 1996; Buard *et al.*, 1998). SESP-PCR involves releasing the minisatellite with restriction enzymes that cut near the repeat array. Following gel electrophoresis, size fractions are recovered which are completely or partially depleted in progenitor molecules. Any length-change mutants within these fractions can be recovered by SP-PCR (Jeffreys *et al.*, 1990). This approach can detect large deletion and gain events as rare as 10^{-7} per cell, and small gain or loss mutants occurring at a frequency as low as 10^{-5} per cell, depending on the position of the fraction screened in relation to the progenitor (Bois *et al.*, 1997; Jeffreys & Neumann, 1997). Mutants isolated using these techniques can be mapped in detail by MVR-PCR of the minisatellite array and RFLP analysis of the flanking DNA (see previous chapter). This allows multiple mutants to be studied in both germline and somatic tissues.

Minisatellite mutation

Somatic

Minisatellite length-change mutants have been detected in a variety of somatic tissues (Armour *et al.*, 1989a; Jeffreys *et al.*, 1994; May *et al.*, 1996; Buard *et al.*, 1998; Nagel *et al.*, 1995; Kiaris *et al.*, 1996), and have been studied in detail at the human minisatellite, MS32 in blood DNA (Jeffreys & Neumann, 1997). Mutation mainly arises by simple, perfect intra-allelic duplications or deletions of repeat blocks, with a slight bias towards loss of repeats. These seem to arise randomly along the repeat array, and mutants are often mosaic, with multiple isolates possessing the same mutant structure. The most likely mechanism of instability in blood is mitotic recombination; either occurring intra-molecularly or by unequal exchange between sister chromatids (Jeffreys & Neumann, 1997). Replication slippage cannot be excluded, although hallmarks of slippage were not seen in blood, i.e. small events which are targeted to homogeneous regions of the repeat array. However, it remains to be seen whether blood mutation processes are representative of somatic mutation as a whole and whether, as seems likely, the frequency of somatic mutation is linked to cellular proliferation.

Germline mutation in females

Minisatellite instability in the female germline can only be explored through pedigree analysis because of the impossibility of obtaining large numbers of oocytes to study. Understanding of

maternal mutation is further limited because most minisatellites mutate preferentially in the male germline. The few maternal mutants characterised to date, four at MS32 and three at MS31a (Jeffreys *et al.*, 1994; Neil, 1994), show polarity of mutation, although no evidence for complex inter-allelic exchange has been observed. This suggests that germline mutation initiates in the same way but is processed differently in the maternal and paternal germline. However this is purely speculative because of the small number of maternal mutants analysed.

Male germline mutation processes

Minisatellite instability in the male germline has been extensively characterised at five different human loci, MS205 (May *et al.*, 1996), MS32 (Jeffreys *et al* 1991a; Jeffreys *et al.*, 1994), MS31a (Neil, 1994), B6.7 (Tamaki *et al.*, 1999), and CEB1 (Buard *et al.*, 1998). There appears to be two different types of mutation at these loci. **Intra**-allelic events are reminiscent of somatic mutation i.e. mostly simple, non-polar deletions or duplications, which generally form a low-level background of mutation in the male germline. **Inter**-allelic events, on the other hand constitute the majority of germline mutants at most minisatellites, representing up to 90% of gains. A substantial proportion of deletion mutants also arise by inter-allelic transfer accompanied by loss of repeats from the recipient allele. These inter-allelic mutation events exhibit a number of common features which appear to be germline specific, that is they have never been observed in the soma (Neil, 1994; Jeffreys & Neumann, 1997).

Firstly, the minisatellite mutation rate in sperm is much higher than in blood, about 250 fold higher at MS32 (Jeffreys & Neumann, 1997), suggesting these processes are germline specific.

Second, sperm exhibit very different, more complicated mutation processes than in blood. Most of the mutation events in sperm involve the gain or loss of small numbers of repeat units, with a bias towards gains. This occurs by the transfer of one or more repeat units from one allele to another in a gene conversion-like process. In a single mutation event this inter-allelic transfer only occurs one way, with one allele acting as the donor and the other acting as recipient, although both alleles can perform either function. Information is copied from the donor during mutation leaving it unchanged, but alterations in the target site of the recipient can be highly complex, involving duplications or deletions, scrambling of donor repeat segments, and the insertion of anomalous repeats. The transfer of information between alleles suggest that mutation is recombination based and probably arises during meiosis.

Third, very rarely have identical mutant molecules, in terms of allele size and structure, of any minisatellite studied been isolated from the germline. This lack of germinal mosaicism for minisatellite mutants is compatible with instability being a meiotic process. Identical mutants that have been isolated are mainly intra-allelic deletions, which probably arose pre-meiotically during replication (mitosis) of proliferative germ cells (Buard *et al.*, 1998).

Fourth, most rearrangements occur at the end of the array showing most population variability. This polarity of mutation varies between minisatellites; at MS32 over 90% of the insertion sites are clustered within the first quarter of the tandem array, whereas at CEB1 only 75% of insertions are found in the first half of the array (Buard *et al.*, 1998).

Fifth, mutation rates can be highly heterogeneous between different repeat arrays of the same minisatellite, indicating that instability is not an intrinsic property of the array. Polarity of mutation suggests it is influenced by *cis*-acting factors in the flanking DNA adjacent to the most variable end of the minisatellite. Such factors have been observed at MS32 with the identification of a G to C transversion located upstream of the minisatellite array which suppresses mutation in *cis*. Repeats cannot be transferred to the C-linked allele (O1C allele) but its ability to donate repeats is not affected (Monckton *et al.*, 1994). The O1C variant suppresses mutation 110 fold compared to O1G alleles, but has no affect on instability in blood (Jeffreys & Neumann, 1997) suggesting that it specifically affects meiotic instability. Similar flanking polymorphisms which appear to be associated either with low mutation rate, and/or with short, similar alleles occurring with high frequency within the population have also been reported at minisatellites pλg3 (Andreassen *et al.*, 1996), MS205 (May *et al.*, 1996). Although it remains to be seen whether these variants are true suppressers of mutation and not simply linked to short alleles which have very low levels of mutation (see next point). *Cis*-acting factors may also promote the alignment of minisatellites prior to mutation which gives rise to the in-phase transfer of repeats between the donor and the recipient observed at both MS31a and MS32 (Neil, 1994; Jeffreys *et al.*, 1994).

Finally, some minisatellites appear to show a threshold level of mutation. For example, at CEB1 and B6.7 it has been demonstrated that mutation rate increases steadily with allele size until it reaches about 40 to 50 repeats and plateaus off (Buard *et al.*, 1998; Tamaki *et al.*, 1999). Pedigree analysis of pλg3 also suggests that the longer alleles of this minisatellite are associated with higher mutation rates (Andreassen *et al.*, 1996). However, MS205 and MS32 have shown no evidence of the effect of allele size and mutation rate, suggesting that the mutation plateau may be lower in these minisatellites (May *et al.*, 1996; Jeffreys *et al.*, 1994).

As with all rules there are also exceptions. At CEB1 and MS205 (Buard *et al.*, 1998; May *et al.*, 1996), the proportion of inter-allelic transfers is comparatively low, 25% and $\geq$ 20% respectively. Buard *et al.* (1998) have shown at CEB1 that intra-allelic events are often complex and occur predominantly in the longer alleles of this minisatellite . These events do not show polarity although intra-allelic duplications do tend to cluster within homogeneous regions of alleles. These features are all reminiscent of trinucleotide repeat instability. However, these results may simply be due to ascertainment bias. CEB1 has an incredibly heterogeneous repeat array (Buard & Vergnaud, 1994), and mutation at MS205 is characterised by the transfer of very small numbers of repeats which is exacerbated by a relatively uninformative MVR-PCR system (Armour *et al.*, 1996). These features make it incredibly difficult to define the origin of

transferred repeats during mutation, and may lead to the bias towards intra-allelic events observed at these minisatellites (May *et al.*, 1996; Buard *et al.*, 1998). Despite these exceptions germline mutation shows a number of features common to all minisatellites. However, little is actually known about what drives minisatellite mutation in the germline and why these processes on the whole are ubiquitous among minisatellites particularly if, as suggested, mutation is not an intrinsic property of the repeat array itself.

Minisatellites and recombination

It is difficult to imagine that minisatellites, as sites of intense conversional activity, have no biological significance. Possible functions have been implied from their distribution in the human genome. *In situ* hybridisation and linkage mapping have shown that hypervariable minisatellites cluster in the pro-terminal regions at, or near the ends of genetic linkage maps (Royle *et al.*, 1988; Armour *et al.*, 1989b; Armour *et al.*, 1990; Amarger *et al.*, 1998). These regions are the sites of initiation of chromosome synapsis and pairing during meiosis (Solari, 1980; Laurie & Hulten, 1985). Minisatellites may therefore be involved in chromosome homologue recognition prior to chromosome homologue pairing. This is thought to occur by strand invasion and homology searching which, in turn promotes recombination between the chromosomes, and is reflected in the elevated rates of recombination in these regions. This ultimately suggests a strong link between minisatellites and recombination. Furthermore, although all autosomes are approximately equally rich in minisatellites, very few minisatellites have been isolated from the sex-specific region of the X chromosome (Donis-Keller *et al.*, 1987; Armour *et al.*, 1990). There are two possible explanations for this. Firstly. the X-chromosome has no partner in male meiosis and therefore can only undergo recombination during female meiosis. Second, minisatellite instability is largely restricted to the male germline (May *et al.*, 1996; Jeffreys *et al.*, 1997; Buard *et al.*, 1998), and minisatellites may not therefore be involved in chromosome synapsis in the female germline. Conversely, the pseudoautosomal region, which is a region of high recombination in male meiosis, is very rich in minisatellite loci (Cooke *et al.*, 1985; Page *et al.*, 1987). It is not clear whether minisatellites have evolved as a consequence of the local action of recombination in the subtelomeric regions, or if they themselves are hotspots for meiotic recombination and have evolved because of this (Jarman & Wells, 1989).

Evidence of de novo *crossing over at minisatellites*

Evidence of recombination has been found at a number of different minisatellites to support the theory which links minisatellites, recombination and chromosome pairing at meiosis. Gene-conversion events which are typical of germline mutation at minisatellites are themselves recombinational and are thought to occur by abortion of inter-allelic crossing over. However, very little evidence of true crossover events has been observed within minisatellite repeat arrays. Although one example has been isolated following pedigrees analysis at MS31a. This length-

change mutant showed evidence of haplotype switching in the DNA closely flanking the repeat array, this was extensively investigated to reveal a true crossover molecule consisting of 10 kb of one allele fused to approximately 2 kb of the other (Jeffreys *et al.*, 1998a). Recently, evidence from MS32 has also shown that recombination within the minisatellite array occurs infrequently compared to gene-conversion, to give rise to equal and unequal crossovers at similar frequencies (Jeffreys *et al.*, 1998b). Crossing-over showed the same polarity as gene-conversion and it was suggested that these processes arise by a common mechanism, implying that minisatellite instability is a by-product of meiotic recombination (Jeffreys *et al.*, 1998b). In addition, several mutants isolated by SP-PCR at CEB1 consist of the beginning of one allele fused to the end of the other, which may also represent true crossover events (Buard *et al.*, 1998).

It must be remembered that analysis of minisatellite mutation has previously been performed on solely on sperm mutants detected by large changes in repeat array length. However, if recombination is truly the main driving force behind mutation at minisatellites then some of these events would be expected to result in isometric (same size), or nearly isometric mutants; which would, by definition, not be identified by conventional methods of minisatellite mutation detection. To overcome this problem, a recombinant detection system has been designed and tested at the human minisatellite, MS32 (Jeffreys *et al.*, 1998b). Studies of repeat turnover and flanking marker exchange in recombinants isolated using this system have defined a crossover hotspot in the flanking DNA immediately adjacent to the 5′ end of the minisatellite array, which showed high levels of population variability and gene-conversion activity in previous studies.

The present work reproduces the analysis described for MS32 to isolate and characterise recombinant molecules from MS31a. This will determine whether the recombination hotspot identified at MS32 is a common feature of minisatellites, and whether this is the driving force behind the instability of the tandem repeat array. Recombinant molecules are specifically amplified from a pool of sperm DNA using allele-specific primers. To do this, the sperm donor chosen had to be heterozygous at as many sites in the flanking DNA of the minisatellite as possible, as well as being heterozygous for minisatellite array length. MS31a is a prime target for this work because pedigree analysis has already identified one true crossover event in the flanking DNA of this minisatellite (Jeffreys *et al.*, 1998a). Furthermore, population studies carried out on variant sites in the flanking DNA of MS31a have provided substantial evidence of recombination around this minisatellite (Chapter 4). In addition MS31a has an informative MVR-PCR system which has shown that germline mutation generally involves relatively simple exchange of large blocks of repeats (Neil, 1994). In contrast, mutation at MS32 involves the exchange of small blocks of repeats which can be quite complicated. Mutation becomes progressively more complicated and difficult to interpret at MS205, CEB1 and B6.7, due to difficulties in identifying the origin of repeats during mutation and the highly complicated nature of mutation processes at these minisatellites (May *et al.*, 1996; Buard *et al.*, 1998; Tamaki *et al.*, 1999). Work in the previous chapter has also identified a number of flanking polymorphic sites

that are essential for the isolation and characterisation of recombinant alleles; and identified a number of possible sperm donors which could be useful for this analysis.

Results

The sperm donor, LRIs137 was chosen for this analysis because of the substantial difference in the size of the MS31a alleles in this individual, and because he is heterozygous at five sites in the 5′ flanking DNA and two sites in the 3′ flanking DNA (Figure 5.4).

Mutation rate estimation in sperm donor LRIs137

To determine whether LRIs137 was suitable for this study, it was necessary to obtain an accurate estimate of mutation rate for this individual and confirm that it is within the expected range for MS31a. This can be done with reasonable accuracy by PCR, but this first requires that the amplification efficiency of the PCR reaction must be ascertained. In this case, amplification efficiency is the probability that amplification of a single molecule will be successfully initiated in each reaction. To do this, the concentration of the sperm DNA from this individual was accurately measured using a fluorimeter (Materials and Methods). Multiple aliquots of DNA, diluted to the equivalent of two diploid genomes (12 pg) per reaction, were amplified across the minisatellite using primers 31C and 31N. Out of a total of twenty reactions, fourteen were shown to be positive for the upper allele and thirteen were positive for the lower allele (data not shown). Assuming a Poisson distribution, the amplification efficiency was calculated to be 56% per molecule.

Mutation rate was accurately calculated by SP-PCR analysis, taking into account the amplification efficiency, as determined above. 10 x 80 molecules, 10 x 160 molecules and 9 x 320 molecules of LRIs137 sperm DNA was amplified using primers 31C and 31N (Figure 5.1). 41 mutants were seen in a total of 5,280 molecules, allowing the mutation rate to be calculated at 0.8% per haploid genome. The use of primers 31C and 31N, located close to the array, was essential to allow small length-change mutant molecules and progenitor alleles to be easily distinguished by size following electrophoresis and Southern hybridisation. However, mutants may still be hidden by the progenitor signal, and also by the background signal generated using a high concentration of input DNA (320 molecules, Figure 5.1). This figure is therefore likely to be an under-estimation of mutation rate. Nevertheless, this estimate is similar to the average mutation rate in the male germline of 1.2% per gamete as estimated by pedigree analysis (Neil, 1994). It is also apparent that using an input DNA of over 160 molecules increases the background and mutants can be difficult to distinguish, so reducing the amount of input DNA would be an improvement for subsequent analysis.

SP-PCR analysis of length-change mutants from LRIs137

This study was continued to examine the structure of mutants in the sperm DNA of LRIs137 to determine whether they are similar to those previously observed at this minisatellite.

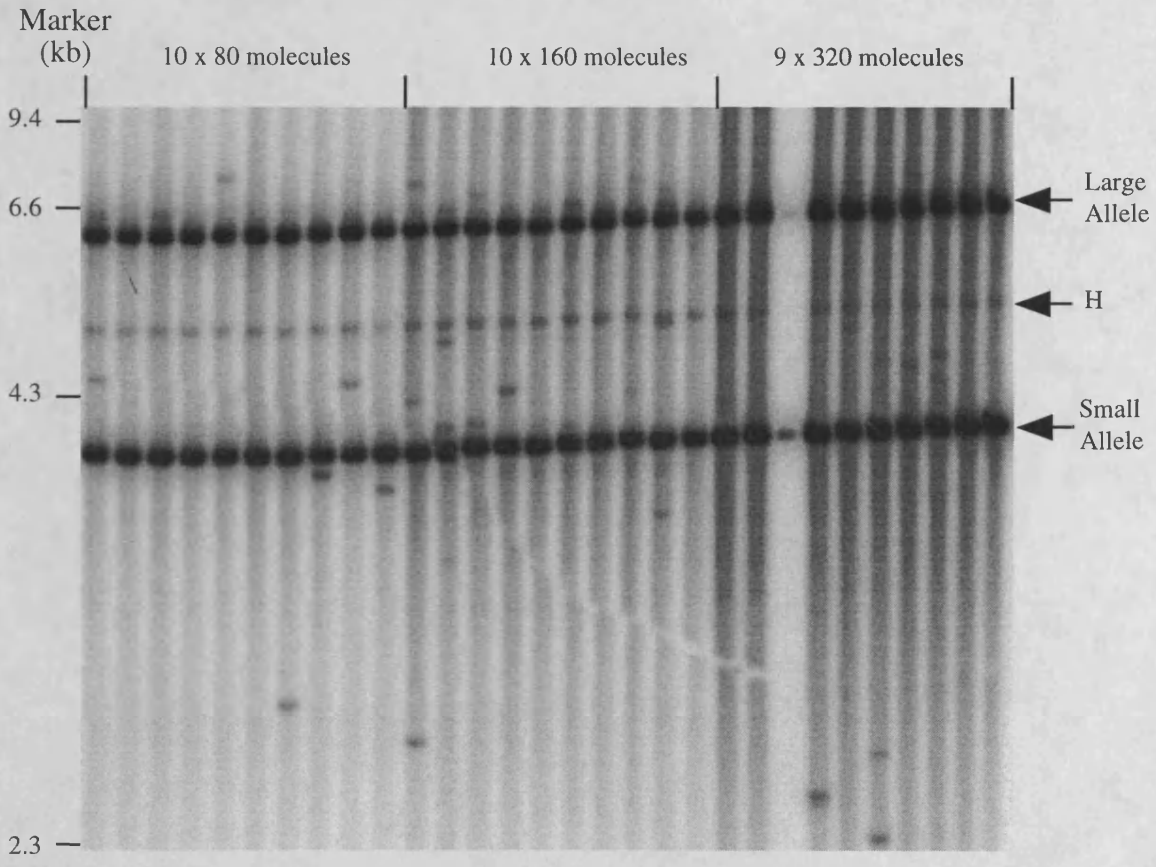


Figure 5.1 Mutation rate estimation in sperm donor LRIs137 by SP-PCR.

SP-PCR, using primers 31C and 31N was carried out on 10 x 80 molecules, 10 x 160 molecules and 9 x 320 molecules of LRIs137 genomic DNA as estimated by Poisson analysis and fluorimetry. The size and location of the small and large alleles are indicated with arrows; a band of heteroduplex DNA (H) is also shown. A total of 31 gain and 10 loss mutants were identified and assigned to the allele closer in size.

Length-change mutants were directly isolated from sperm DNA following SP-PCR. Thirty reactions containing 100 molecules each of LRIs137 sperm DNA were amplified using primers 31C and 31N in a 7 µl reaction. 1 µl of product from the mutant positive PCRs was reamplified for 6 cycles in a 20 µl reaction, and length-change mutants identified following gel electrophoresis and Southern hybridisation (Figure 5.2A). The remaining 6 µl from the first round of amplification was separated by electrophoresis under identical conditions. The smaller allele was visible following ethidium bromide staining and the position of mutant bands in the gel could be estimated by comparison with the autoradiograph. Three fractions around the estimated size of the mutants were excised from the gel, and the DNA recovered by electroelution into dialysis membrane (Materials and Methods), followed by ethanol precipitation. The DNA pellet was redissolved in 15 µl water. Nested PCR using primers 31A and 31M showed a large degree of contamination with the small progenitor allele of this individual. To reduce this contamination, the fractions containing the highest ratio of mutant to progenitor were taken and the excision and recovery of mutant bands was repeated twice. These mutants were then analysed by three-state MVR-PCR (Materials and Methods).

The structure of six length-change mutants from around the size of the smaller progenitor allele were analysed by MVR-mapping (Figure 5.2B). Unfortunately, because of the run of eleven identical repeats (underlined) at the beginning of both arrays it was difficult to determine whether some of these events were intra- or inter-allelic in origin. This is exacerbated because of the polarity of mutation so that almost all of these mutation events occur adjacent to or within this region. All mutation events show strong polarity to within the first 24 repeats (480 bp) of two arrays which are 113 and 228 repeats long and one mutant, 18b shows perfect alignment between the two alleles with respect to the position of the donated segment of repeats. Mutant 9 involves deletion of 13 repeats from the beginning of the array. The ambiguous origin of the ten repeats immediately upstream from the deleted region means this could be either an inter-allelic gene conversion followed by a deletion in the recipient, or a simple intra-allelic deletion from the small allele. This is the first case of paternal germline deletion observed at MS31a. Four mutants appear to have arisen by interallelic gene-conversion. Two of these mutants (1b and 17) are simple and the upstream boundary of conversion was not mapped within the repeat array. These mutants may therefore have arisen in three ways; by simple conversion directly into the start of the array; by co-conversion of part of the array plus some flanking DNA; or by unequal exchange. Examination of upstream markers would be necessary to confirm which explanation is appropriate. The three remaining gene conversion events (18a, 18b and 1a) were complex, involving duplication (arrows) of the insertion site in the recipient (green) allele. However these duplications are also found within the donated (red) segment of repeats, and may therefore be derived from either allele. Fortunately the surrounding repeats allow the allele of origin to be determined, so the triplications (18a and b) are actually duplications and the duplication (1a) is a simple gene-conversion event. The extent of conversion cannot be determined in 18a and requires characterisation of upstream markers to confirm whether co-conversion of the flanking DNA has occurred. The complexity of the mutant structure indicates that it has not arisen by unequal crossing over.

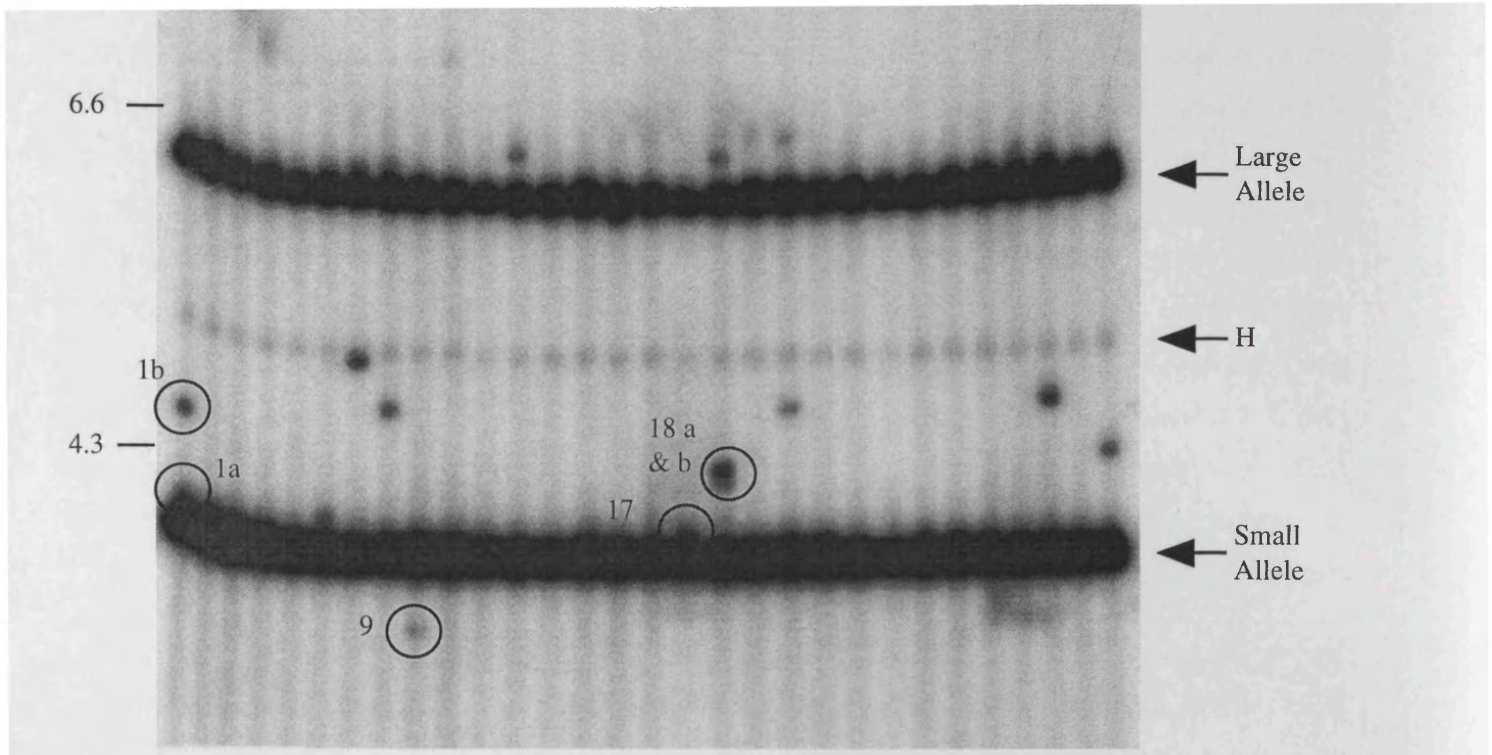


Figure 5.2. LRIs137 sperm mutants isolated by SP-PCR.

SP-PCR was carried out on 30 aliquots each containing 100 amplifiable molecules of sperm DNA from LRIs137 using primers 31C to 31N. Mutants were isolated by three rounds of excision and reamplification using nested primers 31A and 31M and analysed by three state MVR-PCR.

A. SP-PCR using primers 31C to 31N. Size and location of the small and large progenitor alleles are shown. Mutants isolated for analysis are circled. A large amount of heteroduplex DNA (H) was also present, as shown.

B. Three-state MVR-PCR of the isolated mutant alleles. Repeats are coded E, y or e according to the three-state system described in Materials and Methods. Repeats from the large allele are shown in red, repeats from the small allele are shown in green. Deleted regions are indicated by "-----"; duplicated regions in the recipient are highlighted in bold and shown by arrows above the sequence; grey arrows indicate additional putative regions of duplication in the donated segment; and "....." indicates allele continues beyond mapped region. Underlined repeats are identical in both progenitor alleles so the origin of these repeats in the mutant is difficult to determine. The repeat copy number and mutation event are detailed for each mutant isolate. Flanking haplotypes of these mutants were not determined.

B.


Large progenitor allele
Small progenitor allele
Mutants
9
-13 rpt internal deletion
18b
interallelic conversion +11 rpts + 9 rpt duplication
18a
interallelic conversion 11rpts + 4 rpt duplication
1a
interallelic conversion 5 rpts
1b
interallelic conversion 23 rpts or unequal exchange
17
interallelic conversion 5 rpts or unequal exchange

Admittedly this analysis is not comprehensive but it does show that SP-PCR can be successfully used to isolate length-change mutants from MS31a. This system is able to detect allele length-changes of five repeat units (see Figure 5.2a), which is probably the limit of resolution of SP-PCR at this locus. Unfortunately the possibility that the simple mutants 9, 1b and 17 are PCR artefacts could not be ruled out. Artefacts can arise by incompletely extended PCR products reannealing at random within the repeat array. These can be extended in the next round of amplification to give products of abnormal length. Artefacts can be distinguished from authentic mutants in several ways. This is generally done by comparing the number of abnormal length bands arising during SP-PCR in both blood and sperm DNA. The mutation rate in blood is much lower at MS31a, possibly as low as 0.008% (Neil, 1994), so if equivalent numbers of abnormal length products were observed in sperm and blood from the same individual the majority of these bands would be artefacts. Unfortunately, blood DNA from this individual is not available because he is a sperm donor. However, other evidence suggests that the majority of these signals are derived from authentic mutants. Firstly, the signal intensities of all bands are equivalent (Figure 5.2a) and are similar to the signal intensity obtained following amplification of a single progenitor molecule under the same conditions. In addition, prolonged autoradiography shows no evidence of additional bands which would be indicative of PCR artefacts arising late in the amplification program. The first SP-PCR experiment (Figure 5.1) showed that dose response of abnormal length bands to input DNA is consistent with the authenticity of these mutants, and the mutation rate estimated from this screen (0.7% per sperm) was in accordance with that estimated from the second (0.8% per sperm), which were similar to mutation estimates (1.4% per sperm) obtained by pedigree analysis (Neil, 1994; Jeffreys *et al.*, 1994). Also the structure of mutant alleles isolated by SP-PCR correspond well with those previously analysed from pedigrees (Neil, 1994; Jeffreys *et al.* 1994) and are difficult to reconcile with artefactual processes.

This work has demonstrated that both the mutation processes and mutation rate of this individual are within the expected range for this minisatellite. Structural analysis of mutants has presented definitive evidence of recombination, in the form of gene-conversion within the repeat array, and possible evidence of crossovers at this locus in the sperm DNA of this individual. This work can now be extended to isolate mutant molecules on the basis of mutant structure and not length differences which, as shown above, can be quite laborious.

Size enrichment of LRIs137 sperm DNA

Following the preliminary analysis of mutation at MS31a, isometric and nearly isometric molecules exhibiting exchange of flanking markers were isolated and characterised. This was achieved using allele-specific PCR on fractions of DNA containing only the small progenitor allele of LRIs137. The two progenitor alleles were separated by size, a process called size-enrichment or size fractionation. This minimises the risk of jumping PCR between partially amplified progenitor alleles in the subsequent analysis which would interfere with recombinant detection. Removing the large progenitor allele from this system means that the majority of the

targets of the upstream allele specific primer are removed. The remaining molecules that will be amplified using both sets of allele specific primers will therefore represent recombinant molecules present in the initial DNA sample; these can also be authenticated by size.

Size enrichment was carried out under PCR clean conditions. All reagents, e.g. restriction endonucleases, digestion buffers, PCR buffer, DNA polymerases were kept separate from general stocks and maintained in a clean environment. DNA manipulations, whenever possible were carried out under a laminar flow hood and, where appropriate all equipment (e.g. electrophoresis tanks, gel trays and other items) was soaked overnight in ~1 M hydrochloric acid to degrade any potential contaminant DNA. All electrophoretic steps were carried out in the absence of ethidium bromide and UV light to minimise the DNA damaging effects of these agents. The loading dye used for size enrichment contained 3 g glycerol, 0.001 g bromophenol blue and 100 µl 10x TBE in 10 ml. This dense buffer reduced the risk of DNA escaping the wells and migrating through the buffer which could complicate fractionation. A large stock of 11x PCR buffer (Materials and Methods) was prepared and used solely for the subsequent analyses.

Approximately 40 µg sperm DNA from donor LRIs137 was digested with 90 units *Dra* I in 10x REact buffer 1 (Gibco BRL) for 2.5 hr. This enzyme excises the small allele as a fragment of 8.4 kb and the large allele as a fragment of 10.5 kb with 3.3 kb flanking DNA 3′ and 2.7 kb 5′ of the minisatellite. After 10 min, an aliquot containing ~2 µg DNA was removed and over-digested for 2 hr in fresh digestion mix. Following incubation, an aliquot containing equivalent amounts of DNA from both reactions were examined by electrophoresis to ensure that the DNA had been completely digested. The concentration of the digested DNA was measured using a fluorimeter. 10 µg was separated by electrophoresis for 32 hours at 120 volts on a 40 cm long 0.6% LE (Sigma Biochemical Co.) agarose gel, with 400 ng each of λ DNA x *Hin*d III and ϕx174RF DNA x *Hae* III markers. The position of marker bands and the lanes containing the digested DNA were determined by ethidium bromide staining of the end of the gel and marker lanes for 20 min. The position of the digested DNA corresponding to fragments between approximately 11.9 kb and 7.8 kb was estimated and divided into 20 gel slices or fractions. A wide range of size fractions were taken because the digested fragments generally run slower than predicted. These fractions contained the small and the large allele, and any mutants within about 2 kb each side of both alleles. The gel slices were isolated and the DNA electroeluted into dialysis membrane, ethanol precipitated and redissolved in 10 µl 5 mM Tris-HCl pH 7.5. A duplicate fractionation of the same DNA was carried out in parallel, under identical conditions.

The size range of the DNA in each fraction was estimated by electrophoresis of a 10th volume of the fractionated DNA which was visualised by Southern hybridisation with ^{32}P-labelled total genomic DNA (Figure 5.3). The degree of enrichment of each fraction was measured by amplification of 10th volume fractionated DNA and a dilution series of unfractionated, digested starting DNA using primers 31DD to 31V (Figure 5.5; see Chapter 6 for calculations). In fractions 13 - 20 (i.e. those containing only the small progenitor allele), a total of 160,000 amplifiable molecules of the small allele were recovered, and almost all (> 99.99%) large allele

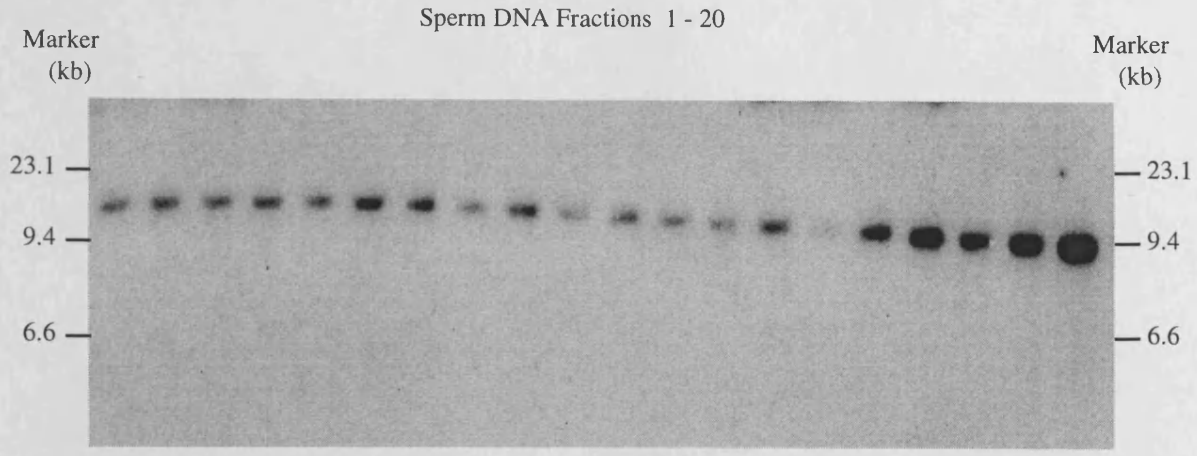


Figure 5.3. DNA from fractions recovered following size enrichment of LRIs137 sperm DNA.

Sperm DNA was digested with *Dra* I, electrophoresed through an agarose gel and 20 size fractions ranging from between approximately 8.0 and 17.2 kb were collected by electroelution. Aliquots of the recovered DNA were electrophoresed and detected by Southern hybridisation with total genomic DNA. The size range of each fraction was estimated by comparison with marker bands. Note the slightly erratic recovery.

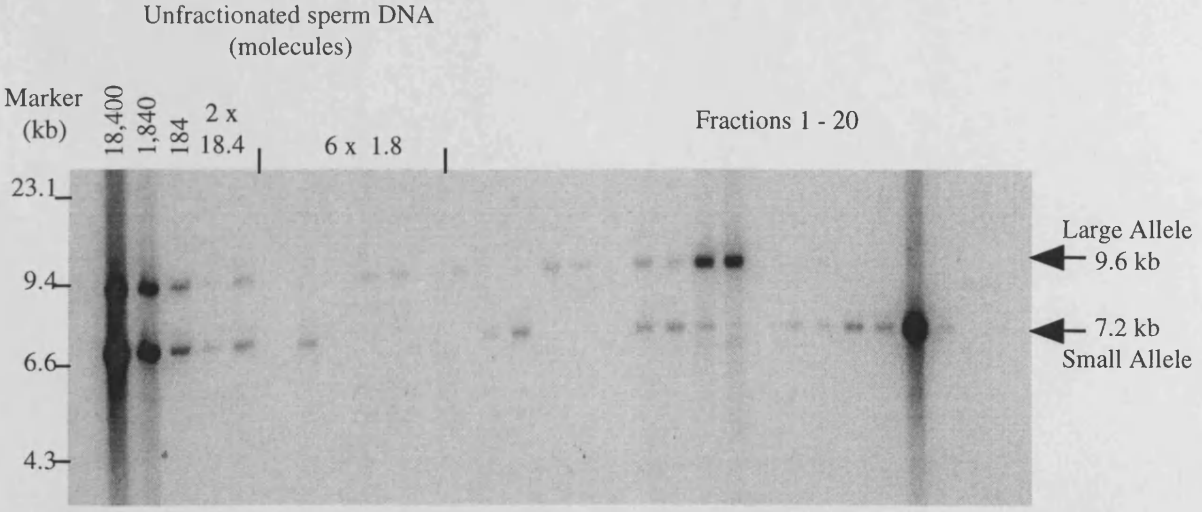


Figure 5.4. Amplification of MS31a in fractions recovered following size enrichment of LRIs137 sperm DNA.

10x diluted aliquots of each size fraction were amplified using primers 31DD and 31V. The degree of enrichment of each allele was estimated by the comparison of the hybridisation signal of each fragment with that of a dilution series of unfractionated digested genomic DNA. Size and location of the large and small alleles is indicated.

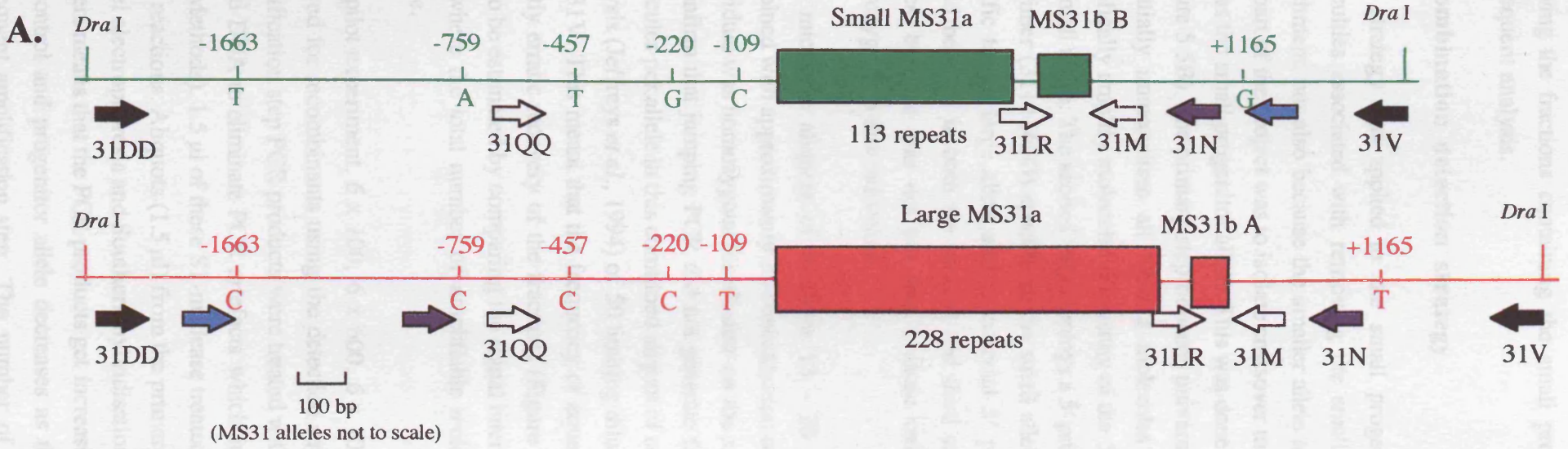
Figure 5.5. Crossover detection strategy.

A. Schematic diagram of the two MS31a alleles and the flanking DNA of LRIs137.

The small allele and flanking DNA is represented in green and the large allele in red. Large boxes represent MS31a alleles with the size of each allele underneath, smaller boxes represent MS31b alleles. Flanking polymorphic sites and the allelic state of each site are indicated. The *Dra* I sites used for size enrichment are indicated and primers are shown as coloured arrows corresponding to the steps in B.

B. The PCR strategy used for detecting crossovers.

The primary (black) step amplifies and "immortalises" all MS31a molecules present, and the subsequent allele-specific (blue and purple) steps select for crossover molecules consisting of the 5′ end of the large (red) allele fused to the 3′ end of the small (green) allele. The colour of each step corresponds to colour of the primers in A. PCR products derived from each step were treated with S1 nuclease to specifically degrade single-strand products. Primers 31QQ and 31M re-amplify selected alleles in order to map the crossover points by RFLP analysis of the 5′ sites (+457, +220, +109), by determining the length of MS31b (31LR and 31M), and by MVR mapping the repeat array.



B.

1° PCR	31DD → 31V	
	S1 Nuclease Treatment	
2° PCR	31- 1663C →	31+1165G
	S1 Nuclease Treatment	
3° PCR	31- 759C →	31N
Final PCR	31QQ →	31M
	Mapping of breakpoints	

molecules were eliminated. The two alleles have now been (almost) completely separated, allowing the fractions containing the small progenitor allele to be pooled and used in the subsequent analysis.

Recombination detection strategy

This strategy was applied to the small progenitor allele fractions mainly because of the difficulties associated with removing the small allele from the large allele fraction by size enrichment, but also because the smaller allele is more amenable to PCR analysis. The aim of this part of the project was to isolate crossover molecules that were the same, or nearly the same size as the small progenitor allele. This was done using a three-step nested amplification process (Figure 5.5B). The primary step between universal (non-allele specific) primers 31DD and 31V essentially immortalises all MS31a molecules with equal efficiency. The next two steps specifically amplify molecules consisting of the 5′ end of the large allele fused to the 3′ end of the small allele. The second step employs a 5′ primer (31-1663C) specific to the large allele and 3′ primer (31+1165G) specific to the small allele. The final step uses a 5′ primer (31-759C) specific to the large allele and a universal 3′ primer (31-N). Ideally, allele-specific primers would be used in both the second and third steps to improve the selection for recombinant alleles, but this was not possible because only one marker in the 3′ flanking DNA was heterozygous in this individual.

Nine microlitre aliquots of fractions 13 - 20 containing the small allele were pooled and combined with approximately 2x concentration of digested DNA from a control individual. This individual was homozygous for all sites on the small test allele and acted as an internal control to confirm that jumping PCR did not generate false recombinants. The number of amplifiable molecules per allele in this combined aliquot of control and test alleles was estimated by Poisson analysis (Jeffreys *et al.*, 1994) of 50 limiting dilutions amplified using universal primers, 31DD and 31V. This meant that the frequency of equal mutation could be determined precisely. The slightly erratic recovery of the fractions (Figure 5.3) meant that the rate of unequal exchange had to be estimated by comparing the signal intensity of fractions 13-15 and 17-20 with fraction 16, where the total number of amplifiable molecules of the small allele has been calculated above.

In a pilot experiment, 6 x 100, 6 x 600, 6 x 1000 and 6 x 3000 amplifiable molecules were assayed for recombinants using the detection strategy outlined in Figure 5.5. Following each amplification step PCR products were treated with S1 nuclease which specifically digests single strand DNA to eliminate PCR artefacts which interfere with recombinant detection (Materials and Methods). 1.5 µl of these S1-nuclease treated PCR products were used to seed subsequent PCR reactions. Aliquots (1.5 µl) from the primary, secondary and tertiary PCRs were examined by gel electrophoresis and Southern hybridisation (Figure 5.6). The use of nested allele-specific primers means that the PCR products get increasingly shorter, and the background signal from the control and progenitor allele decreases as the selection against them increases with each subsequent amplification step. The number of recombinants also increases with increasing

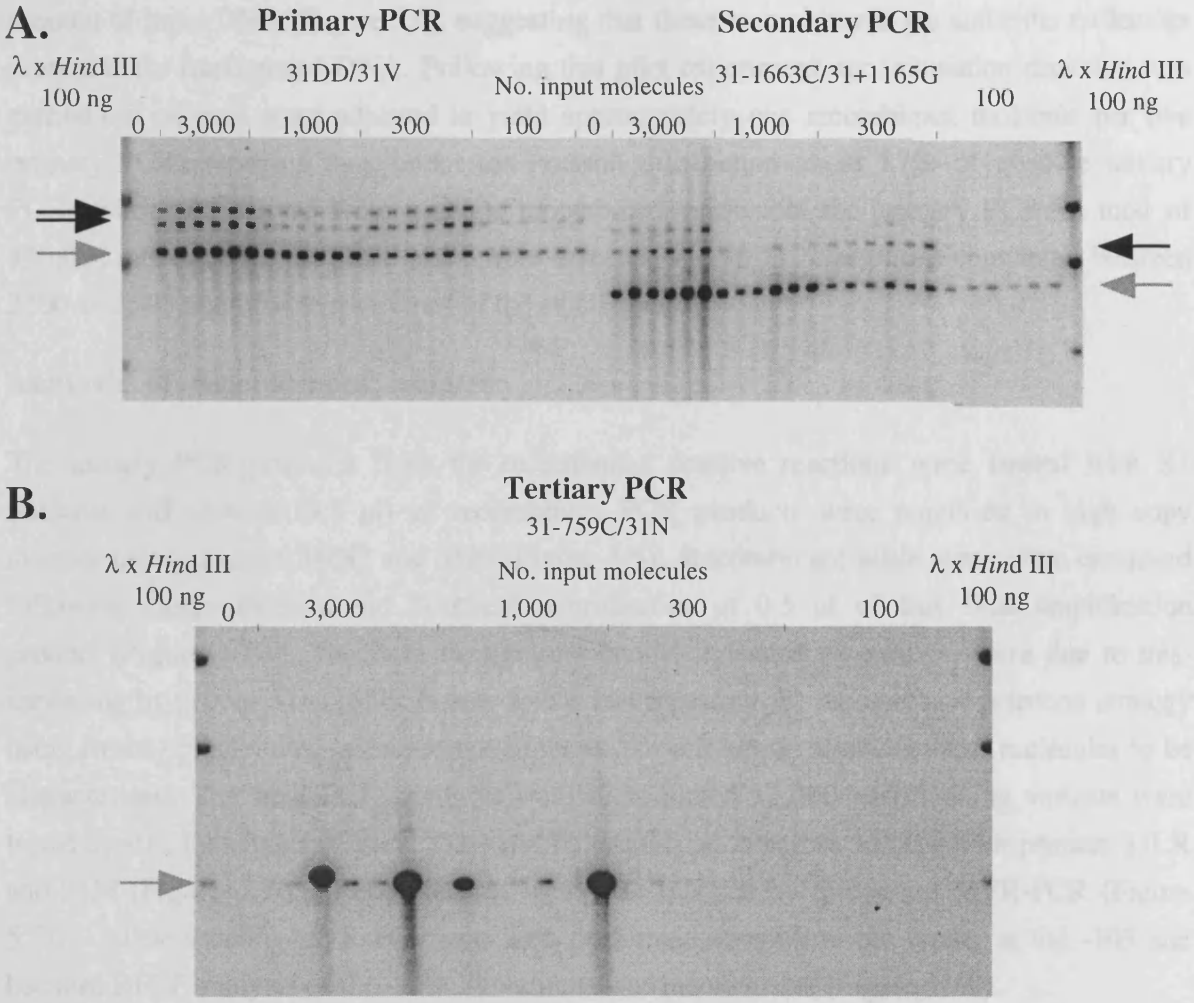


Figure 5.6. Pilot screen to isolate crossover molecules from sperm DNA.

6 x 100, 6 x 600, 6 x 1,000 and 6 x 3,000 amplifiable molecules of the small allele (MS31a array 113 repeats long) were assayed for recombinants. These were mixed with an equivalent concentration of control DNA (MS31a array sizes 218 and 138 repeats long) from an individual homozygous for sites from the small allele. This provides an internal control to ensure jumping PCR does not occur. 1.5 µl from each 7 µl PCR reaction was analysed by gel electrophoresis and Southern hybridisation.

A. PCR products obtained from the primary and secondary steps of this crossover detection strategy (Figure 5.5B). The first step was the "immortalisation" step between universal primers 31DD and 31V, the second step was the first allele specific PCR step designed to select for crossover molecules. The small and large alleles of the control individual are indicated by black arrows, the position of the small test allele is indicated by grey arrows. The PCR products from the second step are noticeably shorter than those from the primary step, the test allele was preferentially amplified over the control alleles in the secondary step because it is shorter.

B. PCR products obtained from the third step of the crossover detection strategy (Figure 5.5B). The four highly intense bands were identified as recombinants, the remainder were background products. All recombinants were isometric apart from the first, which was a gain mutant. More recombinants were observed with higher concentrations of input DNA, which was consistent with the authenticity of these crossover molecules.

amount of input DNA (Figure 5.6), suggesting that these recombinants are authentic molecules present in the fractionated DNA. Following this pilot experiment, recombination detection was carried out on pool sizes adjusted to yield approximately one recombinant molecule per five primary PCRs ensuring that, under the Poisson distribution about 87% of positive tertiary PCRs would be derived from a single recombinant present in the primary PCR. A total of 146,000 molecules of the small allele were screened in 126 PCR reactions containing between 3100 and 603 amplifiable molecules of the small allele.

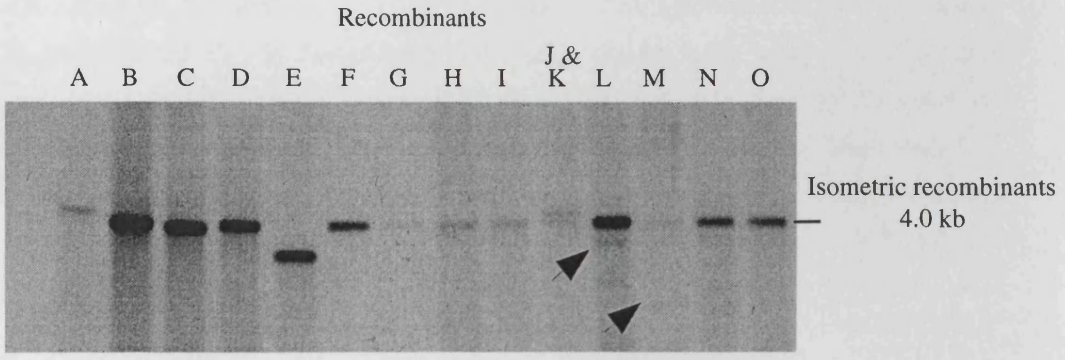
Analysis of recombinant isolates

The tertiary PCR products from the recombinant positive reactions were treated with S1 nuclease and aliquots (0.5 μl) of recombinant PCR products were amplified to high copy number using primers 31QQ and 31M (Figure 5.5). Recombinant allele sizes were estimated following electrophoresis and Southern hybridisation of 0.5 μl of this final amplification product (Figure 5.7A). The faint background bands, indicated by arrows, were due to mis-annealing by primer 31+1165G. It was found that repeating the recombinant detection strategy using freshly precipitated primer removed these spurious bands, allowing these molecules to be characterised. The final PCR products were then diluted x1,000 and flanking variants were typed by RFLP analysis (Figure 5.7B) and by amplification across MS31b with primers 31LR and 31M (Figure 5.7C). Recombinants were also mapped by three-state MVR-PCR (Figure 5.7D). Allele-specific MVR-PCR was also performed to confirm the typing at the -109 site because RFLP analysis of the -109C/T variants was inconclusive (Figure 5.7E).

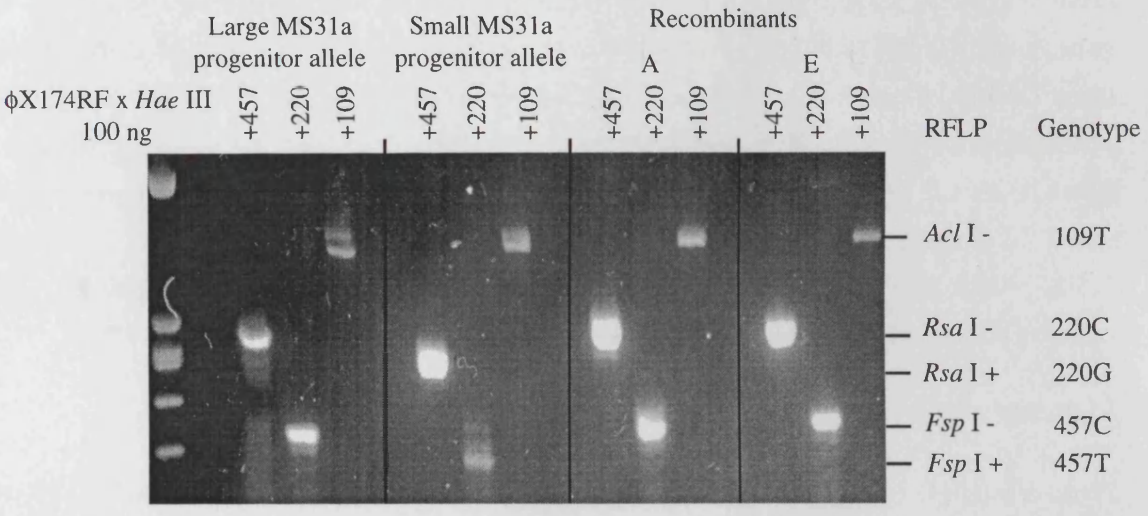
Structure of crossover mutants

From a total of 146,000 molecules of the smaller allele from sperm donor LRIs137, fifteen recombinants were isolated and characterised (Figure 5.8). No false recombinants arising by jumping PCR between the test and control alleles were observed. This established that most or all of these recombinants were authentic. All the exchange events localised outside the minisatellite array were isometric (B-D, F-I, K-O), whereas all the exchange events characterised within the array involved a change in length of the minisatellite (A and E). This work identified two interallelic unequal crossover mutants with recombinant repeat arrays (A and E); eleven isometric recombinants with the smaller allele unchanged in length but linked to distal markers from the larger allele (B and C, F-I, K-O); one complex equal crossover mutant (D); and one which could not be fully characterised (J). Recombinant J was amplified from the same primary PCR as recombinant K (Figure 5.7A). Allele-specific MVR-mapping using primer 31-109C was successful for a short distance into the minisatellite array of recombinant K (Figure 5.7E), but primer 31-109T was unable to successfully map recombinant J. However, because it is a gain mutant, it is likely to have arisen by interallelic unequal crossover within the minisatellite array. Isometric recombinants G and H consist of the flanking markers of the large (red) allele linked to the minisatellite array of the small (green) allele. The breakpoints of both molecules probably lie between site -109 and the minisatellite array, this could not be confirmed for recombinant G because the first two repeats could not be mapped.

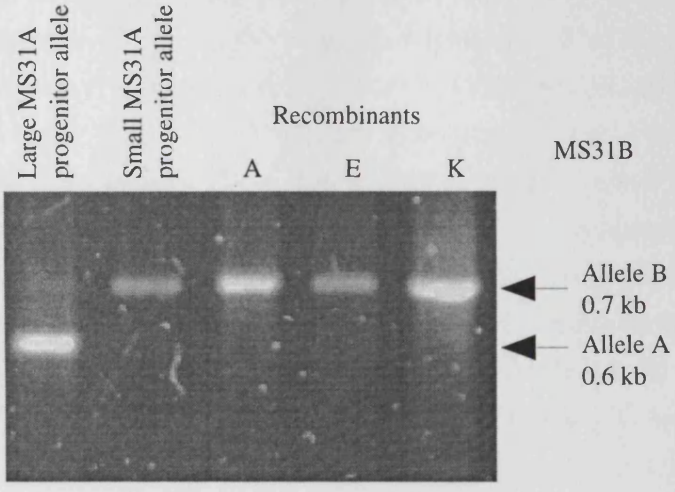
A.


Recombinants
A B C D E F G H I J & K L M N O
Isometric recombinants
4.0 kb

B.


Large MS31a progenitor allele
Small MS31a progenitor allele
Recombinants
A
E
ϕX174RF x Hae III 100 ng
+457 +220 +109 +457 +220 +109 +457 +220 +109 +457 +220 +109
RFLP
Genotype
Acl I - 109T
Rsa I - 220C
Rsa I + 220G
Fsp I - 457C
Fsp I + 457T

C.


Large MS31A progenitor allele
Small MS31A progenitor allele
Recombinants
A E K
MS31B
Allele B
0.7 kb
Allele A
0.6 kb

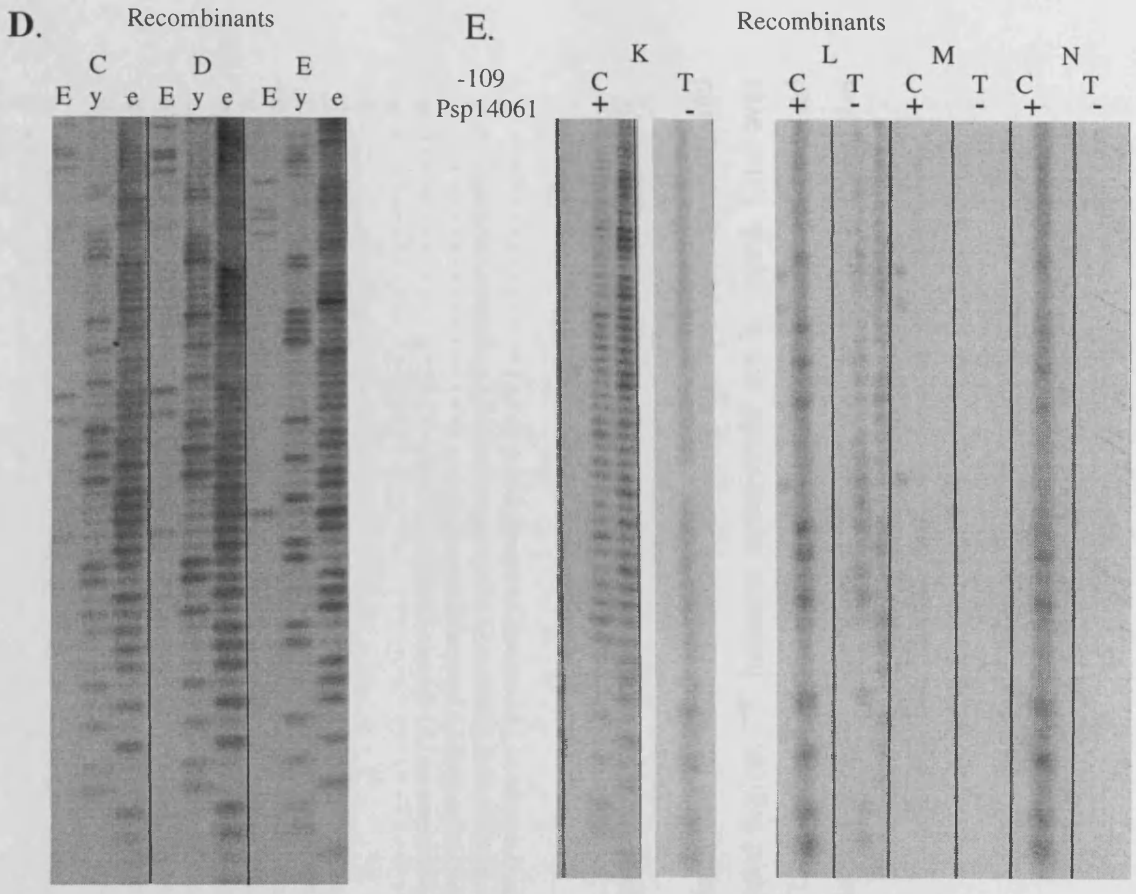


Figure 5.7. Characterisation of MS31a crossover molecules.

Recombinant molecules isolated were characterised by size and the locations of the crossovers were mapped by RFLP analysis of the 5′ sites (+457, +220, +109), by determining the length of MS31b and by MVR analysis of the repeat array.

A. Determination of the size of the recombinants isolated from sperm DNA. Isolates were reamplified between primers 31QQ and 31M and an aliquot (0.5 µl) used to determine the size of the recombinants by gel electrophoresis and Southern hybridisation. Three non-isometric (two gain, one loss) and twelve isometric recombinant molecules were isolated. The position and size of the small progenitor allele is indicated. Arrows indicate spurious bands arising during PCR.

B. Locating crossover breakpoints in the 5′ flanking DNA of MS31a by RFLP analysis. Markers -457T/C, *Fsp* I+/-; -220G/C, *Rsa* I+/-; and -109C/T, *Acl* I+/- were genotyped in the two progenitor alleles and in two recombinants, A and E. Sets of digests are separated by a black line and the position of the cut and uncut fragment is indicated for each RFLP. *Acl* I failed to cut in all samples but the remaining sites can be typed, for the large progenitor allele: 457C, 220C; the small progenitor allele: 457T, 220G; recombinant A: 457C, 220C; recombinant E: 457C, 220C.

C. MS31b allele-length analysis in recombinants A, E and K by amplification between primers 31LR and 31M. MS31b from the large and the small MS31a progenitor allele were amplified as controls. Size and position of MS31b alleles A and B are indicated.

D. MVR-PCR mapping of recombinants. All recombinants were typed by three-state MVR-PCR using the universal primer 31A. The distinguishing sequence of the MVR primer used (AT, GC or GT) and the code they represent (**E**, **y** or **e**) are indicated.

Figure 5.8. Structure of recombinants isolated from the sperm DNA of LRIs137.

The minisatellite alleles and the informative flanking polymorphic positions are indicated for each crossover molecule and for the large (red) and small (green) progenitor alleles. "....." indicates allele continues beyond mapped region. "?" indicates repeat could not be typed. Crossover molecules are defined by size (pink), and as either isometric (I), gain (G) or loss (L) mutants. The repeat array of recombinant J could not be typed by MVR-PCR. The first two flanking sites -1663 and -759 are the same in all molecules because allele specific primers complementary to these sites were used to select for the recombinants.

Most of these events are simple, that is crossovers can be mapped to a single interval between two polymorphic sites. However, recombinant D appears to be much more complex, apparently crossing over in three intervals; between -457 and -220, -220 and -109, -109 and the minisatellite array. It is unlikely that multiple crossovers have occurred within such a short region of DNA so these are probably the result of a crossover with an adjacent conversion event (Borts & Haber, 1987). This could have occurred by formation of a Holliday junction between the alleles, probably in the interval between -457 and -220, which then underwent branch migration towards the minisatellite (Figure 5.9). Resolution of this complex somewhere between -109 and the third repeat of the array means that the migration region would have been in the heteroduplex state. If this was then repaired by a patchwork mismatch repair process, the pattern of crossover and adjacent conversion in recombinant D would be a conceivable product.

A crossover hotspot?

The frequency of crossovers within the flanking DNA and the MS31a repeat array was compared to the rate expected if crossovers were located randomly in the male genome at meiosis (Figure 5.10). This is done by calculating the crossover efficiency (C) of each interval using: C = OC / EC. OC is the Poisson corrected number of crossovers observed in each interval, EC is the expected number of crossovers per interval as calculated from the average frequency of crossovers in human autosomes at male meiosis, which is 8.9×10^{-3} per Mb (Weissenbach *et al.*, 1992; Gyapay *et al.*, 1994). From this figure, the number of crossovers expected in a screen of 146,000 molecules was 1.3×10^{-3} per base pair. Therefore the crossover efficiency in the interval between -759 and -457 was $4 / (1.3 \times 10^{-3} \times 302 \text{ bp})$, or 10.19 cM / Mb. The crossover point of unequal exchanges was taken as the mean of the breakpoints in the two alleles.

Crossovers are distributed non-randomly within this region ($\chi^2 = 24.16$, 5 d.f., $p = 0.002$). The majority (80%) of exchanges are located in the flanking DNA but these decline rapidly within the minisatellite array itself (Figure 5.10). There appears to be a peak of crossing over between -220 and -109, approximately 160 bp upstream of the minisatellite array, although this clustering in the flanking DNA is not significant and may be due to chance ($\chi^2 = 4.35$, 3 d.f, $p = 0.23$). This demonstrates that this is a crossover hotspot within the flanking DNA immediately adjacent to the most unstable end of the MS31a repeat array. However, the dimensions of this hotspot could not be defined using these data. It is possible that there may be a very short region of crossover activity, alternatively, the hotspot may have a wide base within the flanking DNA of MS31a. These two scenarios can only be distinguished by the characterisation of additional crossover events both within this region and further upstream of the tandem repeat array.

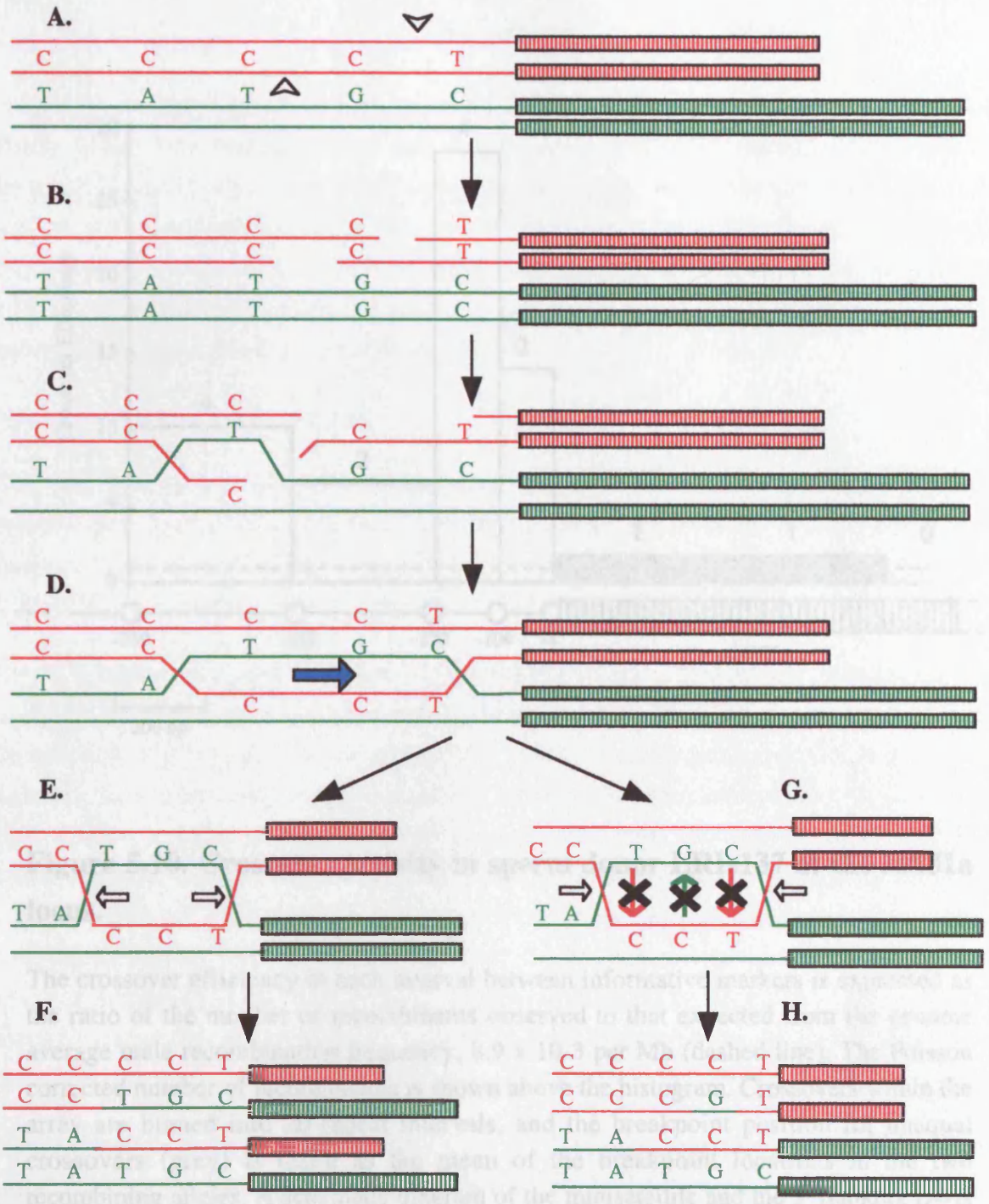


Figure 5.9. Formation of patchwork repair mutants.

A. Staggered single-strand nicks are introduced into the flanking DNA of the large (red) minisatellite allele. **B.** The break expands into a gap and the two alleles pair in register. **C.** The gap is bridged by strand-invasion from the small (green) allele forming a Holliday junction. **D.** Branch migration (blue arrow) occurs to expand the region of heteroduplex DNA. **E.** The strand invasion complex is isomerised and the Holliday junctions are resolved (white arrows). **F.** Repair synthesis yields two reciprocal equal exchange recombinant alleles. **G.** Alternatively, Holliday junctions are resolved (white arrows) and mismatch repair within the heteroduplex uses either allele at random as a template. **H.** This yields one patchwork repair mutant and one equal exchange recombinant .

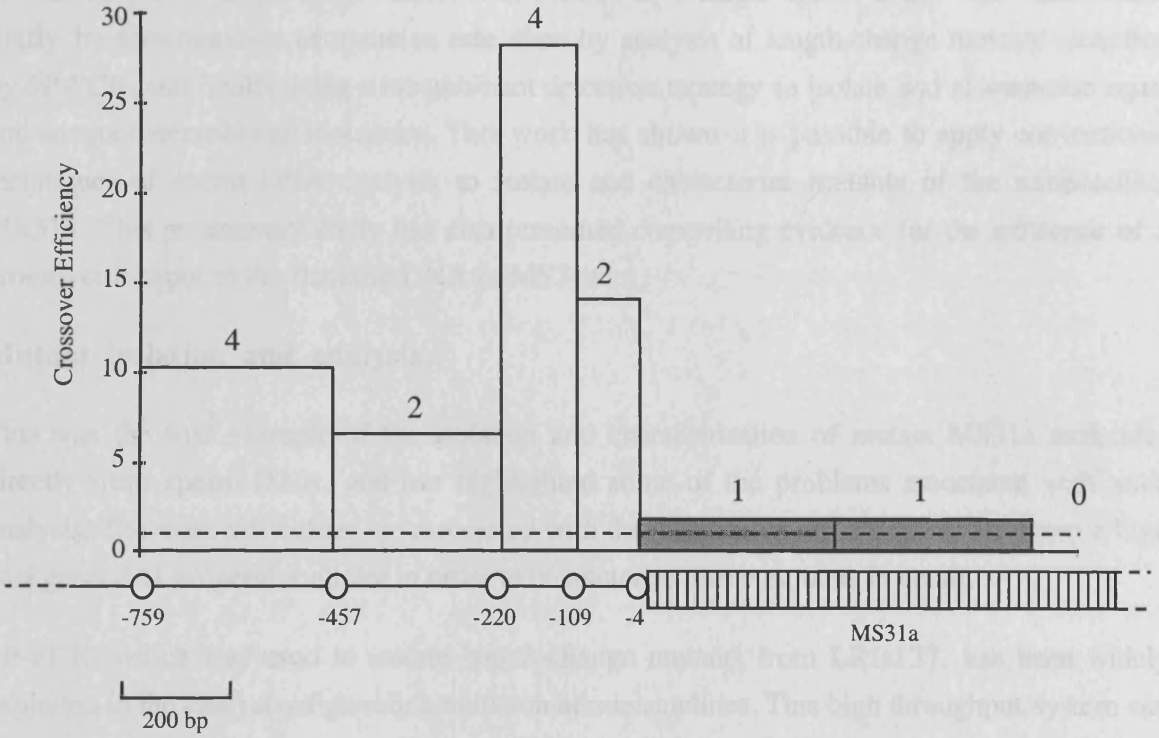


Figure 5.10. Crossover activity in sperm donor LRIs137 at the MS31a locus.

The crossover efficiency in each interval between informative markers is expressed as the ratio of the number of recombinants observed to that expected from the genome average male recombination frequency, 8.9 x 10-3 per Mb (dashed line). The Poisson corrected number of recombinants is shown above the histogram. Crossovers within the array are binned into 20-repeat intervals, and the breakpoint position for unequal crossovers (grey) is taken as the mean of the breakpoint locations in the two recombining alleles. A schematic diagram of the minisatellite and the 5′ flanking DNA is shown underneath.

Discussion

A detailed profile of germline mutation at MS31a in a single sperm donor was constructed; firstly, by determination of mutation rate, then by analysis of length-change mutants identified by SP-PCR, and finally using a recombinant detection strategy to isolate and characterise equal and unequal recombinant molecules. This work has shown it is possible to apply conventional techniques of sperm DNA analysis to isolate and characterise mutants of the minisatellite, MS31a. This preliminary study has also presented compelling evidence for the existence of a crossover hotspot in the flanking DNA of MS31a.

Mutant isolation and analysis

This was the first example of the isolation and characterisation of mutant MS31a molecules directly from sperm DNA, and has highlighted some of the problems associated with such analysis. The main difficulties are associated with the isolation of mutant molecules from a high background of progenitor alleles in order to characterise these mutants in detail.

SP-PCR, which was used to isolate length-change mutants from LRIs137, has been widely exploited in the analysis of germline mutation at minisatellites. This high throughput system can be used to screen large amounts of sperm DNA to identify minisatellite germline length-change mutants. It is a relatively simple technique, using PCR, gel electrophoresis and Southern hybridisation to detect mutants occurring at high frequencies. However, because it is PCR based, isometric mutants, large length change mutants and alleles close in size to the progenitor cannot be detected. This work has demonstrated that allele-length changes as small as 5 repeat units can be detected and characterised by SP-PCR. An additional drawback of this technique, as noted here, was that the use of primers immediately flanking the minisatellite means that few flanking variant sites are available to characterise these events in more detail. The ability to do this would have permitted the full classification of mutants 1b and 17 (Figure 5.2).

Conventional methods of mutant analysis, such as SP-PCR are only useful for the identification and isolation of length-change mutants, and are unable to identify crossover mutants similar in size to the progenitor. This can be overcome using pairs of allele-specific primers to specifically select for recombinant molecules from a pool of DNA enriched for one progenitor allele. This is an incredibly powerful system which allows single mutant molecules to be isolated from a background of up to 3100 progenitor alleles. The use of allele-specific PCR to isolate mutants has three implications. Firstly, that mutants are isolated from progenitor molecules on the basis of sequence. Secondly, multiple copies of the mutant allele are generated with each subsequent step. Third, the use of fixed primer sites irreversibly restricts the analysis to a particular region which, in this case was not sufficient to characterise the entire crossover hotspot. This system is also dependent on size-enrichment which allows the authenticity of mutants to be validated by size. However, size-enrichment is labour intensive and technically demanding, requiring the dedication of large amounts of sperm DNA, which is not feasible when using limited stocks of

sperm DNA. This system also requires a sperm donor who meets certain strict specifications, that is heterozygous for both minisatellite allele size and flanking markers, these are often difficult to meet and may not always be possible. However, this is also a requirement for detailed analysis of the mutant alleles and will therefore be impossible to overcome.

Male germline mutation at MS31a

Germline mutation at MS31a has been analysed by pedigree analysis and, as shown here, by direct analysis of sperm DNA by SP-PCR. Various evidence suggests that the majority of these mutants are genuine and not PCR artefacts. The rate of mutation to new length alleles in this individual was calculated at 0.8% per sperm, which is comparable with the estimate of 1.2% per gamete, obtained following pedigree analysis. This was not unexpected as variation in mutation rates between minisatellite alleles have been previously described at other loci (Jeffreys *et al.*, 1994; May *et al.*, 1996; Buard *et al.*, 1998). These studies also demonstrated that mutation is dominated by length gain mutants arising by gene conversion-like processes which copy blocks of repeats from one (donor) allele into the other (recipient) allele. These events are polarised towards the 5′ end of the minisatellite array, which was also found to be the most variable following population studies (Neil, 1994). This type of mutation appears to be representative of minisatellite mutation in sperm, suggesting that processes of male germline mutation to new length alleles is reasonably well conserved between minisatellites (Jeffreys *et al.*, 1997). There was also some evidence of unequal crossover occurring at this locus, although the possibility that these are artefacts cannot be ruled out.

Evidence of crossing over at MS31a in the male germline

The second half of this study has provided strong evidence of meiotic crossing over at the MS31a locus. The majority of crossovers occur within the flanking DNA to give rise to equal recombinants, whereas a smaller number of crossover events within the minisatellite array all generate unequal recombinants. These unequal exchange events are simple, involving none of the complex rearrangements, or insertion of anomalous repeats observed following SP-PCR analysis of this locus. The one complex event observed (recombinant D, Figure 5.8) can be explained by branch migration and subsequent resolution of the recombination complex.

True crossover events at the MS31a locus have a similar distribution to crossover events characterised at MS32 (Jeffreys *et al.*, 1998b). This preliminary work suggests that the crossover hotspot associated with MS31a may be centred about the same distance upstream of the minisatellite as the MS32-associated hotspot (160 bp at MS31a, 200 bp at MS32). However, the exact dimensions of the MS31a-associated hotspot are not known because of the limited region used in this study and the small number of breakpoints isolated. This small dataset may also explain the preponderance of unequal events within the minisatellite array of MS31a, whereas a more even balance between isometric and non-isometric events within the array was observed at MS32 (Jeffreys *et al.*, 1998b). The rate of crossing over at MS31a was lower in this individual than for all the alleles studied at MS32. At MS31a, the rate of equal

crossover events within the flanking DNA is 0.008% per small progenitor allele, which is three times lower than the average rate of equal crossover events within the flanking DNA of MS32. However, it is comparable to the rate of 0.009% per allele tested observed at the O1C allele, which is suppressed for mutation at MS32. These results could be indicative of a comparatively low frequency of crossing over associated with MS31a. Alternatively, this may be due to substantial variation in the levels of recombination between different MS31a alleles, which has also been observed at MS32 (Jeffreys *et al.*, 1998b). Preliminary work at MS32 suggests that differences between alleles can extend deeper than this (A. J. Jeffreys, pers. comm.). It has been observed that there is a breakdown of crossover symmetry between different alleles from the same individual, although crossover rates remain the same. It appears that the crossover hotspot becomes more diffuse as the length of the associated minisatellite allele increases. This can be explained if the resolution of branch migration by isomerisation and repair is biased by array length, although it is not clear how these features might be reconciled. If branch migration is a relatively common feature of crossover then it is possible that crossover initiation is confined to a relatively small region, which is subsequently extended by branch migration and repair into the crossover hotspots observed at MS31a and MS32.

Evidence for crossing over has also been observed at CEB1 (J. Buard, pers. comm.) which suggests that flanking crossover hotspots may be a common feature of minisatellites. It therefore seems likely that this hotspot drives germline instability within the minisatellite repeat array. This also explains the observed polarity of conversion events. The relationship between these two process is further reinforced by evidence showing that the O1C variant at MS32 is able to suppress both gene-conversion and crossing over (Jeffreys *et al.*, 1998b). Gene-conversion events within the repeat array are thought to be the products of aborted crossover events. It is likely that the complex events typical of the majority of length-change mutants previous identified in sperm, such as duplication of the target site in the recipient and appearance of anomalous repeat are a direct consequence of this abortion. The role of minisatellites may therefore be to prevent the indefinite extension of unstable recombination complexes into regions of the genome where they could be potentially damaging. This strict regulation of minisatellite meiotic crossing over is suggestive of a potential role for minisatellites, possibly in chromosome synapsis and pairing during meiosis. This role has previously been inferred following the localisation of most minisatellites to the sub-telomeric regions of chromosomes (Royle *et al.*, 1988), which are known to be involved in these processes. Evidence from studies on mouse also suggest that the distribution of mouse minisatellites is associated with chromosomal regions implicated in the pairing of homologous chromosomes during meiosis (Carpenter *et al.*, 1987). This implies that minisatellites may have a conserved role in the mammalian genome.

Future work

Crossover analysis at MS31a is still in a highly preliminary stage and there is much work that can be done. Firstly, sequence analysis must be extended further upstream of MS31a to identify additional polymorphic sites to increase the region that can be used for this analysis. This will

enable the full extent of the hotspot to be defined. This strategy could then be applied to examine multiple different alleles of MS31a for a better understanding of meiotic crossover events at this locus and at minisatellites in general. Analysis of crossing over in both alleles of an individual would be necessary to determine whether the allele size dependent asymmetry of crossing over observed at MS32 is also applicable to MS31a. This could give some idea of the frequency of branch migration, resolution and repair events and may offer an explanation for the bias of these events observed at MS32. Although the actual breakpoints of recombination will only be identified by the direct examination of germ cells actively undergoing meiosis. Additional work can be carried out in conjunction with population studies to identify target alleles of potential interest e.g. the short, stable alleles identified in the Japanese and African populations (Huang *et al.*, 1996; Chapter 3). Eventually, these studies will probably lead to in depth studies of the biochemical mechanisms of recombination, perhaps resolving differences between mitotic and meiotic processes. This may also provide some explanation for differences observed in the frequency but not the type of minisatellite mutation in the male and female germline.

Chapter 6

Alu-mediated *de novo* deletion in the human genome

Summary

A significant proportion of genetically inherited diseases are caused by Alu-mediated recombination, giving rise to deletions, inversions and duplications of large regions of DNA. To date, *de novo* recombination events involving Alu repeats have been characterised solely by the investigation into the molecular basis of such diseases. To understand more about these low frequency events in the human genome it would be advantageous to uncouple the analysis of Alu-mediated recombination in the germline from the characterisation of the underlying molecular basis of disease. Two Alu-rich regions were used in this analysis, the first is located in non-coding DNA upstream of the human hypervariable minisatellite, MS32, and the second spans exon 4 of the C1-inhibitor gene. The first step in this process was to perform an evolutionary investigation into mutation at these loci. This involved amplification and comparative analyses of these regions in orang-utan, chimpanzee, gorilla and human. Evidence for base substitutions and, to a lesser extent, structural rearrangements were found at, or adjacent to both loci. The degree of instability was then examined in more detail by human population analysis which showed both loci to be monomorphic. From these combined data the upper estimates of mutation rate at both loci were estimated to be very low, in the order of 10^{-6} per allele in all populations. Despite the apparent inherent stability of these loci, these figures suggested that it was theoretically possible to access deletion mutation at these loci although the isolation of gain mutants was unlikely. To this end, a system was developed to isolate and characterise *de novo* deletion events in germline DNA. This involved physical size-selection and PCR recovery of single mutant molecules from sperm DNA. Previous use of this technique has shown that it is capable of detecting large deletion events occurring at a frequency of over 10^{-7}. No deletion mutants derived from either locus were isolated from sperm DNA. This analysis was also repeated on somatic tissue (blood) from the same individual with the same results. This indicated that large deletion events at these loci occurs at a rate lower than that which can be detected by the most sensitive mutation analysis system currently available.

Introduction

Alu repeats are the most abundant of the short interspersed repetitive elements (SINEs), making up around 6% of the human genome. The consensus target site of Alu integration is 5′ TTAAA 3′ (Jurka & Klonowski, 1996) which is often found within the poly-A regions of the element itself, explaining the propensity of Alus to cluster together. Alus evolved from DNA complementary to processed 7SL RNA (Kapitonov & Jurka, 1996) and have spread throughout

the genome by retrotransposition of a small number of "master" Alu elements which arose at different times during the evolution of primates (Deininger & Daniels, 1986; Clough *et al.*, 1996). The modern day descendants of these master elements have been divided into 12 subfamilies on the basis of shared substitutions at diagnostic positions, with each subfamily member showing approximately 80% sequence similarity to any other (Saffer & Thurston, 1989). Alus also exhibit internal homology because they are heterodimeric sequences (Figure 6.1B). They appear to have no general biological function, although they are sometimes expressed, usually as a part of a larger transcript (e.g. Chang *et al.*, 1994; Makalowski *et al.*, 1994; Margalit *et al.*, 1994).

Alu retrotransposition and function

Both Alus and L1s retrotranspose by the same mechanism, via an RNA intermediate (Britten *et al.*, 1988), all the proteins for which are encoded by LINE elements. SINEs do not have this ability to autonomously replicate and probably hijack the L1 retrotransposition machinery which appears to recognise and bind poly-A tails of these elements (Moran *et al.*, 1996). Alu RNA has retained the ability of its ancestor 7SL RNA to bind proteins of the signal recognition particle (SRP) with high affinity. This may allow it to position its poly-A tail to interact with the nascent L1 reverse transcriptase enzyme as it emerges from the SRP (Boeke, 1997). This may be a general mechanism of transposition because most SINEs are derived from RNA components of the translational machinery and typically have 3' tails (Smit, 1996). Alus and L1s also appear to integrate into very similar specific target regions (Jurka & Klonowski, 1996; Jurka, 1997) which is curious because L1s are found mainly in AT rich DNA and Alus preferentially integrate into GC rich regions.

The majority of dispersed repeats are believed to be parasitic and to represent junk DNA in host organisms. However, in certain cases, these elements appear to have gained novel functions and have become involved in biological processes (Martignetti & Brosius, 1993; Hambor *et al.*, 1993; Britten, 1996; Shimamura *et al.*, 1998). Recently, a number of genome wide roles of these elements have been proposed e.g. genomic imprinting (Rubin *et al.*, 1994; Liu *et al.*, 1994), protein translation (Bovia *et al.*, 1997; Chang *et al.*, 1994; Chu *et al.*, 1998), and the cell stress response (Russanova *et al.*, 1995; Liu *et al.*, 1995b). The suppression of mutation in potential protein binding regions further suggests that Alus have a sequence dependent function that is selectively important (Britten, 1994; Vansant & Reynolds, 1995). Alus have been used in evolutionary analysis (e.g. Knight *et al.*, 1996); in population studies of human genetic variation (Batzer *et al.*, 1996); in individual identification in forensic analysis (Novick *et al.*, 1995; and in physical mapping (e.g. Goldberg *et al.*, 1993; Mahadevan *et al.*, 1993).

Alu elements and genome instability

The abundance and sequence similarity of Alu elements makes them ideal targets for recombination events which result in gross genomic rearrangements. These events can occur both between and within alleles, causing deletions, duplications and inversions, with deletions

being the most common. Rearrangements can be homologous, occurring between pre-existing Alus, or non-homologous, where one breakpoint does not involve Alu elements. Alu-mediated recombination was discovered by virtue of the fact that it contributes to a number of genetic diseases, e.g. thalassemia (Nicholls *et al.,* 1987), premature artherosclerosis (Karathanasis *et al.*, 1987) and hereditary angioedema (Stoppa-Lyonnet *et al.*, 1990, 1991; Ariga *et al.*, 1990), and has recently been shown to be a frequent type of mutation in cancer susceptibility genes (Nystrom-Lahti *et al.*, 1995; Mauillon *et al.*, 1996, Petrij-Bosch *et al.*, 1997; Swensen *et al.*, 1997; Puget *et al.*, 1997; Levran *et al.*, 1998).

Timing of mutation

Two examples of Alu-mediated mutation in the soma have been characterised recently in the blood of leukaemia patients (Jeffs *et al.*, 1998; Strout *et al.*, 1998). In the *ALL1* gene, tandem duplications of sequences which end within Alu elements give rise to acute myeloid leukaemia (Strout *et al.*, 1998). These duplications are reminiscent of somatic mutation observed at other repeat types (Jeffreys & Neumann, 1997), and may have occurred by unequal sister chromatid exchange or by intra-strand recombination. It is difficult to differentiate between the two events, but the lack of evidence for the reciprocal deletion suggests that these duplications probably do not arise by unequal exchange. In cases of chronic myeloid leukaemia, Alu elements are found at, or near the 3′ breakpoint within the *BCR* gene on chromosome 22, with the other breakpoint found in a non-Alu region within the *ABL* gene on chromosome 9 (Jeffs *et al.*, 1998). In this context Alu repeats are thought to bring dispersed chromosome regions together and help stabilise a putative recombination complex (see below).

In the germline, Alu-mediated mutation is thought to occur mainly by inter-allelic crossing over during meiosis. This is based solely on evidence that XX maleness is caused by Alu-mediated recombination between the X and Y chromosomes (Rouyer *et al.*, 1987), which only pair at meiotic prophase I. Whether Alu-mediated germline recombination is restricted solely to meiosis remains to be seen.

Alus as mediators of mutation

There may be a substantial number of rearrangements in which Alu elements act to promote recombination. This is incredibly difficult to quantify but may occur by these elements bringing together distant regions which themselves show little or no sequence homology. In fact, many larger Alu mediated-deletions involve very little homology, between 2-7 bp at the breakpoints (Jalanko *et al.*, 1995), which is a common feature of large deletions (Morris & Thacker, 1993). In these and other examples, recombination breakpoints are not always found within Alus even though the breakpoint region may be highly enriched for them (e.g. Jeffs *et al.*, 1998; Nicholls *et al.*, 1987; Campbell *et al.*, 1995; Vnencak-Jones & Phillips, 1990). This illegitimate recombination suggests that Alu elements represent efficient sites for recombination initiation (for example regions of accessible chromatin), but not necessarily crossover resolution.

Most breakpoints identified in Alu-rich regions of the genome are homologous, occurring between two of these repeats. This may be due to recombinogenic sequences within the Alu element itself. Rüdiger *et al.* (1995) observed that breakpoints of Alu-mediated recombination often occurred within a common 26 bp core sequence (Figure 6.1C), and suggested that this sequence may be a requirement for recombination. This core contains part of the *chi* sequence that stimulates recBC mediated recombination in *E.coli*, and may represent a functional similarity. This core sequence is highly conserved in Alu repeats which undergo homologous recombination, and some sequence identity to this region was also found in non-homologous Alu-mediated recombination (Rüdiger *et al.*, 1995). In fact, the homology between different Alu sequences is far greater within this core region (~96.7%) than outside this region (~80%) (Rüdiger *et al.*, 1995). Interestingly, this Alu core sequence also contains a putative substrate site for the enzyme, topoisomerase I (Figure 6.1C) which is known to mediate illegitimate recombination *in vitro* (Konopka, 1988).

Features of Alu-mediated recombination

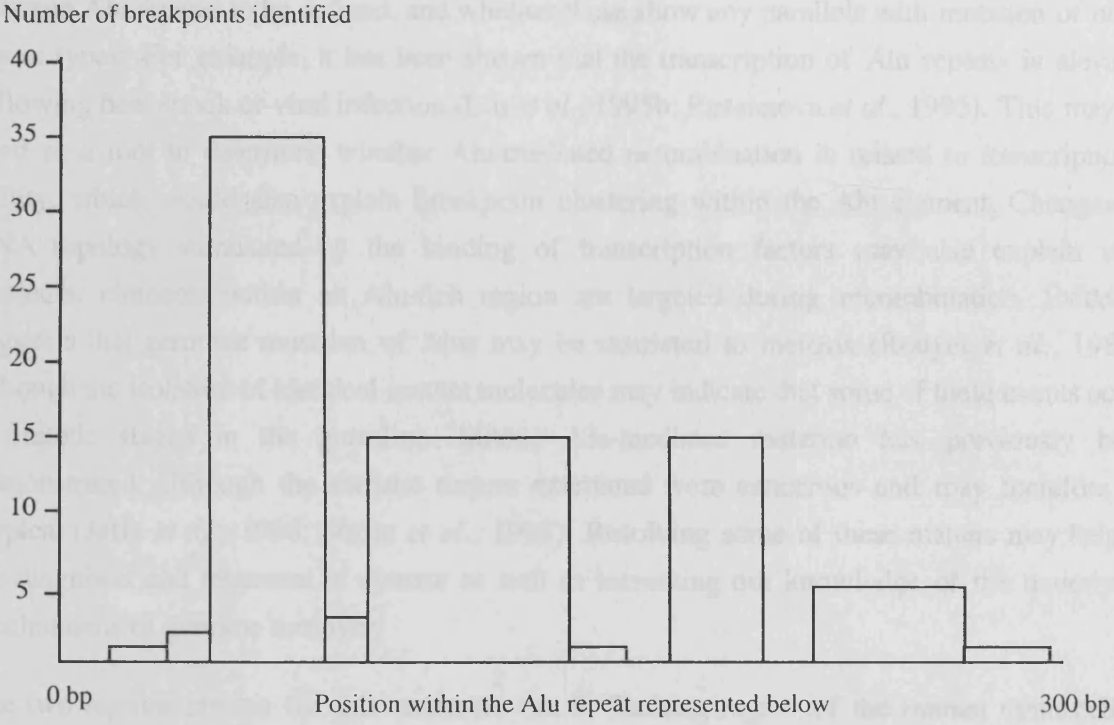
Frequently, one particular Alu element within a cluster appears to be predisposed to recombination (Stoppa-Lyonnet *et al.*, 1991; Levran *et al.*, 1998; Strout *et al.*, 1998; Chae *et al.*, 1997), and in some cases multiple rearrangements appear to be identical (Berkvens *et al.*, 1990; Jiang *et al.*, 1997; Hori *et al.*, 1995; Pousi *et al.*, 1994; Heikkinen *et al.*, 1994). The reason behind this targeting is unclear; it could be some sort of positional effect, or perhaps due to strong recombinogenic sequences within the element. However, it must be remembered that the breakpoint may not represent the initiating crossover of recombination, it may simply be the resolution point of structures such as Holliday junctions formed during recombination.

In most cases of Alu-mediated recombination, breakpoints are clustered between the A and B boxes of the RNA polymerase III promoter in the left monomer (Figure 6.1A). This may be a transcriptional effect because the left hand monomer is more efficiently transcribed *in vitro* than the right (Schmid, 1996). This suggests that only those Alus with transcriptional ability, namely the younger Alu subfamilies will recombine, which is often, but not always the case. It may be the process of transcription itself or changes in DNA topology caused by binding of transcription factors which promote recombination. The best known example of Alu-mediated mutation which does not exhibit breakpoint clustering is found in the C1-inhibitor (*C1NH*) gene, which is rearranged in the genetic disease hereditary angioedema. Here, the breakpoints are spread randomly over the entire length of the Alu element (Stoppa-Lyonnet *et al.*, 1991; Ariga *et al.*, 1990). This may be due to some influence of the flanking DNA or gene-specific features that affect recombination (Stoppa-Lyonnet *et al.*, 1991).

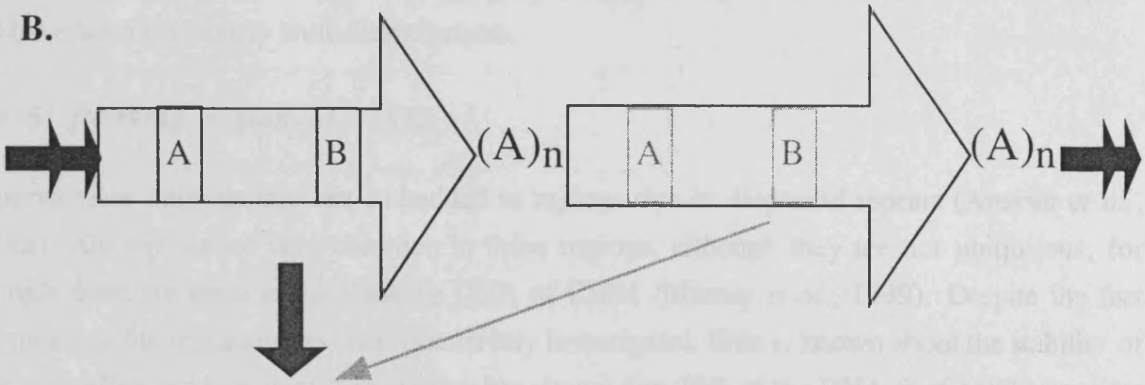
Mutation analysis in Alu-rich regions

Little is known about the mechanisms or the frequency of recombination between Alu elements in the human genome. This work attempts to address this using techniques which have successfully been applied to loci with high *de novo* mutation rates, such as minisatellites (see

A.



B.



C.

CCTGTAATC**CCAGC**ACTTTGGGAGGC

CCA**CCAGC** complementary *chi* sequence

CCT putative topoisomerase site

Chapter 5). Instead of examining the molecular basis of mutation in patients, the aim was to directly isolate mutants occurring in Alu-rich regions from germline (sperm) DNA. Detailed characterisation of these mutants would hopefully allow some of the features of recombination between Alu repeats to be defined, and whether these show any parallels with mutation of other repeat types. For example, it has been shown that the transcription of Alu repeats is elevated following heat shock or viral infection (Liu *et al.*, 1995b; Russanova *et al.*, 1995). This may be used as a tool to determine whether Alu-mediated recombination is related to transcriptional ability, which would also explain breakpoint clustering within the Alu element. Changes in DNA topology stimulated by the binding of transcription factors may also explain why particular elements within an Alu-rich region are targeted during recombination. Evidence suggests that germline mutation of Alus may be restricted to meiosis (Rouyer *et al.*, 1987), although the isolation of identical mutant molecules may indicate that some of these events occur in mitotic stages in the germline. Mitotic Alu-mediated mutation has previously been demonstrated, although the somatic tissues examined were cancerous and may therefore be atypical (Jeffs *et al.*, 1998; Strout *et al.*, 1998). Resolving some of these matters may help in the diagnosis and treatment of disease as well as increasing our knowledge of the underlying mechanisms of genome turnover.

The two regions chosen for this research, the 5′ flanking region of the human minisatellite, MS32 (Figure 6.2A) and the C1 inhibitor gene (Figure 6.2B) are both enriched in Alu repeats and have been previously well characterised.

The 5′ flanking region of MS32

Hypervariable minisatellites are embedded in regions rich in dispersed repeats (Armour *et al.*, 1989b). Alu repeats are very common in these regions, although they are not ubiquitous, for example there are none in the flanking DNA of CEB1 (Murray *et al.*, 1999). Despite the fact that minisatellite mutation has been extensively investigated, little is known about the stability of their immediate environment. Sequencing has shown that 29% of the DNA 5′ of MS32 consists of Alu elements (Murray *et al.*, 1999), which suggested that this may be an ideal target for recombination. In addition, Jeffreys *et al.*, (1998b) have identified a recombination hotspot within the 5′ flanking region of MS32, which was not delineated until after this project was completed (see discussion). The original target for this work was an 8.4 kb fragment including a cluster of thirteen Alu repeats; 2 partial and 1 complete element in the forward orientation, and 1 partial and 9 complete elements in the reverse orientation (Figure 6.2A). These include some of the oldest, (subfamily J) and some of the youngest (subfamily Sb2) Alu repeats in the human genome (Figure 6.2A). A total of 78 potential deletion products due to Alu-Alu recombination can be predicated from this region.

The C1-inhibitor gene

The *C1NH* gene was chosen as the second site for analysis because Alu-mediated mutations within this gene have been previously identified in patients with the autosomal dominant

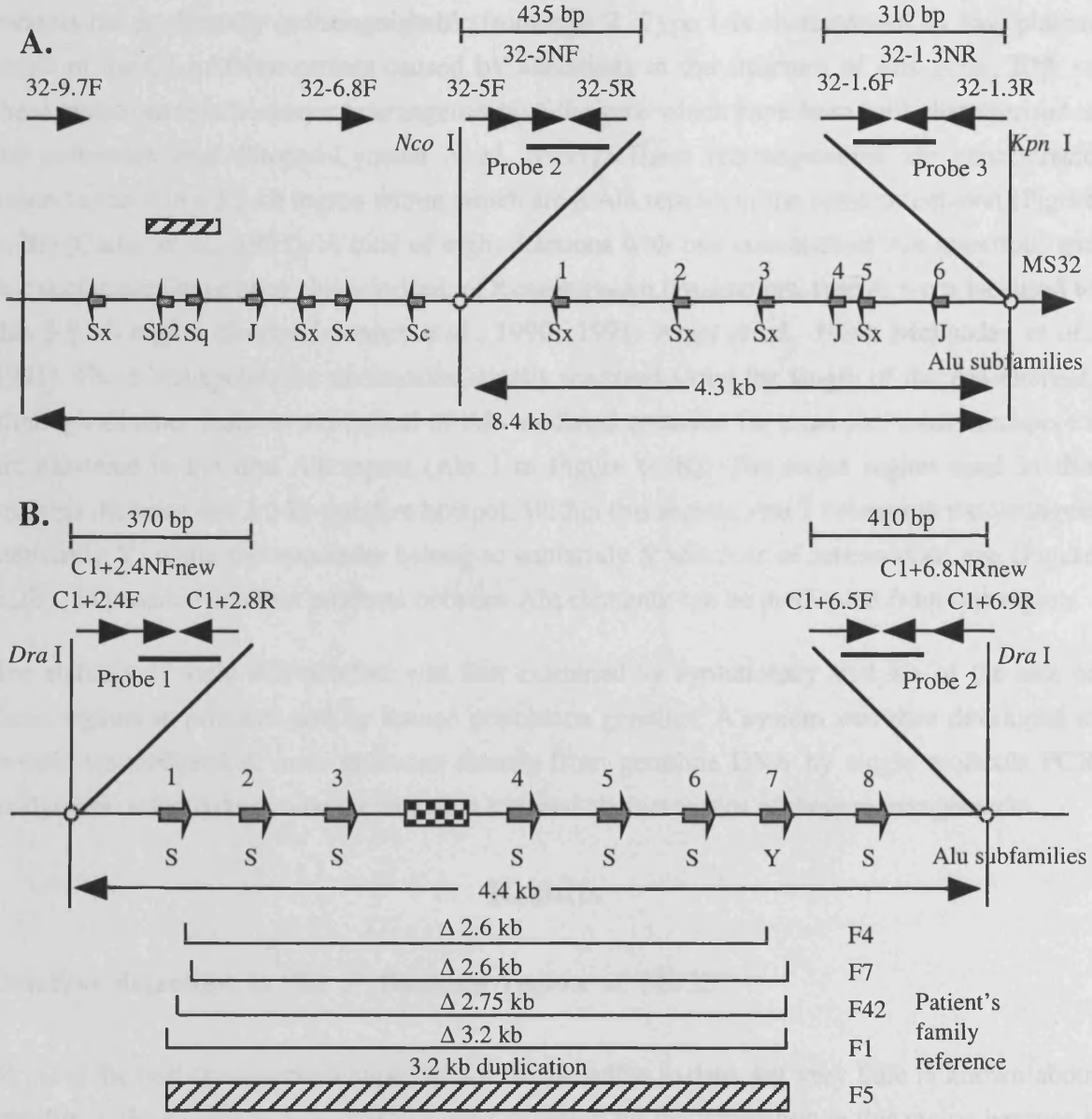


Figure 6.2. Alu-rich regions of the human genome analysed for instability.

Thick arrows represent the position and orientation of full length Alu elements. Subfamilies of all Alus and their designated number are also shown. Thin arrows represent the position and orientation of primers. Recognition sites for restriction enzymes used in size enrichment are indicated. Oligonucleotide probes are indicated by a thick line. Expanded regions give more detail of primer location and the region detected by the synthesised probes, size of these regions are indicated.

A. Schematic diagram of the 8.4 kb Alu-rich region in the 5′ flanking DNA of MS32 (Murray, *et al.*, 1999). A hatched bar shows the unamplifiable region with two nearest primers. The 4.3 kb region used in this analysis lies between the recognition sites for the restriction enzymes *Nco* I and *Kpn* I as indicated.

B. The 4.4 kb target region of the C1 inhibitor gene (*C1NH*). Exon 4 is represented by a chequered box (Carter *et al.*, 1991). Brackets denote the location and extent of each family specific rearrangement in HAE patients (Stoppa-Lyonnet *et al.*, 1987, 1991; Ariga *et al.*, 1990; McPhaden *et al.*, 1991). The hatched bar in family F5 shows the extent of a tandemly duplicated region (Stoppa-Lyonnet *et al.*, 1991).

disease, hereditary angioedema (HAE). There are two types of HAE, type 1 represents 85% of patients but is clinically indistinguishable from type 2. Type 1 is characterised by low plasma levels of the C1-inhibitor protein caused by alterations in the structure of this gene. 20% of these mutations involve major rearrangements of the gene which have been well characterised at the molecular level (Stoppa-Lyonnet *et al.*, 1991). These rearrangements are concentrated around exon 4 in a 3.5 kb region within which are 8 Alu repeats in the same orientation (Figure 6.2B) (Carter *et al.*, 1991). A total of eight deletions with one concomitant Alu insertion, and one duplication have been characterised; of these eighteen breakpoints, twelve were localised to this 3.5 kb region (Stoppa-Lyonnet *et al.*, 1990, 1991; Ariga *et al.*, 1990; McPhaden *et al.*, 1991). These breakpoints are uncharacteristically scattered along the length of the Alu element, although all other features are typical of Alu-mediated mutation for example, most breakpoints are clustered in the first Alu repeat (Alu 1 in Figure 6.2B). The target region used in this analysis includes this 3.5 kb putative hotspot. Within this region, Alu 7 belongs to the youngest subfamily Y, while the remainder belong to subfamily S which is of intermediate age (Figure 6.2B). 28 possible deletion products between Alu elements can be predicated from this region.

The stability of these Alu-rich loci was first examined by evolutionary analysis of the size of these regions in primates and by human population genetics. A system was then developed to isolate Alu-mediated *de novo* deletions directly from germline DNA by single molecule PCR analysis in order to determine the mutation rate and characteristics of these rearrangements.

Results

Deletion detection in the 5′ flanking region of MS32

MS32 is the best characterised hypervariable minisatellite to date, yet very little is known about stability of the flanking DNA. There is great potential for recombination in this region because it is rich in different repeat types, and it has been demonstrated that recombination extends upstream of the minisatellite repeat array into the 5′ flanking DNA (Jeffreys *et al.*, 1998b). The precise region used in this analysis lies 1.3 to 10 kb upstream of the unstable end of MS32 and includes a cluster of Alu elements.

Unfortunately, attempts to amplify the entire 8.4 kb region between primers 32-9.7F to 32-1.3NR were unsuccessful in all DNA types, including cosmid clones of this region (Murray *et al.*, 1999). Alternative pairs of primers within this region were therefore used to identify a better target for this analysis and to examine the reason for these difficulties during PCR (Table 2.4, Materials & Methods). The problematic region was located between primers 32-9.2R and 32-6.8F which includes two Alu repeats in opposite orientations (Figure 6.2A). Difficulties associated with amplification in this region may be due to "snap back" of these repeats to form a hairpin which could inhibit PCR. Amplification across this region was eventually achieved using primers 32-9.7F and 32-1.3R under long range PCR conditions, but this was difficult and highly prone to mis-priming giving many spurious bands. The region finally selected for

analysis was located between 1.3 kb and 5 kb upstream of the minisatellite which could be efficiently amplified using long-range PCR. Locus-specific probes 2 and 3 (Figure 6.2A) were used for detection of the amplified product during Southern hybridisation of this region. Probes were located at the extreme ends of the region in order to detect all potential Alu-mediated deletion products. Two sets of primers were essential for these analyses to prevent contamination of the primary stock of genomic DNA. The outside primers 32-5F and 32-1.3R were used for single molecule PCR during mutant screening; and nested primers, 32-5NF and 32-1.3NR were designed for evolutionary analysis, population screening and reamplification of this region following single molecule PCR. Primer pairs were designed to be as efficient as possible to maximise the probability of detecting deletion mutants arising at low frequency. This also minimised the risk of jumping PCR by inter-Alu annealing between incompletely extended PCR products. Products of jumping PCR can be preferentially amplified in subsequent cycles because they are often shorter than the progenitor and can interfere with mutant detection.

Evolutionary analysis of the Alu-rich region 5′ of MS32

Variation over evolutionary time was measured by comparative analysis of this locus in human, chimpanzee, gorilla and orang-utan. Semi-nested amplification and digestion of the region between primers 32-5F and 32-1.3NR in these primates showed no gross genomic rearrangement (Figure 6.3). Shared bands between the different primates (particularly following *Sma* I digestion) indicated that the region amplified in each species showed high sequence similarity with the human locus. This was confirmed by Southern hybridisation of these regions using the locus-specific probe 2 (data not shown). Subtle differences between the digests were interpreted as base substitution events indicated by the loss or gain of a restriction site. Examples of base substitution are demonstrated following restriction digestion with *Hind* III, *Bcl* I and particularly *Xba* I. Amplification of a region further upstream between primers 32-6.8F and 32-5R revealed that this was over 1 kb shorter in humans than in any of the great apes (Figure 6.4). Despite this evidence of instability, no further data was collected on this region because it was difficult to amplify in humans (see previous section).

This showed that either this region was highly stable over long periods of evolutionary time, or that any rearrangements which had occurred have not been fixed in any of these species. Human, chimpanzee and gorilla are thought to have diverged from a common ancestor approximately 7 million years ago (Koop *et al.*, 1986) and orang-utan diverged from this lineage about 3 million years previously. Therefore, these data has shown that no structural rearrangements have occurred in a total of ~34 million years of evolution of these primates, so the maximum mutation rate at this locus can be estimated at $< 3 \times 10^{-7}$ per generation (assuming an average generation time for these primates of approximately 10 years). However, the indel observed further upstream of the target region in primates indicates that rearrangements can occur at this locus. This rearrangement is probably relatively recent and suggests that this locus may be more unstable in human populations than other primates, which is certainly true of MS32, the downstream minisatellite (Gray & Jeffreys, 1991).

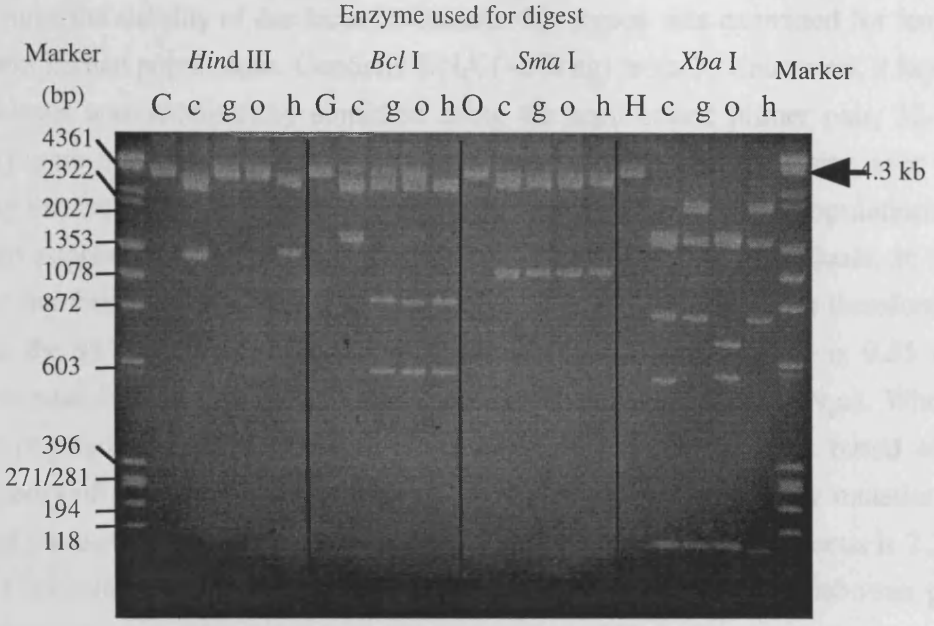


Figure 6.3. Amplification and comparative restriction digestion of the MS32 target region in primates.

Genomic DNA was amplified using primers 23-5NF and 32-1.3NR and digested with restriction enzymes indicated. Vertical lines separate sets of digests performed on amplified DNA from chimpanzee (c), gorilla (g), orang-utan (o) and human (h). 100 ng undigested amplified DNA from chimpanzee (C), gorilla (G), orang-utan (O) and human (H), all of 4.3 kb (arrow) was also included. Changes in the pattern of restriction enzyme digestion between species indicate that base substitution has occured within this region but the similarities of the patterns preclude any major structural rearrangements.

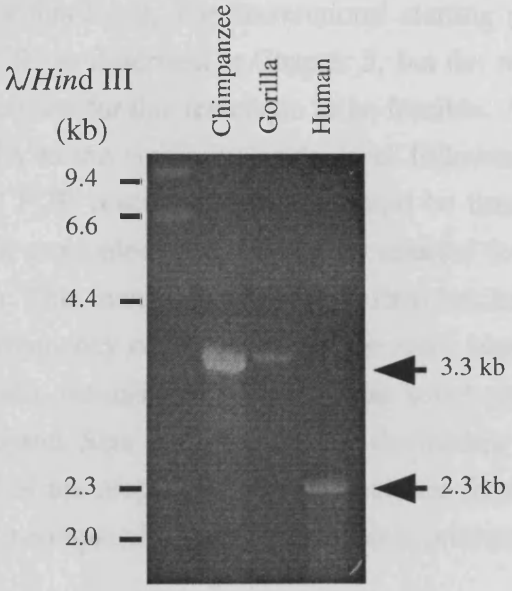


Figure 6.4. Evidence for structural rearrangement upstream of the MS32 target region in humans.

~100ng of genomic DNA from chimpanzee, gorilla and human was amplified using primers 32-6.8F and 32-5R demonstrating that the human sequence is 1.0 kb shorter, probably representing a deletion in the human genome.

Human population screening in this region

To determine the stability of this locus in humans this region was examined for length-changes in different human populations. Genomic DNA (~200 ng) from 20 Caucasian, 9 Japanese and 9 Zimbabweans was successfully amplified using the semi-nested primer pair, 32-5F and 32-1.3NR (Figure 6.5). All individuals examined were monomorphic for size over this region, indicating that no gross genomic rearrangements had occurred in these populations. Using the Caucasian population as an example, no mutants were seen in 20 individuals, so the Poisson corrected number of mutants was < 3 in 20. The heterozygosity rate (H) is therefore < 3 / 20 = < 0.15 at the 95% confidence limit, and the level of homozygosity (H_o) is 0.85 (1-H). This allows the mutation rate (μ) of this region to be calculated ($H_o = 1 / 1+4N_e\mu$). Where N_e is the effective population size, estimated at about 20,000. This calculation is based on Kimura's (1983) theory of neutral drift which states that the rate of creation of new mutation is equal to the rate of fixation of variants. The upper estimate of mutation rate at this locus is 2.2×10^{-6} per allele in Caucasians, and 6.3×10^{-6} per allele in both Japanese and Zimbabwean populations. The different estimates of rearrangement in these populations is likely to reflect the numbers of alleles screened in each population, rather than actual differences in mutation rate. These estimates of mutation rate are similar to that obtained following evolutionary analysis of primate DNA. However, these figures do not suggest that the inherent stability of this locus will preclude the detection of germline mutants by single sperm analysis.

Size enrichment for deletion mutants

Evolutionary studies and population genetics have demonstrated that the MS32 locus is stable, but techniques developed for the isolation of length change mutants by single sperm analysis may still be effective at this locus. The conventional starting point for the analysis of sperm DNA is small pool PCR, as described in Chapter 5, but the mutation rate predicted from the preceding work was too low for this technique to be feasible. A second technique involves the dilution of sperm DNA to the single molecule level followed by analysis of each separate molecule in individual PCR reactions, but this would be time-consuming and unproductive. Instead, deleted mutant molecules were specifically selected for by size-enrichment and single molecule amplification. This incredibly sensitive system has been used to detect large deletion events occurring at a frequency of 10^{-7}, and may be more sensitive than this (Jeffreys *et al.*, 1997). This technique also validates mutant molecules which must fall within the size range of the fraction being analysed. Size fractionation also diminishes the likelihood of swamping the mutant signal with that of the progenitor, and prevents the formation of spurious bands due to inter-Alu annealing of incompletely extended progenitor products.

Size enrichment was carried out under PCR clean conditions. All reagents, e.g. restriction endonucleases, digestion buffers, PCR buffer, DNA polymerases were kept separate from general stocks and maintained in a clean environment. DNA manipulations whenever possible were carried out under a laminar flow hood and, where appropriate all equipment (e.g. electrophoresis tanks, gel trays and other items) was soaked overnight in ~1 M hydrochloric

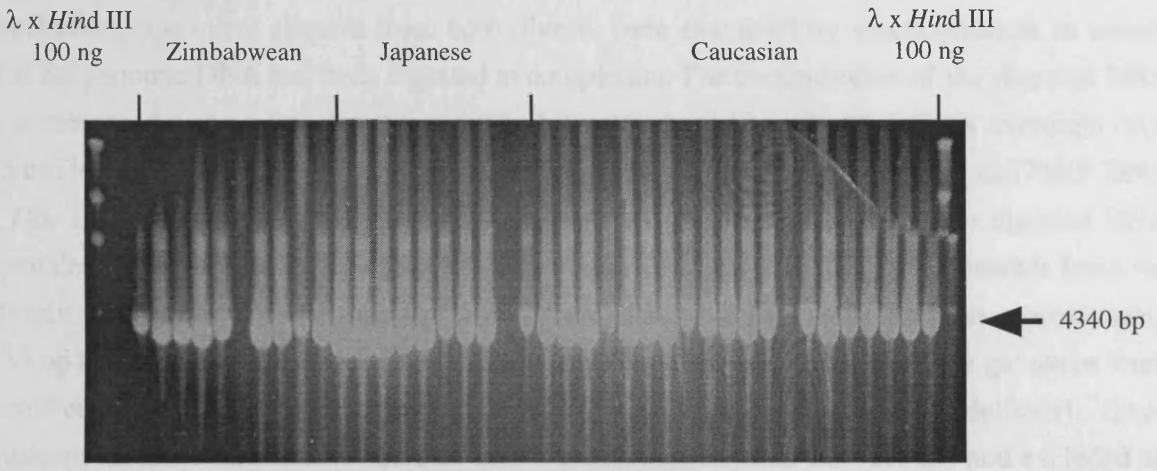


Figure 6.5. Human population screen for length polymorphisms at the MS32 locus.

~200 ng genomic DNA from 20 Caucasian, 10 Japanese and 10 Zimbabwean individuals was amplified using primers 32-5F and 32-1.3NR. All individuals gave a fragment of expected size, 4.3 kb (arrow) except DNA from one Zimbabwean and one Japanese individual which failed to amplify due to the reduced concentration of DNA in these samples. The lack of polymorphism observed in these individuals was used to estimate an upper limit of mutation rate for this locus in all three popultions (see text).

acid to degrade any potential contaminant DNA. All electrophoretic steps were carried out in the absence of ethidium bromide and UV light to minimise the DNA damaging effects of these agents. The loading dye used for size enrichment contained 3 g glycerol, 0.001 g bromophenol blue and 100 µl 10x TBE in 10 ml. This dense buffer reduced the risk of DNA escaping from the wells and migrating through the buffer which could complicate fractionation.

A total of approximately 60 µg sperm DNA (equivalent to about 20 million molecules) was digested with the restriction endonucleases *Nco I* and *Kpn I* to release the region of interest (Figure 6.2A). Digestion was performed in REact buffer 1 (GibCo BRL) with 3 units of *Nco* I and 2 units *Kpn* I per 1 µg input DNA, for 2.5 hr. After 10 min an aliquot containing ~1 µg of this digest was removed and over-digested for 2 hr in fresh digestion mix. Following incubation, equivalent aliquots from both digests were examined by electrophoresis to ensure that the genomic DNA had been digested to completion. The concentration of the digested DNA was measured using a fluorimeter, and 30 µg was separated by electrophoresis overnight on a 35 cm long 0.8% agarose gel with 400 ng each of phage λ DNA x *Hin*dIII and ϕx174RF DNA x *Hae* III markers. The position of marker bands and the lanes containing the digested DNA were determined following ethidium bromide staining of the end of the gel and marker lanes for 30 min. The position of the digested DNA corresponding to fragments between approximately 950 bp and 4.0 kb was estimated and divided into 10 gel slices or fractions. The gel slices were removed and the DNA electroeluted onto dialysis membrane (Materials & Methods). These fractions included all possible Alu-Alu deletion mutants (between 4.3 - 1.4 kb) and excluded all progenitor molecules (5.0 kb). A total of 30 µg of blood DNA from the same individual was digested and fractionated in parallel under the same conditions.

Analysis of the recovered size-enriched DNA

The concentration and size range of the DNA in each fraction was estimated by comparison of a 100th volume of the fractionated DNA to a dilution series of the unfractionated, digested starting DNA. Fragments were visualised by electrophoresis and Southern hybridisation with ^{32}P-labelled total genomic DNA (Figure 6.6). Yield was calculated by:

yield = amount of starting DNA with equivalent signal intensity x dilution factor

This demonstrated that the yield was uniform between fractions and that the fraction sizes were sufficiently discrete. The yield of the fraction closest in size to the progenitor (fraction 10) is < 5,000 ng/µl or ~1.7 x 10^6 molecules/µl, if the amount of starting DNA with equivalent signal intensity is < 50 ng/µl and the dilution factor is 0.01. Fraction yield varied between ~1.7 x 10^6 molecules/µl in the fraction closest to the progenitor, to ~1.0 x 10^7 molecules/µl in the smallest fraction, with an average recovery of ~6.7 x 10^6 molecules/µl. Following ethanol precipitation, the fractions were then dissolved in 20 µl 5 mM Tris-HCl pH 7.5 with 5 µg/ml herring sperm DNA as carrier. One tenth volume of the fractionated DNA and a dilution series of unfractionated, digested starting DNA were amplified using semi-nested primers 32-5F and 32-1.3NR to estimate the degree of enrichment following fractionation. Enrichment was calculated by:

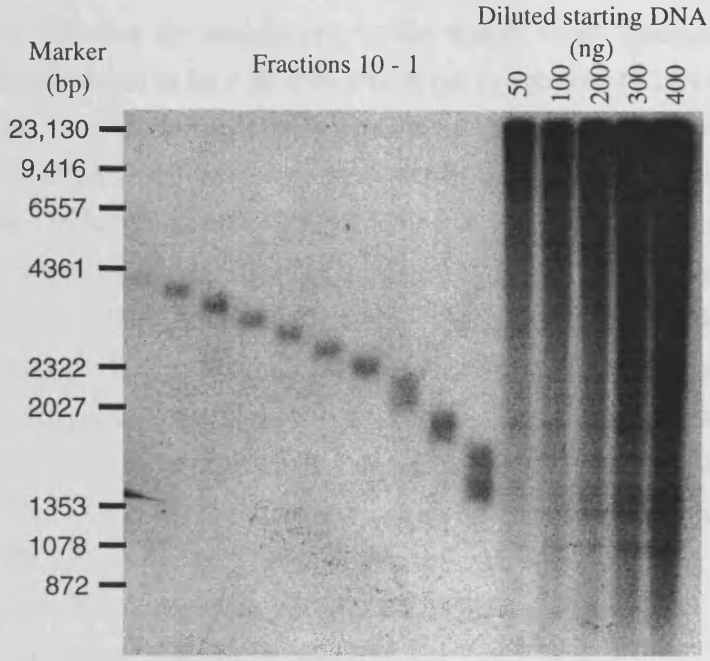


Figure 6.6. DNA from fractions recovered following size enrichment of sperm DNA.

Sperm DNA was digested with *Nco* I and *Kpn* I, electrophoresed through an agarose gel and size fractions ranging from 4.3 to 1.4 kb were collected by electroelution. Aliquots of the recovered DNA were electrophoresed and detected by Southern hybridisation with total genomic DNA. The size range of each fraction was estimated by comparison with marker bands. Concentration of the recovered DNA was estimated by comparison of the signal intensity in each fraction and the signal intensity in the equivalent size range from a dilution series of DNA of known concentration.

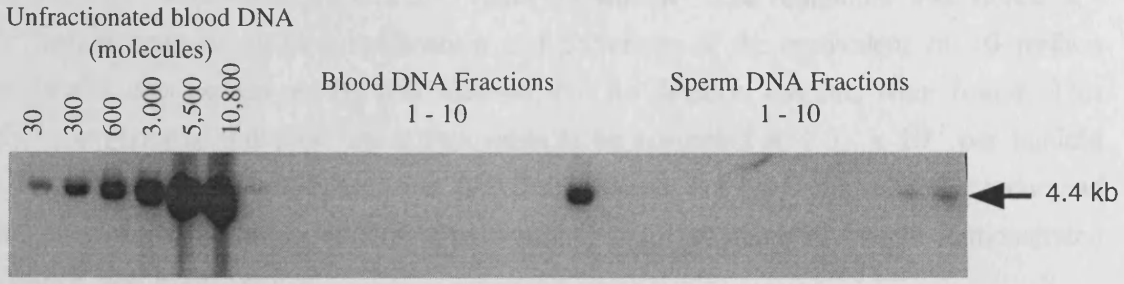


Figure 6.7. Amplification of the MS32 region in fractions recovered following size enrichment of blood and sperm DNA.

1/100 diluted aliquots of each size fraction were amplified using primers 32-5F and 32-1.3NR. The degree of enrichment of each fraction was estimated by comparison of hybridisation signal with signals from 30-10,800 molecules of unfractionated blood DNA digest. Size and location of progenitor is indicated by the arrow.

$$Enrichment = 100 - \left[\left(\frac{proportion\ of\ DNA\ loaded\ to\ gel}{yeild} \right) x \begin{array}{c} number\ molecules\ in\ starting\ DNA \\ with\ equivalent\ signal\ intensity \end{array} \right]$$

From the above equation the enrichment in the sperm DNA fraction closest in size to the progenitor, was calculated to be < 99.6%, where the proportion of DNA loaded is 0.1, the yield is 1.7 x 10^6, and the signal intensity is equivalent to < 75 molecules of starting DNA. The remaining sperm DNA fractions were over 99.9% depleted in progenitor (Figure 6.7). It is possible that the DNA can become degraded by this process but unfortunately, the quality of this DNA is not known. DNA quality can be tested by single molecule analysis across a fraction containing progenitor DNA, but such fractions were not collected during this work. However, the analysis of multiple loci in this laboratory has demonstrated that very little DNA damage is incurred using this technique. Comparison of the signal intensity obtained following semi-nested amplification and hybridisation of 1 ng unfractionated DNA in the presence and absence of 1 ng fractionated DNA ensured that there were no substances in the fractionated DNA which could inhibit PCR (Figure 6.8). This also showed that the equivalent of 2 μg unfractionated DNA of fractionated DNA could be put into the PCR reaction without detection of progenitor, indicating that enrichment was almost 100%.

Screening the fractionated sperm DNA for deletion mutants

Fractions were screened for deletion mutants using the outside primers 32-5F and 32-1.3R to minimise the risk of external contamination from previously amplified PCR products. Between 250 - 975 ng DNA (the equivalent of 8.3 x 10^4 - 3.3 x 10^5 progenitor molecules) was screened in a single 7 μl PCR reaction. Fractions were amplified separately or in a pool with other fractions of similar size, and in the presence of 1 μg/ml herring sperm DNA as carrier. In all screens a positive control of 10 ng unfractionated, digested starting DNA was also amplified. Half the volume of each PCR reaction was analysed by electrophoresis and Southern hybridisation with ^{32}P-labelled probe 2 or 3 (data not shown). The remainder was stored at -20°C for further analysis. Bulk amplification and screening of the equivalent of 10 million sperm molecules was carried out by this method, but no deletion mutants were found. This allows the true germline mutation rate at this locus to be estimated at < 3 x 10^{-7} per haploid genome. This work demonstrated that large germline deletions are not common at this locus and confirmed previous evolutionary studies and human population analysis which demonstrated that this region was highly stable.

Analysis of somatic mutation at this locus

As a comparison, fractionated blood DNA from the same individual was also screened for deletion mutants. Analysis of the equivalent of 5 million molecules of this somatic DNA revealed no mutants. The somatic rate of mutation at this locus was therefore estimated to be < 5 x 10^{-6} per progenitor molecule. This was expected because to date, somatic Alu-mediated rearrangements have only been reported in cancerous tissue (Jeffs *et al.*, 1998; Strout *et al.*, 1998).

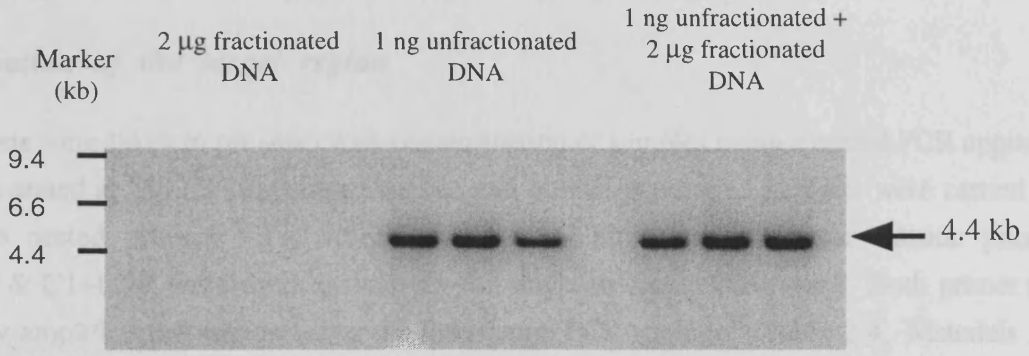


Figure 6.8 "Seed" experiment to show PCR was not inhibited by any component within the fractionated DNA.

DNA from fractionated and unfractionated DNA of known concentration were amplified using primers 32-5F to 32-1.3R and detected by Southern hybridisation. Amplification of the equivalent of 2 µg DNA from fraction 1 alone showed no progenitor. Signal intensities obtained following amplification of 1 ng progenitor DNA alone and 1 ng progenitor DNA mixed with 2 µg fractionated DNA are equivalent, showing there was no inhibition of amplification from the fractionated DNA. Size and location of the progenitor band is indicated with an arrow.

It is possible that the deletion rate at this locus is too low to be detected by this system in germline or somatic tissue. This analysis was therefore extended to a locus at which deletions have been well characterised, and for which positive deletion controls can easily be obtained.

Detection of mutation in the C1-inhibitor gene

The Alu-rich region around exon 4 of the C1-inhibitor gene is a common target for Alu-mediated rearrangements in HAE patients (Stoppa-Lyonnet *et al.*, 1991). In fact, most of the breakpoints are clustered within the first Alu repeat of this region, designated Alu 1 (Figure 6.2B). The 4.4 kb target region used in this analysis includes the 3.5 kb putative recombination hotspot which contains 8 Alu repeats in the same orientation (Figure 6.2B).

Amplification of the target region

Precautions were taken to prevent cross-contamination of samples using a nested PCR approach as demonstrated at MS32. All primate studies and human population analysis were carried out using the nested primers C1+2.4NFnew and C1+6.8NRnew, while the outside primers C1+2.4F & C1+6.9R were used exclusively for single-molecule PCR work. Both primer sets efficiently amplified this region using the long-range PCR protocol (Table 2.4, Materials and Methods). Locus-specific probes 1 and 2 (Figure 6.2B) were end probes designed to be able to detect any predicted Alu-mediated deletion event by Southern hybridisation. The advantage of using a known disease locus was that the PCR system could be tested before sperm analysis was attempted. To do this, blood DNA obtained from HAE patients (kindly donated by Mario Tosi, Pasteur Institute, Paris) was digested using the restriction enzyme, *Dra* I (Figure 6.2B) and diluted to the single molecule level. The DNA was then amplified using the outside primers C1+2.4F and C1+6.9R, and amplified fragments were detected by electrophoresis and Southern hybridisation. This demonstrated that the assay system was capable of detecting single copies of normal and deleted molecules of this region from genomic DNA in the same amplification reaction (Figure 6.9).

Structural arrangement of the C1NH *region in primates*

To give some indication of the mutation rate at the *C1NH* locus this region was examined for structural rearrangements over evolutionary time. Analysis of this locus in chimpanzee, gorilla, human and orang-utan was carried out by amplification using nested primers C1+2.4NFnew and C1+6.8NRnew. Restriction digestion of this amplified product in human, gorilla and chimpanzee revealed no structural rearrangements, but in orang-utan this locus was ~200 bp smaller than in the other primates (Figure 6.10). Several sites of base substitutional polymorphism were identified (Figure 6.11), for example the loss of a *Bgl* I site in orang-utan, and the gain of an additional *Eco* RI site in gorilla can be explained by base substitution. The smaller *Eco* RI and *Rsa* I fragments obtained following digestion of orang-utan DNA are consistent with the loss of ~200 bp predicted at this locus. The overall similarity between

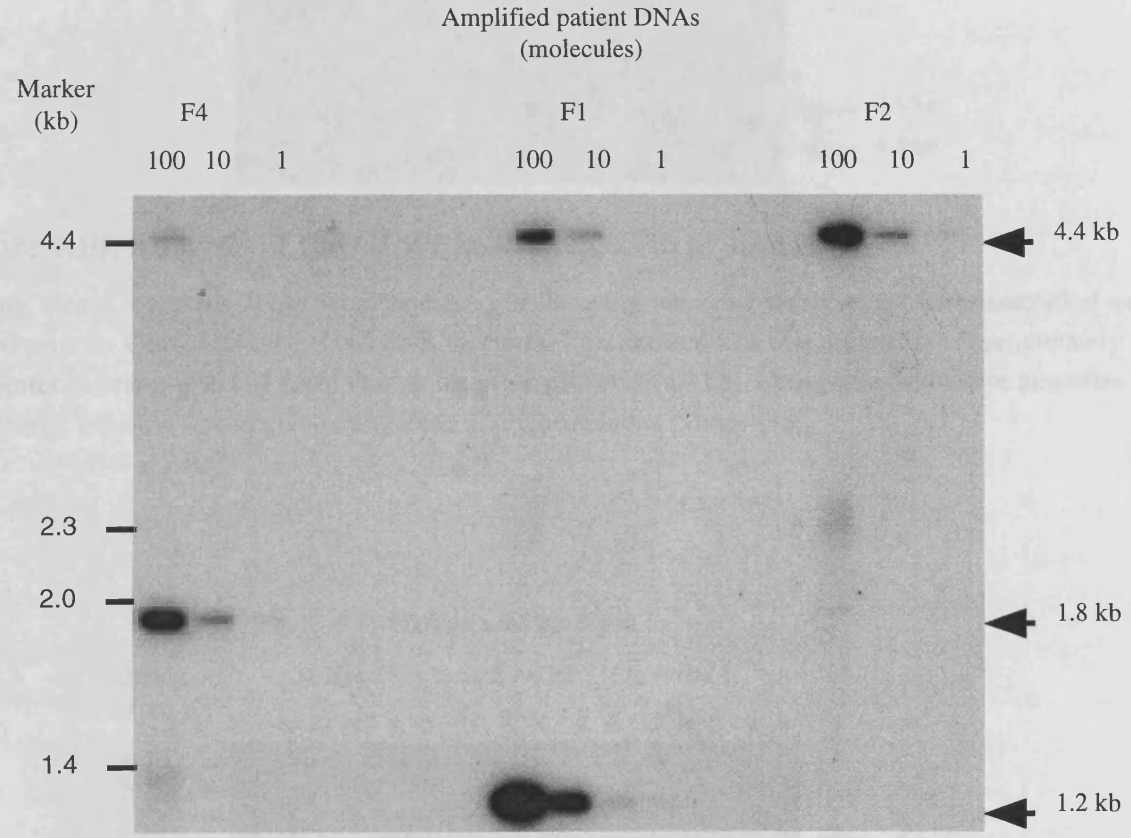


Figure 6.9. Detection of normal and deleted alleles at the single molecule level in the target region of the C1-inhibitor gene.

DNA from three HAE patients was digested with *Dra* I, diluted to the single molecule level and amplified with nested primers C1+24NF & C1+6.8NR. Patient F4 has a 2.6 kb deletion in this region and one normal allele. Patient F1 has a 3.2 bp deletion in this region and one normal allele. Patient F2 has a deletion in the *C1NH* gene outside this target region and therefore has two normal alleles. Normal (4.4 kb) and deleted (1.8 and 1.2 kb) alleles are indicated with arrows. See Figure 6.2B for position and extent of breakpoints (Stoppa-Lyonnet *et al.*, 1987 & 1991; Ariga *et al.*, 1990; McPhaden *et al.*, 1991).

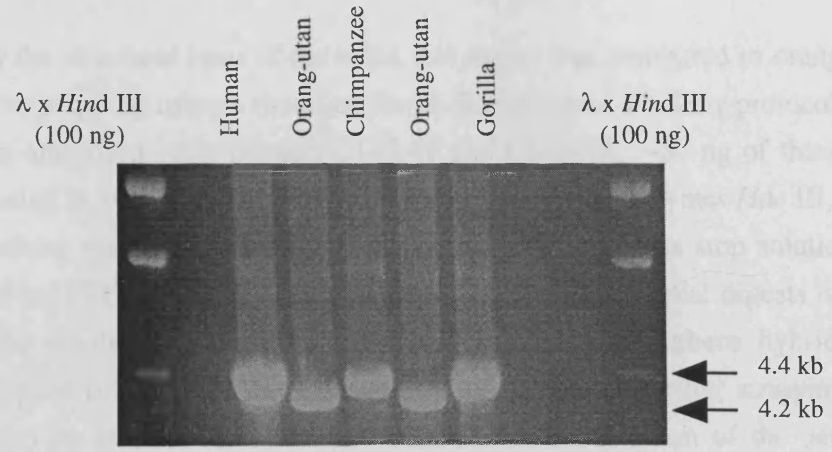


Figure 6.10. Analysis of the *C1NH* target region in primates.

100 ng *Dra* I digested DNAs from human, gorilla,orang-utan and chimpanzee were amplified using nested primers C1+2.4NF and C1+6.8NR (arrows). This showed that this region was approximately 200 bp shorter in orang-utan (~4.2 kb) than in the other primates (4.4 kb). Orang-utan DNA also amplifies less efficiently, possibly due to greater sequence divergence in the primer sites.

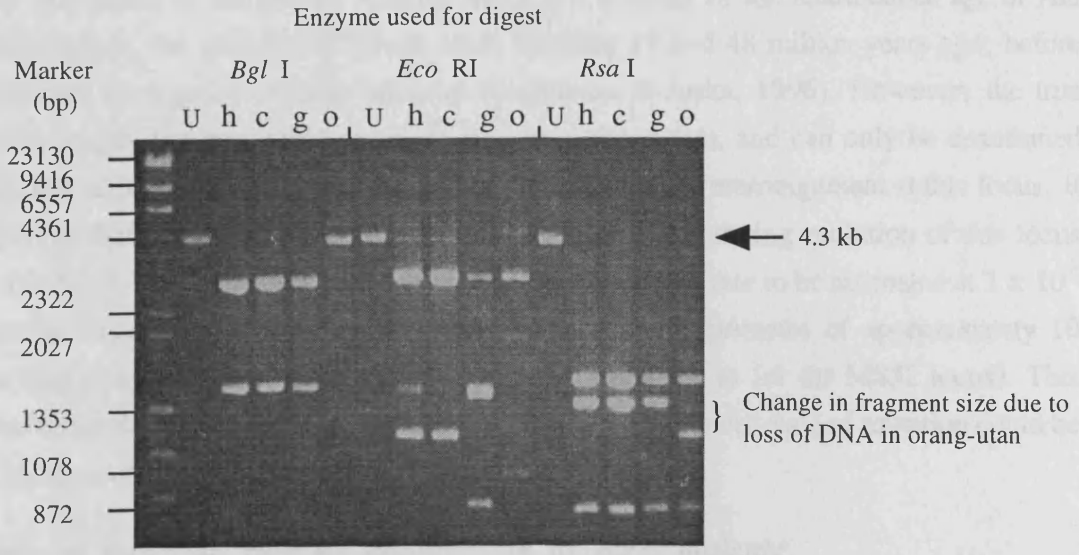


Figure 6.11. Amplification and comparative restriction digestion of the C1 target region in primates.

DNA from human (h), chimpanzee (c), gorilla (g) and orang-utan (o) was amplified using primers C1+2.4NF and C1+6.8NR and digested using restriction enzymes indicated. The undigested human fragment of 4.3 kb was also included (U). Evidence of base substitution can be seen in orang-utan following digestion with *Bgl* I and following *Eco* RI digestion in gorilla. Following *Rsa* I digestion the 1.2 kb fragment seen in orang-utan instead of the 1.4 kb fragment (indicated by a bracket) was consistent with the ~200 bp smaller size of this locus (Figure 6.10).

fragments obtained following restriction enzyme digestion showed that these amplified products were derived from the locus identified as the *C1NH* gene in humans.

To identify the structural basis of the indel, this region was compared in orang-utan and human by restriction mapping using a modified Smith-Birnstiel end-labelling protocol. DNA from both species was amplified using primers C1+2.4F and C1+6.9R. ~50 ng of these amplified DNAs were incubated in 10 µl reaction volume with the restriction enzymes *Hae* III, *Hinf* I, *Mbo* I and *Sty* I. Reactions were stopped after 80 sec by the addition of 5x stop solution (10 mM EDTA pH 8.1, 3.6 mM Tris-HCl pH 7.5, 5 x loading buffer). The partial digests obtained from both species were resolved by electrophoresis and detected by Southern hybridisation, firstly to probe 2 (Figure 6.12B) and then to probe 1 (Figure 6.12C) after stripping all radiolabelled probe 2 from the membrane (Materials & Methods). Comparison of the patterns obtained by partial digestion of human and orang-utan DNA showed the region absent in orang-utan was between approximately 600 - 1207 bp from the 5′ end of the human amplicon (Figure 6.12C). The location of the indel cannot be defined more accurately because this technique depends on the position of the restriction enzyme sites, which were not informative enough for more precise estimates. Within this 600 bp predicted region of sequence difference between the species lies the element designated Alu 2 (length 265 bp) in the human sequence. This rearrangement may therefore represent insertion of this element after the divergence of orang-utan from the other primates, or loss of existing DNA from the orang-utan locus. The most likely explanation is deletion of this region in orang-utan because Alu 2 is a member of the intermediate age of Alu subfamilies (Alu S), the majority of which arose between 19 and 48 million years ago, before the evolutionary divergence of these primates (Kapitonov & Jurka, 1996). However, the true cause of this length change could not be inferred from these data, and can only be determined by sequencing, although this does provide evidence for structural rearrangement at this locus. It is also possible that other rearrangements may also have occurred during evolution of this locus but were not fixed. These data allowed an upper limit of mutation rate to be estimated at 3×10^{-7} per generation (assuming an average generation time for these primates of approximately 10 years in a total of 34 million years of evolution of these primates; as for the MS32 locus). This analysis investigated long periods of evolutionary time but a better estimate of mutation could be obtained by measuring variation in human populations.

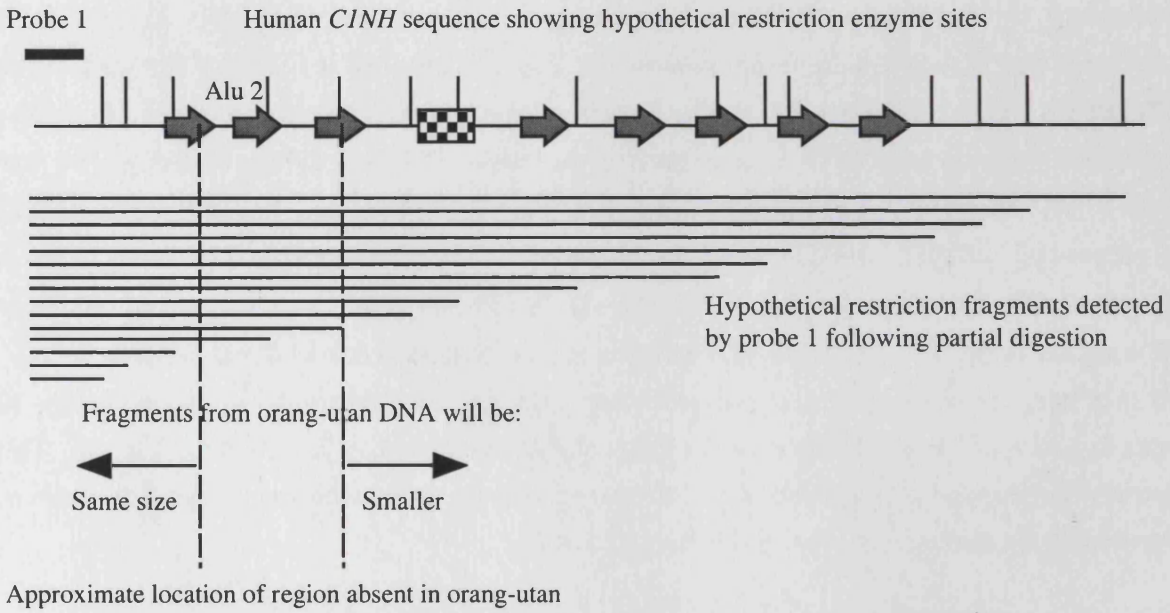
Estimation of mutation rate by examination of HAE patients

The majority of rearrangements at this locus are related to disease allowing mutation rate at this locus to be derived from known frequencies of HAE. The frequency of HAE in the world population (F_{HAE}) is between 5×10^{-4} and 1×10^{-5} which is divided unequally between Type 1 and 2 forms of the disease. Type 2 HAE is due to regulatory defects and so is not important for this analysis. Type I HAE represents about 85% of cases, of which about 20% are due to major structural changes in the *C1NH* gene (Stoppa-Lyonnet *et al.*, 1991). The rate of *de novo* mutation is estimated to be about 20% of all known HAE mutations (Agostoni & Cicardi, 1992), but unfortunately the proportion of these which are paternal in origin is not known. From this information the *de novo* rate of major structural rearrangements can be calculated to

Figure 6.12. Comparison of the *C1NH* region in orang-utan and human by restriction analysis.

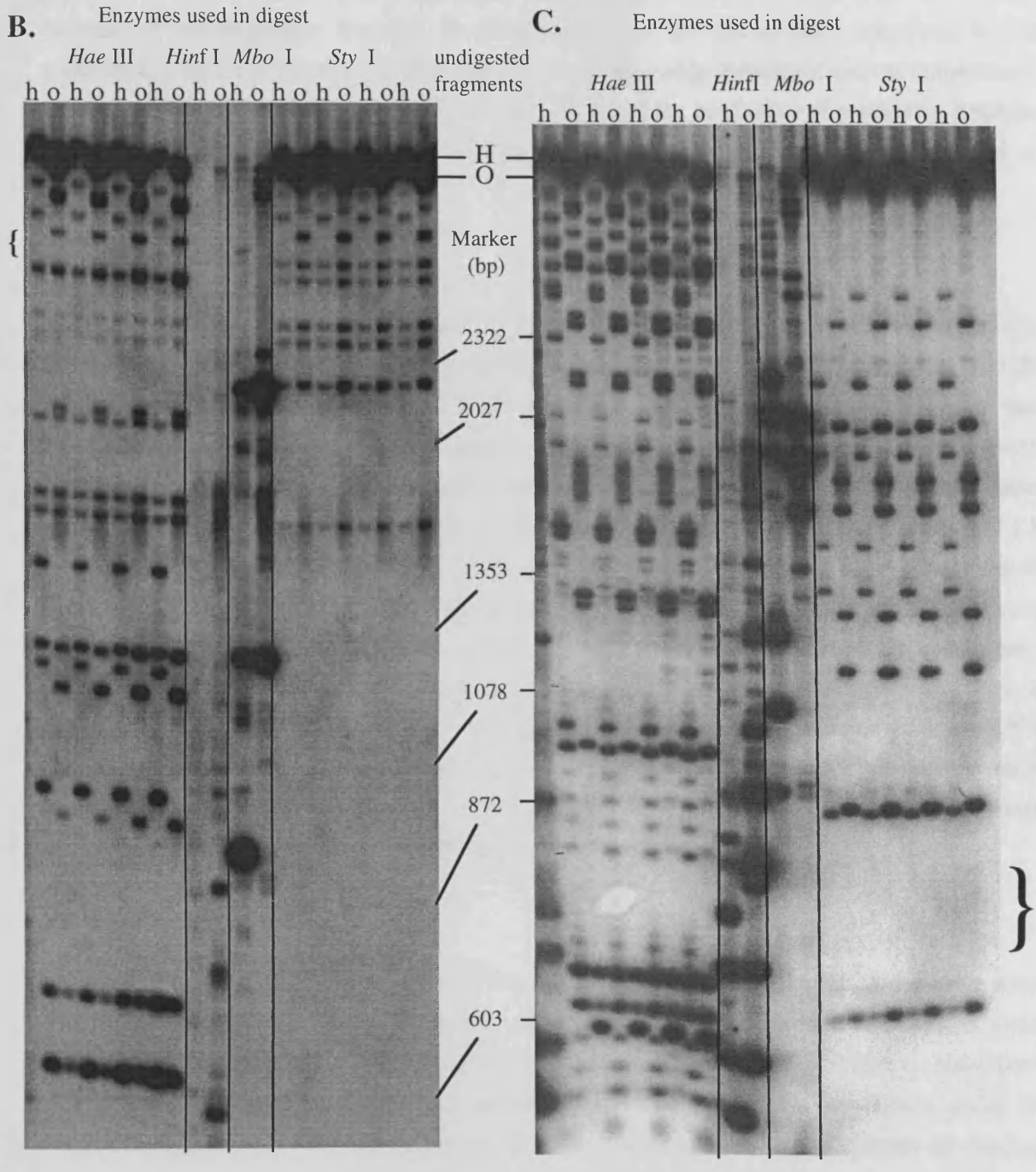
To detect the location of the 200 bp difference between the human and orang-utan sequences, this region was amplified between primers C1+2.4NFnew and C1+6.8NRnew in both species. These regions were then partially digested by incubating the amplified fragments with different restriction enzymes for a limited time. The digested products were then separated by electrophoresis and Southern hybridised to a ^{32}P-labelled end probe to detect a series of different size fragments which extend from one end of the sequence (B). Bands shared between the sequences of the two species indicate sequence similarity, and break-down of this band sharing defines the region of difference between the two sequences, in this case the smaller size of the orang-utan sequence. These fragments can then be probed using the end probe from the opposite end of the region to obtain an inverse pattern of partially digested fragments (C). However, sequence divergence (see Figure 6.11) between the two species can complicate the interpretation of these results.

A.



A. Schematic diagram explaining this technique.

The human *C1NH* sequence is shown with a number of hypothetical restriction enzyme sites (vertical lines above the sequence). Thick grey arrows represent the position and orientation of Alu elements, and the chequered box represents exon 4 of the *C1NH* sequence. A series of hypothetical fragments visualised following partial digestion and detection with probe 1 (thick line) are represented as horizontal lines below the sequence. The approximate region not present in orang-utan is shown by dashed lines, and includes Alu 2 (labelled) of the human *C1NH*. Changes in fragment size detected by probe 1 due to this "missing region" in orang-utan DNA are indicated by thin arrows.



B. Detection of partial fragments obtained by Southern hybridisation using 32P labelled probe 2 (3′ end). Human (h) and orang-utan (o) digests have been loaded in alternate lanes along the gel. The undigested human (H) and orang-utan (O) fragments can be seen at the top of the gel and the size difference is clear. The approximate region where the difference between the human and orang-utan sequences ends and the shared bands start is shown by a bracket to indicate the start of the differences between the two sequences.

C. Detection of partial fragments obtained by stripping and re-probing the same membrane as in B with ^{32}P labelled probe 1 (5′ end). Human (h) and orang-utan (o) sequences are therefore in the same order along the gel and the undigested fragments (H and O) are indicated. The approximate region where the similarity between the human and orang-utan sequences ends and the shared bands start, is shown by a bracket to indicate the start of the differences between the two sequences.

be between 8.5 x 10^{-5} and 1.7 x 10^{-6} per individual (F_{HAE} x 0.85 x 0.2) This may be an under-estimate of rearrangement because, as demonstrated by the above data, selectively neutral variation can occur in this region. This estimation will also only include *de novo* rearrangements which manifest as the HAE disease phenotype. To avoid the confusion of predicting genotype using phenotype, the structure and stability of this locus should be directly examined in humans.

Population screening

Human population screening was carried out to investigate the maximum level of heterozygosity at this locus. A total of 51 unrelated individuals, including 6 African, 5 Japanese and 39 Caucasian individuals were successfully amplified using primers C1+2.4NFnew and C1+6.8NRnew. All individuals were monomorphic for size in this region (data not shown). Applying Kimura's theory of neutral drift as for MS32, the maximum mutation rate at this locus was calculated to be 1.0 x 10^{-6} per allele in Caucasians, 1.3 x 10^{-6} per allele in Africans and 1.9 x 10^{-6} per allele in the Japanese. The different estimates of rearrangement in these populations is likely to reflect the numbers of alleles screened in each population, rather than actual differences in mutation rate. These mutation rates correlate well with those estimated by evolutionary analysis and from extrapolation of HAE disease frequencies. This again suggests that there is strong selection against mutation in this region, implying that rearrangements in the DNA immediately flanking exon 4 may not be selectively neutral. The final method that can be used to investigate mutation at this locus is by direct analysis of germline DNA which is not possible with oocytes but can be done with sperm DNA.

Small pool PCR

The estimates of mutation rate determined by population and evolutionary analysis may be gross underestimates of the true germline mutation rate at this locus. To determine whether this locus is unstable in sperm but for some reason these mutants do not survive in pedigrees, small pool PCR (SP-PCR) was carried out using sperm DNA. It was hoped that if mutants could be detected and analysed by SP-PCR, size enrichment could be avoided as a technically demanding and time consuming technique. SP-PCR can detect mutants occurring at a frequency of 10^{-6} and above (Jeffreys *et al.*, 1997), and the mutation estimates for this locus are within this range. The potential problem with this technique is jumping PCR between incompletely extended progenitor molecules, which yields PCR products indistinguishable from mutants. To minimise this risk DNA was amplified using the most efficient primer pair, C1+2.4F and C1+6.8NR, with herring sperm DNA (5 μg/ml) included as carrier. Blood DNA was also analysed as a negative control. 16 reactions each containing different dilutions of sperm (500 amplifiable molecules) and blood DNA (600 amplifiable molecules) were analysed for mutants by PCR and Southern hybridisation. This revealed 4 and 6 abnormal length molecules in blood and sperm respectively (Figure 6.13). This indicated that the frequency of gross genomic deletion in this region is 10^{-3} per progenitor molecule, about three orders of magnitude higher than previous estimates. However, some or all of these could be PCR artefacts, particularly as abnormal

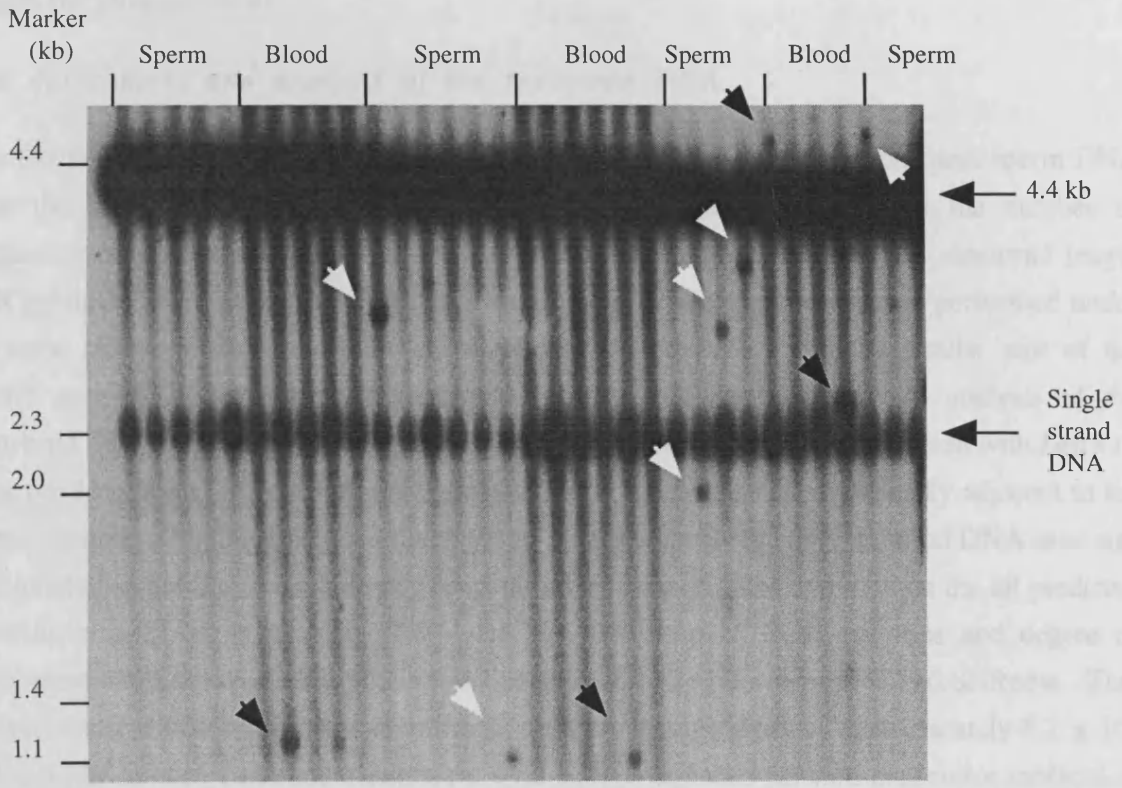


Figure 6.13. Small pool (SP-) PCR of the *C1NH* region.

Dra I digested blood and sperm DNA was amplified with primers C1+2.4F and C1+6.8NR to give a 4.4 kb fragment (arrow). 16 reactions each containing 500 amplifiable molecules of sperm DNA and 12 reactions each containing 600 amplifiable molecules of blood DNA were loaded alternately on the gel. The diffuse bands which migrate between 2.0 kb and 1.3 kb represent single stranded DNA (arrow). 4 putative blood mutants (black) and 6 putative sperm mutants (white) are highlighted with arrows.

length PCR products were also, rather unexpectedly observed in blood DNA. Pilot experiments had shown that abnormal size PCR products appeared at increasing frequency with increasing amounts of input DNA, suggesting that these were authentic mutants. However, the high concentration of input DNA and the large amount of single-stranded DNA produced indicated that these abnormal length products are likely to be PCR artefacts. The only way to determine whether these are true mutants was to remove the progenitor molecules which provide potential targets for jumping PCR.

Size enrichment and analysis of the recovered DNA

Size fractionation and single molecule PCR was carried out on genomic blood and sperm DNA from the same sperm donor used for SP-PCR. This technique depletes the number of progenitor molecules in the starting DNA, in order to determine whether the abnormal length PCR products seen during SP-PCR were true mutants. Size enrichment was performed under the same stringent single-molecule clean conditions as at MS32 and the similar size of the *C1NH* and MS32 progenitor fragments also meant that fractionation and analysis of the recovered DNA was identical. The exception was that genomic DNA was digested with *Dra* I (2 units per 1 μg input DNA in the appropriate buffer), which cleaves immediately adjacent to the outside primers (Figure 6.2B). A total of 20 μg of this digested sperm and blood DNA was size fractionated as before, to exclude any progenitor fragment (5.1 kb) and include the all predicted Alu-Alu recombination products (1.9 - 4.1 kb). Inhibition of PCR enzymes and degree of enrichment were assayed using the nested primers, C1+2.4NFnew and C1+6.8NRnew. This showed that fraction recovery was uniform, with an average yield of approximately 8.2 x 10^6 molecules/μl in each fraction, which were over 99.7% enriched for non-progenitor molecules. Unfortunately, the quality of the recovered DNA was not tested, though previous use of the technique suggests that genomic DNA is fairly robust.

Screening of the fractionated DNA

As for MS32, several screens of the fractionated DNA were carried out to detect deletion mutants. Up to the equivalent of 5 x 10^5 amplifiable progenitor molecules were screened in a single 7 μl PCR reaction. DNA was generally screened in pools which contained approximately the same concentrations of DNA from 3 fractions. Fractions only partially depleted in progenitor DNA were amplified separately at lower DNA concentrations in order to avoid swamping any signal from potential mutants with progenitor signal. Amplification of the fractionated DNA was carried out using the outside primers, C1+2.4F and C1+6.9R, and 5 μg/ml herring sperm as carrier. A positive control consisting of 10 ng unfractionated, digested starting DNA was included in each screen performed. Half the volume of each PCR reaction was analysed by electrophoresis and Southern hybridisation with ^{32}P-labelled probe 1 or 2, the remainder was stored at -20°C for subsequent analysis. The equivalent of 4.3 million molecules of fractionated sperm DNA were screened revealing no deletion mutants, so the rate of deletion at this locus is < 4.3 x 10^{-6}. This estimate is similar to those obtained by evolutionary analysis and human population screening, and indicates that the abnormal length molecules observed

following SP-PCR were PCR artefacts. Therefore the true rate of rearrangement at this locus is below the level which is detectable using the most sensitive systems available at this time.

Analysis of somatic mutation at this locus

Somatic mutation was investigated at this locus by the fractionation and analysis of blood DNA from the same sperm donor used for both the *C1NH* and the MS32 work. Screening of the equivalent of 5 million molecules detected no deletion mutants, so the rate of somatic deletion at this locus can be estimated at less than 5×10^{-6}. The rate of rearrangement in blood was expected to be low because Alu-mediated mutation is thought to be largely restricted to the germline. This adds further weight to the argument that the abnormal length PCR products observed following SP-PCR analysis on blood and sperm DNA were PCR artefacts.

Discussion

During this work a system was developed for identifying rare *de novo* deletion events at two Alu-rich loci using physical size-selection and PCR recovery of single molecules. The first target locus was a 4.3 kb fragment of sequence flanking the hypervariable minisatellite, MS32; the second was the region around exon 4 of the *C1NH* gene. Both loci have previously been associated with instability. MS32 is one of the most well characterised hypervariable minisatellites in the human genome with a male germline mutation rate to new length alleles of around 0.8% (Jeffreys *et al.*, 1994), and Alu-mediated rearrangements have previously been characterised in HAE patients at the *C1NH* gene (Stoppa-Lyonnet *et al.*, 1987, 1990, 1991; Ariga *et al.*, 1990). This system was shown to be suitable for detecting single deletion mutant molecules by amplification and detection of mutant bands in blood DNA obtained from HAE patients. The equivalent of 10 million sperm and 5 million blood molecules were screened for deletion mutants in the region flanking MS32; and the equivalent of 5 million blood and 4.8 million sperm molecules were analysed for deletion mutants in the *C1NH* gene. No deletion events were detected for either locus, so the rate of gross ($\geq$ 200 bp) deletion in the germline is less than 3×10^{-7} at MS32 and less than 4.3×10^{-6} at the C1 inhibitor gene. The maximum rate of somatic mutation at these loci is 5×10^{-6} at MS32 and 5×10^{-6} at the C1 inhibitor gene. This showed that these regions are stable in the soma and the germline.

The stability of the Alu-rich region associated with MS32 may be explained by recombination studies that have been carried out at this locus. This work by Jeffreys *et al.* (1998b) identified a recombination hotspot centred approximately 200 bp 5′ of MS32. Further upstream, recombination was found to be reduced at least 100-fold relative to the hotspot, and at least 2-fold relative to the average genomic rate. This suppression extends approximately 1 kb into the Alu-rich region investigated here, and may extend further which would explain the stability observed in the present study. The suppression of recombination outside the hotspot may also affect retrotransposition which would explain the paucity of younger Alu subfamilies flanking the minisatellite. If this is true then the age of the hotspot could be determined from the time of

origin of the most recent Alu subfamily, assuming that suppression of recombination is directly due to the influence of the hotspot. Alu Sx was the most recent insertion directly into target region investigated here, with Alu Sb2 a more recent insertion further upstream (Figure 6.2). These subfamilies arose 30 and 1.2 million years ago, respectively (Kapitonov & Jurka, 1996). Therefore, depending on how far upstream the influence of the hotspot extends, the maximum age of the hotspot is either 30 or 1.2 million years. In addition, because MS32 is monomorphic in great apes (Gray & Jeffreys, 1991) and is therefore unlikely to be associated with a hotspot in these primates, the upper age limit of this hotspot is equivalent to the time of divergence of humans from chimpanzee and gorilla, approximately 7 million years ago (Koop *et al.*, 1986).

The lack of rearrangement seen within the C1 inhibitor gene may be more difficult to interpret. The most likely explanation is that there is selection against recombination at this locus, which affects both disruptive and non-disruptive events. It is also important to remember that it is currently impossible to determine the parental origin of the *de novo* rearrangements at the C1 inhibitor gene in HAE patients. Germ cell mutation can show a sex bias in many different repeat types (Jeffreys *et al.*, 1997; Nelson, 1993), so it is possible that mutation at this locus is exclusively maternal. Unfortunately, large amounts of oocyte DNA are not available to be analysed in the same way as sperm DNA.

This work has demonstrated that the most sensitive detection system, size-enrichment and single molecule PCR available at the present time is unsuitable for the isolation of deletion mutants occurring below a rate of 10^{-8}. Furthermore, this technique is much less suitable for the detection and isolation of gain mutants. This is due to "streak back" of progenitor molecules during electrophoresis so that it is difficult to separate these mutants from progenitor molecules.

These two major problems must be solved before the mechanisms of low level mutation in the genome can be studied in more detail. The first possibility would be to use this system on a locus with a higher deletion rate. However, these two loci were chosen because of their accessibility and suitability for this study. Alternative targets, for example the LDL receptor gene (Lehrman *et al.*, 1986, 1987a & b) where rearrangements may have a much higher population frequency were impractical because, in this case mutation is spread over a very large (55 kb) region. These difficulties demonstrate the need for a more powerful and flexible system of mutation detection at the single molecule level. Until such methods are discovered the analysis of mutation in disease, for example HAE patients, will provide the only means of accessing low level mutation in the human genome.

Chapter 7

Discussion

The ultimate aim of this work was to achieve a better understanding of instability in the human genome at two different repeat types, Alu elements and minisatellites. Both have a well documented history of instability in the human genome and both are amenable to PCR analysis. One area of these studies which clearly requires attention is to improve techniques used for the examination of events which can be rare, such as the Alu deletion work, or events hidden by conventional detection methods such as the isometric crossovers at MS31a.

Population studies using minisatellites

Population studies on minisatellites and their immediate surroundings have been used to indicate new areas of mutation research. Pedigree studies first hinted at the high levels of variability which are displaced towards one end of the hypervariable minisatellites, and this opened up the exciting field of tandem repeat biology. More recently, population studies that have been carried out to make predictions about mutation at these loci have focused on the sequences flanking the minisatellite (Jeffreys *et al.*, 1998b; Chapter 4). Superficially, these appeared to be normal, non-coding regions of DNA, often rich in repeat sequences and SNPs. The SNPs were initially used to design allele-specific primers to amplify single alleles and create single-allele MVR maps. Subsequently, more detailed, comparative analysis of these polymorphic sites in different populations have hinted at high levels of recombination in the flanking DNA. Fluctuations in levels of linkage disequilibrium between these sites were used to predict the presence of a putative recombination hotspot in the flanking DNA. However, this was often masked in non-European populations, presumably because of the small sample sizes analysed. The region of lowest linkage disequilibrium and highest recombination was located immediately upstream of MS31a, and resembled the pattern of linkage disequilibrium in the 5′ flanking DNA adjacent to MS32 (Jeffreys *et al.*, 1998b). These flanking sites also provided the key to the isolation and characterisation of crossover molecules from sperm DNA using pairs of allele-specific primers.

Little is known about the stable end of minisatellites, and it has been generally assumed that this end of the array will show low levels of variability corresponding to the lack of mutation events isolated in these regions. However, design of a reverse MVR mapping system and the identification of flanking variant sites at the 3′ (stable) end of MS31a showed that this was not the case. In fact, population studies have predicated high levels of recombination in the flanking DNA at both ends of the minisatellite, and if recombination is the driving force behind minisatellite instability this suggests that mutation at MS31a may be bipolar. Studies of mutants identified in pedigrees suggest that this instability will probably be slightly higher in the 5′ flanking DNA than in the 3′ flanking regions. In contrast, a similar study of the less variable

end of MS205 showed a very limited repertoire of MVR maps. These were often strongly associated with distinct flanking haplotypes suggesting that mutation is highly polarised towards one end of this minisatellite (C. A. May, pers. comm.). Additional studies of different minisatellite loci will be required to establish whether minisatellites are generally located between two recombination hotspots or whether MS31a is an exceptional example.

Extensive population studies will also be necessary to investigate whether minisatellites experience some sort of lifecycle. The theory behind this is that a hypervariable minisatellite evolves from a highly stable, monomorphic state by rapid stochastic expansion in repeat numbers, and that its existence in this hypervariable state has a somewhat stochastically defined lifespan (Gray & Jeffreys, 1991). This idea originates from the observation that there are substantial numbers of monomorphic and stable minisatellites (e.g. MS31b) in the human genome, and because minisatellites that are hypervariable in humans are often monomorphic and highly stable in other primates (Gray & Jeffreys, 1991). The latest studies on minisatellite mutation suggest that instability, and therefore the lifespan of the minisatellite is determined by the presence of a crossover hotspot, although other factors may also be important. For example, with the bias towards a gain of repeats at minisatellites, the seven million year divergence time (Koop *et al.*, 1986) since humans and great apes is long enough for MS32 alleles 7,000 repeats long to have been generated, depending on exactly when these human hypervariable minisatellites arose. However, the upper limit of array size for both MS32 and MS31a seems to be about 600 repeats implying that indefinite expansion of these loci may be prevented. This may happen by rare compensatory deletions, possibly in the female germline (Neil, 1994); by truncating selection on long arrays (Caskey *et al.*, 1992; Orr *et al.*, 1993, Huntingdon's Disease Collaborative Research Group, 1993); or by mutational silencers, such as the flanking variants discussed below, that sweep through a population by meiotic drive. These effects may occasionally be catastrophic, reducing the minisatellite to a single, presumably stable repeat, and thereby ending the lifecycle of the hypervariable minisatellite. The combined studies of population genetics and instability at and around minisatellites will hopefully be able to determine how these hypervariable repeats have evolved in the human genome.

Investigation into the size and structure of minisatellite alleles in different populations is essential to identify both very unusual and very common alleles, which may reflect underlying mutation processes. For example, minisatellite alleles which have spread to high population frequency may be associated with mutational silencers which, because they do not mutate can spread quickly through an interbreeding population. Previously, a subset of short MS31a alleles with very similar MVR maps which show little evidence of variability has been reported at relatively high frequency in the Japanese and Zimbabwean populations (Huang *et al.*, 1996). The comparative study of allele sizes in the different populations carried out here (Chapter 3) suggests that these short, stable alleles are likely to be present in the Japanese and African populations at high frequency, but are negligible in the Afro-Caribbean and Caucasian populations investigated. These alleles may be similar to the short stable alleles identified at CEB1 and B6.7 (Buard *et al.*, 1998; Tamaki *et al.*, 1999), or they may be associated with a

mutational silencer. At MS32, such alleles are associated with flanking variants which suppress minisatellite mutation in *cis* (Monckton *et al.*, 1994). This so-called O1C site, disrupts a mirror palindrome which may represent a binding site for a protein necessary for the initiation of recombination (Monckton *et al.*, 1994). This function cannot be ubiquitous however, because conserved features could not be identified in either the primary or secondary sequence of the flanking DNA of hypervariable minisatellites (Murray *et al.*, 1999). Evidence has demonstrated that this suppression is not associated with array length because alleles as short as 19 repeats which mutate "normally" have been identified. There is also evidence from other minisatellite loci that short, stable alleles may be associated with flanking variants (Andreassen *et al.*, 1996; May *et al.*, 1996) although it is not clear whether this stability is an allele length effect as shown at CEB1 and B6.7 (Buard *et al.*, 1998; Tamaki *et al.*, 1999). Recombination in yeast can also be suppressed by flanking variants which disrupt protein binding sequences, suggesting the processes of recombination in humans and yeast may be analogous (Szostak *et al.*, 1983), which could be invaluable in further studies of human meiotic recombination.

Population studies remain essential for the identification of potential new areas of research which allow tandem repeats to be examined in a wider context. There is also information that can only be derived from population work in the first instance, that can be used to direct future mutation studies, as discussed. However, the more detailed analysis of instability processes at and around minisatellites must necessarily be performed directly on germline DNA.

Minisatellite mutation and recombination

From the detailed characterisation of mutants isolated from many different minisatellites, a general picture of germline mutation at minisatellites is emerging. All minisatellites appear to mutate in sperm simultaneously by intra- and inter-allelic processes. Intra-allelic duplications and deletions generally form a background of low-level mutation, reminiscent of that seen in the soma. This type of mutation is probably mitotic, most likely occurring by intra-allelic recombination, such as unequal sister chromatid exchange. Conversely, inter-allelic exchanges in the male germline are dominated by polarised gene conversion-like events with a bias towards gain of repeats. Minisatellites CEB1 and MS205 are unusual in that the rate of intra-allelic mutation appears to be elevated above that of inter-allelic exchange at these loci. However, this may be due to the difficulties in identifying the allele of origin of transferred repeats at these loci. Differences in mutation profile between minisatellites are probably influenced by factors such as the immediate flanking sequence, the genomic location, the proximity of functional components of the chromosome for example active genes, and methylation status. These will be manifested as variations in repeat sequence, length and copy number and in allele size ranges of different minisatellites. Comparison of these differences and similarities will permit the rules that govern minisatellite instability to be identified.

Recent evidence suggests that the polar instability of minisatellites is driven by a crossover hotspot which is at its most intense immediately adjacent to the most variable end of the minisatellite (see Chapter 5; Jeffreys *et al.*, 1998a & b; J. Buard pers. comm.). Obviously, this

hotspot has not yet been demonstrated in all hypervariable minisatellites studied, but it seems likely that it will be. At MS32, O1C alleles have shown co-suppression of both meiotic crossover and conversion, suggesting that both processes originate from the same initiation complex (Jeffreys *et al.*, 1998b). In fact, it has long been suggested that gene conversion events are the result of aborted meiotic recombination events, and now this confirmation has allowed the model of minisatellite mutation to be updated. In this model, mutation is initiated by the introduction of a double strand break or staggered nicks in the DNA. Some of these occur within this restricted hotspot region of the minisatellite flanking DNA and are immediately resolved. The majority of breaks, however initiate within this region and migrate into the minisatellite array where they are aborted and resolved as gene-conversion events. Firm evidence of breakpoint migration has been seen at MS31a (recombinant D, Chapter 5) and MS32 (Jeffreys *et al.*, 1998b). Judging from the profile of the crossover hotspot, breakpoints are prevented from migrating too far, possibly by a generic defence mechanism which avoids excess disruption of the genome. This mechanism may bias branch migration, which preferentially drives these unstable recombination complexes into the repeat array. Here, they are prevented from further migration by the abortion of inter-allelic crossover into gene conversion events. These events may be mediated by the mismatch repair systems which can identify and abort recombination between heterologous repeats (Rayssiguier *et al.*, 1989; Borts *et al.*, 1990). In yeast it has been demonstrated that an increase in sequence divergence between alleles reduces meiotic recombination (Borts and Haber, 1987; Borts *et al.*, 1990). Contrary to this however, at MS32 the degree of mismatch between alleles at the point of crossover does not affect the efficiency of recombination (Jeffreys *et al.*, 1998b). Alternatively, it may be the repetitive nature of the minisatellite array which promotes resolution of these intermediate stages as gene conversion events. This model is likely to invite more questions than it is able to answer, but hopefully with the further study of tandem repeat biology, in time these questions will be addressed.

Now that the link between minisatellites and recombination has been proven, speculation has been revived about the role of minisatellites in chromosome synapsis. It has long been thought that minisatellites mark chromosomal sites actively involved in the homology searches that are essential for homologous chromosome pairing during meiosis (Carpenter, 1987). These searches are thought to involve strand exchange and invasion between duplexes to allow homologous regions to be sensed prior to chromosome pairing. This process of strand exchange and invasion provides substrates for the initiation of meiotic recombination and, consequently gene-conversion. The link between minisatellites and chromosome synapsis is reinforced by the fact that the majority of minisatellites cluster near the ends of chromosomes (Royle *et al.*, 1988) in regions that are involved in initiating meiotic chromosome pairing. Identifying a ubiquitous role for these elements may explain why hypervariable minisatellites show such uniform and intense mutational activity which, in some cases, appears to be regulated by flanking variants (e.g. O1C), and to a certain extent by size. Such a role may also explain why minisatellites have been so exquisitely preserved from normal genomic activity which tends to break up regions of homology, either by fragmenting regions of similar

sequences around the genome, or by the introduction of base substitutional polymorphisms, and transposable elements (Radman *et al.*, 1993). At minisatellites even the insidious creep of base substitution appears to be suppressed within the repeat array, for example at MS31a there are only two variant bases within the 20 bp repeat unit throughout the entire array. Minisatellites have therefore revealed themselves to be useful not only for the study of mechanisms of tandem repeat turnover, but also promise to reveal further details about complex meiotic processes.

Mutation at dispersed repeats

A substantial number of deleterious mutations which give rise to human genetic disease are the result of gross genomic recombination events between dispersed repeats, particularly Alu repeats. These represent a class of relatively low frequency mutations, about which very little is known. The most striking feature of this class is that mutation is not as random as might first appear (as discussed in Chapter 6). Recombination appears to be targeted to particular regions of particular repeats, and some repeat rich regions also seem to be more prone to deleterious recombination events than others. The explanation for this is unknown, but may depend on the efficacy of genome wide mechanisms which suppress unwanted homologous recombination, for example by genomic imprinting (Klar, 1998), or by reducing homology between elements through the introduction of point mutations (Radman *et al.*, 1993). The aim of this part of the project was to isolate and characterise mutant molecules directly from the male germline and, by using information also available from the molecular analysis of disease, understand more about this class of mutation. Unfortunately, the two regions chosen for this study were not ideal, given the suppression of recombination subsequently demonstrated at MS32, and the fact that mutation at the C1 inhibitor gene locus is probably too rare to detect by the methods used. Hopefully, the analysis of low level mutation will be possible in the future, perhaps using more suitable regions revealed by the human genome sequencing project, or by the application of more sensitive techniques, if and when they become available.

Future directions

Previously, recombination hotspots could only be provisionally defined by indirect methods such as linkage disequilibrium studies (Chapter 4; Chakravarti *et al.*, 1984; Charmley *et al.*, 1990; van Endert *et al.*, 1992; Jorde *et al.*, 1993), by high resolution physical mapping of recombinants detected in pedigrees (Bowcock *et al.*, 1987; Grimm *et al.*, 1989; Cullen *et al.*, 1995), or by linkage analysis in single sperm (Hubert *et al.*, 1994). These methods are not sensitive enough to predict the intensity and extent of recombination hotspots, so virtually nothing is known of the fine scale distribution of meiotic crossovers in human chromosomes. However, the techniques recently developed by Jeffreys *et al.* (1998b), and applied here at MS31a allow the high-resolution mapping of crossovers in human sperm permitting the detailed delineation of a recombination hotspot. This ability to map recombination hotspots at the molecular level has opened up a host of opportunities for the investigation of recombination. Known recombination hotspots can now be defined in great detail and it will be interesting to determine whether such intense clustering of crossover events is a general feature of human

meiotic recombination, as it is in yeast (Baudat & Nicolas, 1997), or whether this localised hotspot activity is restricted to the flanking DNA of minisatellites. The only limitation to this work is that potential hotspot regions have to be intensively characterised to identify a number of different SNPs within the region of interest.

Immediate prospects for research at MS31a include the isolation and characterisation of recombinant molecules derived from a number of different alleles. Before this can be achieved however, more advanced strategies are required for the isolation of recombinant molecules from bulk germline DNA. This requires reducing the labour intensive process of size enrichment as much as possible, both to make this selection process more efficient and to reduce the demand on primary materials. This is particularly important with respect to precious stocks of sperm DNA, and may permit the analysis of limited stocks of sperm DNA, for example from different populations. Improved methods for breakpoint mapping are also being developed at MS31a using labelled ASOs to determine the allelic state of multiple recombinants in a single hybridisation reaction (Chapter 5). This will greatly increase the throughput of this analysis, so that multiple recombinants can be analysed in a semi-automated manner.

Very few alleles of MS31a and MS32, have been examined for recombination activity using these techniques. Furthermore, these methods have yet to be attempted at other minisatellites, although analysis of recombination at CEB1 is underway (J. Buard, pers. comm.). It will be interesting to discover if and how these recombination profiles vary between minisatellites, for example it has already been demonstrated that CEB1 alleles can undergo flanking co-conversion of the minisatellite array giving rise to length change mutants, which has not been observed at MS32 (Jeffreys *et al.*, 1998a & b) or MS31a (Chapter 5). The ability to define recombination hotspots in such detail also permits further investigation into the link between minisatellites, recombination, and meiotic events such as chromosome homologue recognition. Comparisons can be made between minisatellites located in the subterminal regions of chromosomes, which are widely associated with such meiotic phenomena, and interstitial minisatellites such as MS32. This may represent a requirement for interstitial points of homologue recognition during meiosis, which has been hitherto unseen in humans, though is well documented in mouse (Carpenter, 1987), and may provide further evidence for a conserved role of minisatellites in mammals. Minisatellites in human, rat and pig all have a subterminal origin (Royle *et al.*, 1988; Amarger *et al.*, 1998) which, in humans is involved in homologous chromosome pairing and synapsis during meiosis. In mouse there also appears to be an association between the distribution of minisatellites and regions involved in pairing of homologous chromosomes during meiosis (Bois *et al.*, 1998). Recombination studies could be invaluable for the identification of proteins involved in this process, and may ultimately even throw some light on the black box that is meiosis.

Finally, it must be remembered that minisatellite mutation is not the only type of instability in the human genome. There are numerous other mutational processes, although the mechanisms and rates governing these events are poorly understood, mainly because of the relative scarcity of such events in man. Previously Monckton (1992) attempted to use PCR based strategies to

measure the rate of *de novo* transposition at a defined locus. In this study, attempts were made to isolate rare deletion mutants from two loci using the incredibly sensitive method of size enrichment coupled with PCR. Also, studies to isolate and characterise mutants at the compound microsatellite wg1c4 have been carried out (C. A. May, pers. comm.). However, none of the studies so far carried out to investigate regions of reduced mutation frequency have successfully identified true mutants. Therefore these rare mutants will probably remain inaccessible until more sensitive detection methods are developed, unless they can be examined in patients with genetic disease (e.g. Reiter *et al.*, 1998; La Spada *et al.*, 1994; Schwartz *et al.*, 1998; Chapter 6) or in the soma of cancer patients (e.g. Jeffs *et al.*, 1998; Strout *et al.*, 1998).

References

Adams RLP, Davies T, Rinaldi A, Easton R (1987) CpG deficiency, dinucleotide distribution and nucleosome positioning. European Journal of Biochemistry 165: 107-115

Agostoni A, Cicardi M (1992) Hereditary and acquired C1-inhibitor deficiency - biological and clinical characteristics in 235 patients. Medicine 71: 206-215

Akgun E, Zahn J, Baumes S, Brown G, Liang F, Romanienko PJ, Lewis S, Jasin M (1997) Palindrome resolution and recombination in the mammalian germ line. Molecular and Cellular Biology 17: 5559-5570

Albertini AM, Hofer M, Calos MP, Miller JH (1982) On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell 29: 319-328

Amarger V, Gauguier D, Yerle M, Apiou F, Pinton P, Giraudeau F, Monfouilloux S, Lathrop M, Dutrillaux B, Buard J, Vergnaud G (1998) Analysis of distribution in the human, pig, and rat genomes points toward a general subtelomeric origin of minisatellite structures. Genomics 52: 62-71

Andreassen R, Egeland T, Olaisen B (1996) Mutation-rate in the hypervariable VNTR λG3 (D7S22) is affected by allele length and a flanking DNA-sequence polymorphism near the repeat array. American Journal Of Human Genetics 59: 360-367

Anthoney DA, McIlwrath AJ, Gallagher WM, Edlin ARM, Brown R (1996) Microsatellite instability, apoptosis, and loss of p53 function in drug-resistant tumor cells. Cancer Research 56: 1374-1381

Arcot SS, Wang ZY, Weber JL, Deininger PL, Batzer MA (1995) Alu repeats - a source for the genesis of primate microsatellites. Genomics 29: 136-144

Ariga T, Carter PE, Davis AE (1990) Recombinations between Alu repeat sequences that result in partial deletions within the C1-inhibitor gene. Genomics 8: 607-613

Armour JAL, Patel I, Thein SL, Fey MF, Jeffreys AJ (1989a) Analysis of somatic mutations at human minisatellite loci in tumors and cell-lines. Genomics 4: 328-334

Armour JAL, Wong Z, Wilson V, Royle NJ, Jeffreys AJ (1989b) Sequences flanking the repeat arrays of human minisatellites - association with tandem and dispersed repeat elements. Nucleic Acids Research 17: 4925-4935

Armour JAL, Povey S, Jeremiah S, Jeffreys AJ (1990) Systematic cloning of human minisatellites from ordered array charomid libraries. Genomics 8: 501-512

Armour JAL, Crosier M, Malcolm S, Chan JCT, Jeffreys AJ (1995) Human minisatellite loci composed of interspersed GGA-GGT triplet repeats. Proceedings Of the Royal Society Of London Series B - Biological Sciences 261: 345-349

Armour JAL, Crosier M, Jeffreys AJ (1996) Distribution of tandem repeat polymorphism within minisatellite MS621 (D5S110). Annals Of Human Genetics 60: 11-20

Ashley CT, Warren ST (1995) Trinucleotide repeat expansion and human disease. Annual Review of Genetics 29: 703-728

Auerbach C (1976) Mutation research: problems, results and perspectives. Chapman and Hall, London

Auge-Gouillou C, Bigot Y, Pollet N, Hamelin MH, Meunier-Rotival M, Periquet G (1995) Human and other mammalian genomes contain transposons of the mariner family. FEBS Letters 368: 541-546

Baird DM, Jeffreys AJ, Royle NJ (1995) Mechanisms underlying telomere repeat turnover, revealed by hypervariable variant repeat distribution patterns in the human Xp/Yp telomere. The EMBO Journal 14: 5433-5443

Barker D, Schafer M, White R (1984) Restriction sites containing CpG show a higher frequency of polymorphism in human DNA. Cell 36: 131-138

Barnes WM (1994) PCR amplification of up to 35 kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proceedings Of the National Academy Of Sciences Of the United States Of America 91: 2216-2220

Bartsocas CS (1984) Aristotle: the father of genetics. Philosophical Inquiry 4: 35-38

Bass HW, Marshall WF, Sedat JW, Agard DA, Cande WZ (1997) Telomeres cluster *de novo* before the initiation of synapsis; a 3 dimensional spatial analysis of telomere positions before and during meiotic prophase. Journal of Cellular Biology 137: 5-18

Bateson W (1894a) Mendel's principles of hereditary, Second Edition (1913) edn. Cambridge University Press, Cambridge

Bateson W (1894b) Materials for the study of variation treated with especial regard to discontinuity in the origin of species. MacMillan, London

Batzer MA, Arcot SS, Phinney JW, Alegriahartman M, Kass DH, Milligan SM, Kimpton C, Gill P, Hochmeister M, Ioannou PA, Herrera RJ, Boudreau DA, Scheer WD, Keats BJB,

Deininger PL, Stoneking M (1996) Genetic variation of recent Alu insertions in human populations. Journal Of Molecular Evolution 42: 22-29

Baudat F, Nicolas A (1997) Clustering of meiotic double-strand breaks on yeast chromosome III. Proceedings of the National Academy of Sciences of the United States of America 94: 5213-5218

Beckmann JS, Weber JL (1992) Survey of human and rat microsatellites. Genomics 12: 627-631

Beletskii A, Bhagwat AS (1998) Correlation between transcription and C to T mutations in the non-transcribed DNA strand. Biological Chemistry 379: 549-551

Bennett ST, Lucassen AM, Gough SCL, Powell EE, Undlien DE, Pritchard LE, Merriman ME, Kawaguchi Y, Dronsfield MJ, Pociot F, Nerup J, Bouzekri N, Cambonthomsen A, Ronningen KS, Barnett AH, Bain SC, Todd JA (1995) Susceptibility to human type 1 diabetes at *IDDM2* is determined by tandem repeat variation at the insulin gene minisatellite locus. Nature Genetics 9: 284-292

Benzer S, Freese E (1958) Induction of specific mutations with 5-bromouracil. Proceedings of the National Academy of Sciences Of the United States Of America 44: 112-119

Benzer S (1961) On the topography of the genetic fine structure. Proceedings of the National Academy of Sciences Of the United States Of America 47: 403-415

Berkvens TM, Vanormondt H, Gerritsen EJA, Khan PM, Vandereb AJ (1990) Identical 3250 bp deletion between two Alu repeats in the *ADA* genes of unrelated ADA-SCID Patients. Genomics 7: 486-490

Bestor TH, Walsh CP, Yoder JA (1997) Does DNA methylation control transposition of selfish elements in the germline? Reply. Trends in Genetics 13: 470-472

Bird A (1997) Does DNA methylation control transposition of selfish elements in the germline? Trends in Genetics 13: 469-470

Boeke JD (1997) LINEs and Alus - the polyA connection. Nature Genetics 16: 6-7

Boeke JD, Pickeral OK (1999) Genome structure - retroshuffling the genomic deck. Nature 398: 108-110

Bois P, Collick A, Brown J, Jeffreys AJ (1997) Human minisatellite MS32 (D1S8) displays somatic but not germline instability in transgenic mice. Human Molecular Genetics 6: 1565-1571

Bois P, Stead JDH, Bakshi S, Williamson J, Neumann R, Moghadaszadeh B, Jeffreys AJ (1998) Isolation and characterization of mouse minisatellites. Genomics 50: 317-330

Borts RH, Haber JE (1987) Meiotic recombination in yeast: alteration by multiple heterozygosities. Science 237: 1459-1465

Borts RH, Leung WY, Kramer W, Kramer B, Williamson M, Fogel S, Haber JE (1990) Mismatch repair-induced meiotic recombination requires the PMS1 gene product. Genetics 124: 573-584

Bovia F, Wolff N, Ryser S, Strub K (1997) The SRP9/14 subunit of the human signal recognition particle binds to a variety of Alu-like RNAs and with higher affinity than its mouse homolog. Nucleic Acids Research 25: 318-325

Bowcock AM, Hebert JM, Wijsman E, Gadi IK, Boyd C, Cavallisforza LL (1987) Linkage of markers at 13q34 - the pro alpha-1-(IV) and pro alpha-2-(IV) collagen genes and D13S3. Cytogenetics and Cell Genetics 46: 585

Boyer JC, Farber RA (1998) Mutation rate of a microsatellite sequence in normal human fibroblasts. Cancer Research 58: 3946-3949

Braman J, Barker D, Schumm J, Knowlton R, Doniskeller H (1985) Characterization of very highly polymorphic RFLP probes. Cytogenetics and Cell Genetics 40: 589

Brenner S, Benzer S, Barnett L (1958) Distribution of proflavin-induced mutations in the genetic fine structure. Nature 182: 983-985

Britten RJ, Baron WF, Stout DB, Davidson EH (1988) Sources and evolution of human Alu repeated sequences. Proceedings Of the National Academy Of Sciences Of the United States Of America 85: 4770-4774

Britten RJ (1994) Evolutionary selection against change in many Alu repeat sequences interspersed through primate genomes. Proceedings Of the National Academy Of Sciences Of the United States Of America 91: 5992-5996

Britten RJ (1996) DNA sequence insertion and evolutionary variation in gene regulation. Proceedings Of the National Academy Of Sciences Of the United States Of America 93: 9374-9377

Buard J, Vergnaud G (1994) Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). EMBO Journal 13: 3203-3210

Buard J, Jeffreys AJ (1997) Big, bad minisatellites. Nature Genetics 15: 327-328

Buard J, Bourdet A, Yardley J, Dubrova Y, Jeffreys AJ (1998) Influences of array size and homogeneity on minisatellite mutation. EMBO Journal 17: 3495-3502

Bulmer M, Wolfe KH, Sharp PM (1991) Synonymous nucleotide substitution rates in mammalian genes - implications for the molecular clock and the relationship of mammalian orders. Proceedings of the National Academy of Sciences of the United States of America 88: 5974-5978

Burwinkle B, Kilimann MW (1998) Unequal homologous recombination between LINE 1 elements as a mutational mechanism in human genetic disease. Journal Of Molecular Biology 277: 513-517

Campbell C, Marondel I, Montgomery K, Krauter K, Kucherlapati R (1995) Unequal homologous recombination of human DNA on a yeast artificial chromosome. Nucleic Acids Research 23: 3691-3695

Carpenter ATC (1987) Gene conversion, recombination nodules and the initiation of meiotic synapsis. Heredity 59: 307

Carter PE, Duponchel C, Tosi M, Fothergill JE (1991) Complete nucleotide sequence of the gene for human C1 inhibitor with an unusually high density of Alu elements. European Journal Of Biochemistry 197: 301-308

Cascalho M, Wong J, Steinburg C, Wabl M (1998) Mismatch repair co-opted by hypermutation. Science 279: 1207-1210

Caskey CT, Pizzuti A, Fu YH, Fenwick RG, Nelson DL (1992) Triplet repeat mutations in human disease. Science 256: 784-789

Centra M, Memeo E, dApolito M, Savino M, Ianzano L, Notarangelo A, Liu JM, Doggett NA, Zelante L, Savoia A (1998) Fine exon-intron structure of the Fanconi anemia group A (FAA) gene and characterization of two genomic deletions. Genomics 51: 463-467

Chae JJ, Park YB, Kim SH, Hong SS, Song GJ, Han KH, Namkoong Y, Kim HS, Lee CC (1997) Two partial deletion mutations involving the same Alu sequence within intron 8 of the LDL receptor gene in Korean patients with familial hypercholesterolemia. Human Genetics 99: 155-163

Chakravarti A, Phillips JA, Mellits KH, Buetow KH, Seeburg PH (1984) Patterns of polymorphism and linkage disequilibrium suggest independent irigins of the human growth-hormone gene-cluster. Proceedings of the National Academy of Sciences of the United States of America - Biological Sciences 81: 6085-6089

Chang DY, Nelson B, Bilyeu T, Hsu K, Darlington GJ, Maraia RJ (1994) A human Alu RNA-binding protein whose expression is associated with accumulation of small cytoplasmic Alu RNA. Molecular and Cellular Biology 14: 3949-3959

Charmley P, Chao A, Concannon P, Hood L, Gatti RA (1990) Haplotyping the human T-cell receptor β-chain gene complex by use of restriction fragment length polymorphisms.

Proceedings of the National Academy of Sciences of the United States of America 87: 4823-4827

Cheng S, Chang SY, Gravitt P, Respess R (1994) Long PCR. Nature 369: 684-685

Choo KHA (1997) The centromere. Oxford University Press, Oxford

Choo KHA (1998) Why is the centromere so cold? Genome Research 8: 81-82

Chu WM, Ballard R, Carpick BW, Williams BRG, Schmid CW (1998) Potential Alu function: regulation of the activity of double stranded RNA-activated kinase PKR. Molecular and Cellular Biology 18: 58-68

Chung MY, Ranum LP, Duvick LA, Servadio A, Zoghbi HY, Orr HT (1993) Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. Nature Genetics 5: 254-258

Ciotta C, Ceccotti S, Aquilina G, Humbert O, Palombo F, Jiricny J, Bignami M (1998) Increased somatic recombination in methylation tolerant human cells with defective DNA mismatch repair. Journal of Molecular Biology 276: 705-719

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide sequence variation in human lipoprotein lipase. American Journal of Human Genetics 63: 595-612

Clarke E (1997) The molecular structure of MS31 in primates. Third Year Project, University of Leicester

Clough JE, Foster JA, Barnett M, Wichman HA (1996) Computer simulation of transposable element evolution - random template and strict master models. Journal Of Molecular Evolution 42: 52-58

Collier S, Tassabehji M, Sinnott P, Strachan T (1994) A *de novo* pathological point mutation at the 21-hydroxylase locus: implications for gene conversion in the human genome. Nature Genetics 4: 101

Cooke HJ, Brown WRA, Rappold GA (1985) Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. Nature 317: 687-692

Cooper DN, Schmidtke J (1984) DNA restriction fragment length polymorphisms and heterozygosity in the human genome. Human Genetics 66: 1-16

Cooper DN, Smith BA, Cooke HJ, Niemann S, Schmidtke J (1985) An estimate of unique DNA sequence heterozygosity in the human genome. Human Genetics 69: 201-205

Cooper DN, Youssoufian H (1988) The CpG dinucleotide and human genetic-disease. Human Genetics 78: 151-155

Cooper DN, Krawczak M (1993) Human gene mutation. Bios Scientific Publishers, Oxford

Costa R, Peixoto AA, Barbujani G, Kyriacou CP (1992) A latitudinal cline in a *Drosophila* clock gene. Proceedings of the Royal Society of London Series B - Biological Sciences 250: 43-49

Crick FHC (1961) The genetic code. Proceedings of the Royal Society of Britain 167: 331-347

Cullen M, Erlich H, Klitz W, Carrington M (1995) Molecular mapping of a recombination hotspot located in the second intron of the human TAP2 locus. American Journal of Human Genetics 56: 1350-1358

Danieli GA, Mioni F, Muller CR, Vitiello L, Mostacciuolo ML, Grimm T (1993) Patterns of deletions of the dystrophin gene in different european populations. Human Genetics 91: 342-346

Daniell E, Roberts R, Abelson J (1972) Mutation in the lactose operon caused by bacteriophage Mu. Journal of Molecular Biology 69: 1-8

Darwin C (1859) On the origin of species by means of natural selection. John Murray, London
de Vries H (1901) Die mutationstheorie. [English translation: The mutation theory, Vol. I (1909), Vol. II (1910). Open Court, Chicago]

de Wind N, Dekker M, Berns A, Radman M, Riele HT (1995) Inactivation of the mouse MSH2 gene results in mismatch repair deficiency, methylation tolerance, hyperrecombination, and predisposition to cancer. Cell 82: 321-330

Deininger PL, Daniels GR (1986) The recent evolution of mammalian repetitive DNA elements. Trends In Genetics 2: 76-80

Dernburg AF, Sedat JW, Cande WZ, Bass HW (1995) Cytology of telomeres. In: Blackburn EH, Grieder CW (eds) Telomeres. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York

Dernburg AF, McDonald K, Moulder G, Barstead R, Dresser M, Villeneuve AM (1998) Meiotic recombination in *C. elegans* initiates by a conserved mechanism and is dispensable for homologous chromosome synapsis. Cell 94: 387-398

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES, Botstein D, Akots G, Rediker KS, Gravius T, Brown VA, Rising MB, Parker C, Powers JA, Watt DE, Kauffman ER, Bricker A, Phipps P, Mullerkahle H, Fulton TR, Ng S, Schumm JW, Braman JC, Knowlton RG, Barker DF,

Crooks SM, Lincoln SE, Daly MJ, Abrahamson J (1987) A genetic linkage map of the human genome. Cell 51: 319-337

Dulbecco R (1949) Reactivation of ultraviolet-inactivated bacteriophage by visible light. Nature 163: 949-950

Economou-Pachnis A, Tsichlis PN (1985) Insertion of an Alu SINE in the human homologue of the MLVI2 locus. Nucleic Acids Research 13: 8379-8387

Eikenboom JCJ, Vink T, Briet E, Sixma JJ, Reitsma PH (1994) Multiple substitutions in the von Willebrand factor gene that mimic the pseudogene sequence. Proceedings Of the National Academy Of Sciences Of the United States Of America 91: 2221-2224

Feinberg AP, Vogelstein B (1983) A technique for radiolabelling DNA restriction endonuclease fragements to high specific activity. Analytical Biochemistry 132: 6-13

Flint J, Rochette J, Craddock CF, Dode C, Vignes B, Horsley SW, Kearney L, Buckle VJ, Ayyub H, Higgs DR (1996) Chromosomal stabilization by a subtelomeric rearrangement involving two closely related Alu elements. Human Molecular Genetics 5: 1163-1169

Fujikawa K, Kamiya H, Kasai H (1998) The mutations induced by oxidatively damaged nucleotides, 5-formyl-dUTP and 5-hydroxy-dCTP, in *Escherichia coli*. Nucleic Acids Research 26: 4582-4587

Gilbertson LA, Stahl FW (1996) A test of the double strand break repair model for meiotic recombination in *Saccharomyces cerevisiae*. Genetics 144: 27-41

Goldberg YP, Rommens JM, Andrew SE, Hutchinson GB, Lin BY, Theilmann J, Graham R, Glaves ML, Starr E, McDonald H, Nasir J, Schappert K, Kalchman MA, Clarke LA, Hayden MR (1993) Identification of an Alu retrotransposition event in close proximity to a strong candidate gene for Huntingtons disease. Nature 362: 370-373

Goldschmidt R (1946) Position effect and the theory of the corpuscular gene. Experientia 2: 250-256

Gray IC, Jeffreys AJ (1991) Evolutionary transience of hypervariable minisatellites in man and the primates. Proceedings of the Royal Society of London Series B - Biological Sciences 243: 241-253

Green PM, Montandon AJ, Bentley DR, Ljung R, Nilsson IM, Giannelli F (1990) The incidence and distribution of CpG to TpG transitions in the coagulation factor IX gene: a fresh look at CpG mutational hotspots. Nucleic Acids Research 18: 3227-3231

Grimm T, Muller B, Dreier M, Kind E, Bettecken T, Meng G, Muller CR (1989) Hot spot of recombination within DXS164 in the Duchenne muscular dystrophy gene. American Journal of Human Genetics 45: 368-372

Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, Bernardi G, Lathrop M, Weissenbach J (1994) The 1993-4 Genethon human genetic linkage map. Nature Genetics 7: 246-339

Haber JE (1998) A locus control region regulates yeast recombination. Trends In Genetics 14: 317-321

Hambor JE, Mennone J, Coon ME, Hanke JH, Kavathas P (1993) Identification and characterization of an Alu containing, T-cell specific enhancer located in the last intron of the human CD8-alpha gene. Molecular and Cellular Biology 13: 7056-7070

Harteveld KL, Losekoot M, Fodde R, Giordano PC, Bernini LF (1997) The involvement of Alu repeats in recombination events at the α-globin gene cluster: characterization of two α(o)-thalassaemia deletion breakpoints. Human Genetics 99: 528-534

Hawley RS, Theurkauf WE (1993) Requiem for distributive segregation - achiasmate segregation in *Drosophila* females. Trends in Genetics 9: 310-317

Hawn MT, Umar A, Carethers JM, Marra G, Kunkel TA, Boland CR, Koi M (1995) Evidence for a connection between the mismatch repair system and the G(2) cell-cycle checkpoint. Cancer Research 55: 3721-3725

Heikkinen J, Hautala T, Kivirikko KI, Myllyla R (1994) Structure and expression of the human lysyl hydroxylase gene (PLOD) intron-9 and intron-16 contain Alu sequences at the sites of recombination in Ehlers-Danlos syndrome Type-VI patients. Genomics 24: 464-471

Hite JM, Eckert KA, Cheng KC (1996) Factors affecting fidelity of DNA synthesis during PCR amplification of d(C-A)(n)·d(G-T)(n) microsatellite repeats. Nucleic Acids Research 24: 2429-2434

Holmes SE, Dombrowski BA, Krebs CM, Boehm CD, Kazazian HHJ (1994) A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. Nature Genetics 7: 143-148

Hori T, Tomatsu S, Nakashima Y, Uchiyama A, Fukuda S, Sukegawa K, Shimozawa N, Suzuki Y, Kondo N, Horiuchi T, Ogura S, Orii T (1995) Mucopolysaccharidosis type IVa - common double deletion in the N-acetylgalactosamine-6-sulfatase gene (GALNS). Genomics 26: 535-542

Hu X, Ray PN, Worton RG (1991) Duplication of three exons in a patient with Duchenne muscular dystrophy caused by intrachromosomal Alu-Alu recombination. American Journal Of Human Genetics 49: 448-448

Huang LS, Breslow JL (1987) A unique AT-rich hypervariable minisatellite 3' to the APO-B gene defines a high information restriction fragment length polymorphism. Journal of Biological Chemistry 262: 8952-8955

Huang XL, Tamaki K, Yamamoto T, Suzuki K, Nozawa H, Uchihi R, Katsumata Y, Neil DL (1996) Analysis of allelic structures at the D7S21 (MS31a) locus in the Japanese, using minisatellite variant repeat mapping by PCR (MVR-PCR). Annals Of Human Genetics 60: 271-279

Hubert R, Macdonald M, Gusella J, Arnheim N (1994) High resolution localization of recombination hotspots using sperm typing. Nature Genetics 7: 420-424

Hulsebos TJM, Bijleveld EH, Riegman PHJ, Smink LJ, Dunham I (1996) Identification and characterization of NF1-related loci on human chromosomes 22, 14 and 2. Human Genetics 98: 7-11

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

Ingram VM (1957) Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. Nature 180: 326-328

Inoue M, Kamiya H, Fujikawa K, Ootsuyama Y, MurataKamiya N, Osaki T, Yasumoto K, Kasai H (1998) Induction of chromosomal gene mutations in *Escherichia coli* by direct incorporation of oxidatively damaged nucleotides - new evaluation method for mutagenesis by damaged DNA precursors in vivo. Journal of Biological Chemistry 273: 11069-11074

Jackson DA, Symons RH, Berg P (1972) Biochemical method for inserting new genetic information into DNA of simian virus 40: circular SV40 molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. Proceedings of the National Academy of Sciences Of the United States Of America 79: 2904-2909

Jagadeeswaran P, Tuan D, Forget BG, Weissman SM (1982) A gene deletion ending at the midpoint of a repetitive DNA sequence in one form of hereditary persistence of fetal haemoglobin. Nature 296: 469-470

Jalanko A, Manninen T, Peltonen L (1995) Deletion of the C terminal end of aspartylglucosaminidase resulting in a lysosomal accumulation disease - evidence for a unique genomic rearrangement. Human Molecular Genetics 4: 435-441

Janicic N, Pausova Z, Cole DEC, Hendy GN (1995) Insertion of an Alu sequence in the Ca^{2+} sensing receptor gene in familial hypocalciuric hypercalcemia and neonatal severe hyperparathyroidism. American Journal of Human Genetics 56: 880-886

Jarman AP, Wells RA (1989) Hypervariable minisatellites - recombinators or innocent bystanders? Trends in Genetics 5: 367-371

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

Jeffreys AJ, Wilson V (1985) Hypervariable regions in human DNA. Genetical Research 45: 213

Jeffreys AJ (1987) Highly variable minisatellites and DNA fingerprints. Biochemical Society Transactions 15: 309-317

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

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

Jeffreys AJ, Macleod A, Tamaki K, Neil DL, Monckton DG (1991a) Minisatellite repeat coding as a digital approach to DNA typing. Nature 354: 204-209

Jeffreys AJ, Turner M, Debenham P (1991b) The efficiency of multilocus DNA fingerprint probes for individualisation and establishment of family relationships, determined from extensive casework. American Journal of Human Genetics 48: 824-840

Jeffreys AJ, Tamaki K, Macleod A, Monckton DG, Neil DL, Armour JAL (1994) Complex gene conversion events in germline mutation at human minisatellites. Nature Genetics 6: 136-145

Jeffreys AJ, Neumann R (1997) Somatic mutation processes at a human minisatellite. Human Molecular Genetics 6: 129-136

Jeffreys AJ, Bois P, Buard J, Collick A, Dubrova Y, Hollies CR, May CA, Murray J, Neil DL, Neumann R, Stead JDH, Tamaki K, Yardley J (1997) Spontaneous and induced minisatellite instability. Electrophoresis 18: 1501-1511

Jeffreys AJ, Neil DL, Neumann R (1998a) Repeat instability at human minisatellites arising from meiotic recombination. EMBO Journal 17: 4147-4157

Jeffreys AJ, Murray J, Neumann R (1998b) High resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. Molecular Cell 2: 267-273

Jeffs AR, Benjes SM, Smith TL, Sowerby SJ, Morris CM (1998) The BCR gene recombines preferentially with Alu elements in complex BCR-ABL translocations of chronic myeloid leukaemia. Human Molecular Genetics 7: 767-776

Jeggo PA, Taccioli GE, Jackson SP (1995) Menage-a-trois double-strand break repair, V(D)J recombination and DNA-PK. Bioessays 17: 949-957

Jennings MW, Jones RW, Wood WG, Weatherall DJ (1985) Analysis of an inversion within the human β globin gene cluster. Nucleic Acids Research 13: 2897-2906

Jiang CK, Hong R, Horowitz SD, Kong XP, Hirschhorn R (1997) An adenosine deaminase (ADA) allele contains two newly identified deleterious mutations (Y97C and L106V) that interact to abolish enzyme activity. Human Molecular Genetics 6: 2271-2278

Jones M, Wagner R, Radman M (1987) Repair of a mismatch is influenced by the base composition of the surrounding nucleotide sequence. Genetics 115: 605-610

Jorde LB, Watkins WS, Viskochil D, Oconnell P, Ward K (1993) Linkage disequilibrium in the neurofibromatosis-I (NFI) region - implications for gene mapping. American Journal of Human Genetics 53: 1038-1050

Junakovic N, Terrrinoni A, Di Franco C, Vieira C, Loevenbruck C (1998) Accumulation of transposable elements in the heterochromatin and on the Y chromosome of *Drosophila simulans* and *Drosophila melanogaster*. Journal of Molecular Evolution 46: 661-668

Jurka J, Klonowski P (1996) Integration of retroposable elements in mammals: selection of target sites. Journal Of Molecular Evolution 43: 685-689

Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proceedings Of the National Academy Of Sciences Of the United States Of America 94: 1872-1877

Kaback DB (1996) Chromosome size dependent control of meiotic recombination in humans. Nature Genetics 13: 20-21

Kaiser P (1980) Pericentriche inversionen menschlicher chromosomen. Thieme, Stuttgart

Kamiya H, Iwai S, Kasai H (1998) The (6-4) photoproduct of thymine-thymine induces targeted substitution mutations in mammalian cells. Nucleic Acids Research 26: 2611-2617

Kan YW, Dozy AM (1978) Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. Proceedings of the National Academy of Sciences Of the United States Of America 75: 5631-5635

Kapitonov V, Jurka J (1996) The age of Alu subfamilies. Journal Of Molecular Evolution 42: 59-65

Kapitonov VV, Holmquist GP, Jurka J (1998) L1 repeat is a basic unit of heterochromatin satellites in cetaceans. Molecular Biology and Evolution 15: 611-612

Karathanasis SK, Ferris E, Haddad IA (1987) DNA inversion within the apolipoproteins AI/CIII/AIV encoding gene cluster of certain patients with premature atherosclerosis. Proceedings Of the National Academy Of Sciences Of the United States Of America 84: 7198-7202

Kasperczyk A, Dimartino NA, Krontiris TG (1990) Minisatellite allele diversification - the origin of rare alleles at the HRASI locus. American Journal of Human Genetics 47: 854-859

Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE (1988) Haemophilia α resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332: 164-166

Kazazian HH, Moran JV (1998) The impact of L1 retrotransposons on the human genome. Nature Genetics 19: 19-24

Kazazian HH (1998) Mobile elements and disease. Current Opinion In Genetics & Development 8: 343-350

Kelner A (1949) Effect of visible light on the recovery of *Streptomyces griseus* conidia from radiation injury. Proceedings of the National Academy of Sciences Of the United States Of America 35: 73-79

Kiaris H, Hatzistamou J, Spandidos D (1996) Instability at the H-ras minisatellite in human atherosclerotic plaques. Atherosclerosis 125: 47-51

Kidwell MG, Lisch DR (1998) Hybrid genetics - transposons unbound. Nature 393: 22-23

Kigawa K, Kihara K, Miyake Y, Tajima S, Funahashi T, Yamamura T, Yamamoto A (1993) Low density lipoprotein receptor mutation that deletes exon-2 and exon-3 by Alu-Alu recombination. Journal Of Biochemistry 113: 372-376

Kimball RF (1987) The development of ideas about the effect of DNA repair on the induction of gene mutations and chromosomal aberrations by radiation and by chemicals. Mutation Research 186: 1-34

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

Kirchgessner CU, Patil CK, Evans JW, Cuorns CA, Fried LM, Carter T, Oettinger MA, Brown JM (1995) DNA-dependent kinase (p350) as a candidate gene for the murine SCID defect. Science 267: 1178-1183

Klar AJS (1992) Developmental choices in mating-type interconversion in fission yeast. Trends in Genetics 8: 208-213

Klar AJS (1998) Propagating epigenetic states through meiosis: where Mendel's gene is more than a DNA moiety. Trends in Genetics 14: 299-301

Knight A, Batzer MA, Stoneking M, Tiwari HK, Scheer WD, Herrera RJ, Deininger PL (1996) DNA sequences of Alu elements indicate a recent replacement of the human autosomal genetic complement. Proceedings Of the National Academy Of Sciences Of the United States Of America 93: 4360-4364

Koch RE (1971) The influence of neighbouring base pairs upon base pair substitution mutation rates. Proceedings of the National Academy of Sciences Of the United States Of America 68: 773-776

Kogan SC, Doherty M, Gitschier J (1987) An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences - application to haemophilia-A. New England Journal of Medicine 317: 985-990

Koivisto UM, Kontula K (1996) A novel deletion/inversion mutation in the low density lipoprotein receptor gene as a cause of heterozygous familial hypercholesterolemia. Human Mutation 8: 326-332

Konopka AK (1988) Compilation of DNA strand exchange sites for nonhomologous recombination in somatic cells. Nucleic Acids Research 16: 1739-1758

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

Korschinsky S (1901) Heterogenesis und evolution. flora 89: 240-363

Krontiris TG, Devlin B, Karp DD, Robert NJ, Risch N (1993) Cancer risk and the HRAS1 minisatellite locus. American Journal of Human Genetics 53: 316

Kudo S, Fukuda M (1989) Structural organization of glycophorin-A and glycophorin-B genes - glycophorin-B gene evolved by homologous recombination at Alu repeat sequences. Proceedings Of the National Academy Of Sciences Of the United States Of America 86: 4619-4623

Kunkel TA (1985a) The mutational specificity of DNA polymerase β during *in vitro* DNA synthesis. Journal of Biological Chemistry 260: 5787-5796

Kunkel TA (1985b) The mutational specificity of DNA polymerase α and γ during *in vitro* DNA synthesis. Journal of Biological Chemistry 260: 12866-12874

Kunkel TA (1990) Misalignment mediated DNA synthesis errors. Biochemistry 29: 8003-8011

Kunkel TA (1992) DNA replication fidelity. Journal of Biological Chemistry 267: 18251-18254

Kwok PY, Deng Q, Zakeri H, Taylor SL, Nickerson DA (1996) Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. Genomics 31: 123-126

La Spada AR, Paulson HL, Fischbeck KH (1994) Trinucleotide repeat expansion in neurological disease. Annals of Neurology 36: 814-822

Lai MD, Beattie KL (1988) Influence of DNA sequence on the nature of mispairing during DNA synthesis. Biochemistry 27: 1722-1728

Laird CD (1987) Proposed mechanism of inheritance and expression of the human fragile X syndrome of mental retardation. Genetics 117: 587-599

Lamarck JBd (1809) Philosophie zoologique, Paris

Lasko D, Cavenee W, Nordenskjold M (1991) Loss of constitutional heterozygosity in human cancer. Annual Review of Genetics 25: 281-314

Laurent AM, Puechberty J, Prades C, Gimenez S, Roizes G (1997) Site-specific retrotransposition of L1 elements within human alphoid satellite sequences. Genomics 46: 127-132

Laurie DA, Hulten MA (1985) Further studies on chiasma distribution and interference in the human male. Annals of Human Genetics 49: 203-214

Lehrman MA, Russell DW, Goldstein JL, Brown MS (1986) Exon-Alu recombination deletes 5 kilobases from the low density lipoprotein receptor gene, producing a null phenotype in familial hypercholesterolemia. Proceedings Of the National Academy Of Sciences Of the United States Of America 83: 3679-3683

Lehrman MA, Goldstein JL, Russell DW, Brown MS (1987) Duplication of seven exons in LDL receptor gene caused by Alu-Alu recombination in a subject with familial hypercholesterolemia. Cell 48: 827-835

Lehrman MA, Russell DW, Goldstein JL, Brown MS (1987) Alu-Alu recombination deletes splice acceptor sites and produces secreted low density lipoprotein receptor in a subject with familial hypercholesterolemia. Journal Of Biological Chemistry 262: 3354-3361

Lehrman AR (1995) Nucleotide excision repair and the link with transcription. Trends in Biomedical Sciences 20: 402-405

Levran O, Doggett NA, Auerbach AD (1998) Identification of Alu-mediated deletions in the Fanconi anemia gene FAA. Human Mutation 12: 145-152

Lewontin RC, Hubby JL (1966) A molecular approach to the study of heterozygosity in natural populations II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54: 594-599

Li W-H, Graur D (1991) Fundamentals of molecular evolution. Sinauer Associates Inc., Sunderland, Massachusetts

Li L, Bray PF (1993) Homologous recombination among three intragene Alu sequences causes an inversion-deletion resulting in the hereditary bleeding disorder Glanzmann thrombasthenia. American Journal Of Human Genetics 53: 140-149

Liang F, Romanienko PJ, Weaver DT, Jeggo PA, Jasin M (1996) Chromosomal double strand break repair in Ku80-deficient cells. Proceedings of the National Academy of Sciences of the United States of America 93: 8929-8933

Lichten M, Goldman ASH (1995) Meiotic recombination hotspots. Annual Review of Genetics 29: 423-444

Liu W-M, Maraia RJ, Rubin CM, Schmid CW (1994) Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. Nucleic Acids Research 22: 1087-1095

Liu B, Nicolaides NC, Markowitz S, Willson JKV, Parsons RE, Jen J, Papdopolous N, Peltomaki P, de la Chapelle A, Hamilton SR, Kinzler KW, Vogelstein B (1995a) Mismatch repair gene defects in sporadic colorectal cancers with microsatellite instability. Nature Genetics 9: 48-55

Liu WM, Chu WM, Choudary PV, Schmid CW (1995b) Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. Nucleic Acids Research 23: 1758-1765

Loidl J (1990) The initiation of meiotic chromosome pairing: the cytological view. Genome 33: 759-778

Lucassen AM, Julier C, Lathrop M, Bell JI (1993) Susceptibility to insulin-dependent diabetes mellitus maps to a 4 kb segment of DNA spanning the insulin gene. American Journal of Human Genetics 53: 183

Ludwig M, Grimm T, Brackmann HH, Olek K (1992) Parental origin of factor IX gene mutations and their distribution in the gene. American Journal of Human Genetics 50: 164-173

Magni GE (1963) The origin of spontaneous mutations during meiosis. Proceedings of the National Academy of Sciences Of the United States Of America 50: 975-980

Mahadevan MS, Foitzik MA, Surh LC, Korneluk RG (1993) Characterization and polymerase chain reaction (PCR) detection of an Alu deletion polymorphism in total linkage disequilibrium with myotonic dystrophy. Genomics 15: 446-448

Mahtani MM, Willard HF (1993) A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation - implications for mechanisms of mutation at short tandem repeat loci. Human Molecular Genetics 2: 431-437

Makalowski W, Mitchell GA, Labuda D (1994) Alu sequences in the coding regions of messenger RNA - source of protein variability. Trends In Genetics 10: 188-193

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

Marcus S, Hellgren D, Lambert B, Fallstrom SP, Wahlstrom J (1993) Duplication in the hypoxanthine phosphoribosyl-transferase gene caused by Alu-Alu recombination in a patient with Lesch-Nyhan Syndrome. Human Genetics 90: 477-482

Margalit H, Nadir E, Bensasson SA (1994) A complete Alu element within the coding sequence of a central gene. Cell 78: 173-174

Martignetti JA, Brosius J (1993) BC200 RNA - a neural RNA polymerase-III product encoded by a monomeric Alu element. Proceedings Of the National Academy Of Sciences Of the United States Of America 90: 11563-11567

Mauillon JL, Michel P, Limacher JM, Latouche JB, Dechelotte P, Charbonnier F, Martin C, Moreau V, Metayer J, Paillot B, Frebourg T (1996) Identification of novel germline HMLH1 mutations including a 22 kb Alu-mediated deletion in patients with familial colorectal cancer. Cancer Research 56: 5728-5733

Maxam AM, Gilbert W (1977) A new method of sequencing of DNA. Proceedings of the National Academy of Sciences of the United States of America 74: 560-561

May CA, Jeffreys AJ, Armour JAL (1996) Mutation rate heterogeneity and the generation of allele diversity at the human minisatellite MS205 (D16S309). Human Molecular Genetics 5: 1823-1833

McClintock B (1951) Chromosome organisation and expression. Cold Spring Harbor Symposium: Quatitative Biology 16: 13-47

McInnis MG, Chakravarti A, Blaschak J, Petersen MB, Sharma V, Avramapoulos D, Blouin JL, Konig U, Brahe C, Matise TC, Warren AC, Talbot CC, C. V, M. L, E. AS (1993) A linkage map of human chromosome 21: 43 PCR markers at average intervals of 2.5 cM. Genomics 16: 562-571

McPhaden AR, Birnie GD, Whaley K (1991) Restriction fragment length polymorphism analysis of the C1-inhibitor gene in hereditary C1-inhibitor deficiency. Clinical Genetics 39: 161-171

Meselson M, Yuan R (1968) DNA restriction enzyme from *E. coli*. Nature 217: 1110-1114

Meselson MS, Radding CM (1975) A general model for genetic recombination. Proceedings Of the National Academy Of Sciences Of the United States Of America 72: 358

Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G (1996) Molecular haplotyping of genetic markers 10 kb apart by allele specific long range PCR. Nucleic Acids Research 24: 4841-4843

Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y (1992) Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Research 52: 643-645

Miki Y, Katagiri T, Kasumi F, Yoshimoto T, Nakamura Y (1996) Mutation analysis in the BRCA2 gene in primary breast cancers. Nature Genetics 13: 245-247

Monckton DG, Neumann R, Guram T, Fretwell N, Tamaki K, Macleod A, Jeffreys AJ (1994) Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. Nature Genetics 8: 162-170

Monkton DG (1992) DNA sequence variation within and around human minisatellites. PhD Thesis, University of Leicester

Monnat RJ, Chiaverotti TA, Hackmann AFM, Maresh GA (1992) Molecular structure and genetic stability of human hypoxanthine phosphoribosyltransferase (HPRT) gene duplications. Genomics 13: 788-796

Moore JK, Haber JE (1996) Capture of retrotransposon DNA at the sites of chromosomal double strand breaks. Nature 383: 644-646

Moran JV, Holmes SE, Naas TP, J. DR, Boeke JD, Kazazian HH (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87: 917-927

Moran JV, DeBerardinis RJ, Kazazian HH (1999) Exon shuffling by L1 retrotransposition. Science 283: 1530-1534

Morgan TH (1912) The physical basis of heredity. J. B. Lippincott, Philadelphia, Pennsylvania
Morris T, Thacker J (1993) Formation of large deletions by illegitimate recombination in the HPRT gene of primary human fibroblasts. Proceedings of the National Academy of Sciences of the United States of America 90: 1392-1396

Muller HJ (1927) Artificial transmutation of the gene. Science 66: 84-87

Muller HJ, Mott-Smith LM (1970) Evidence that natural radioactivity is inadequete to explain the frequency of "natural" mutations. Proceedings of the National Academy of Sciences Of the United States Of America 16: 277-285

Muratani K, Hada T, Yamamoto Y, Kaneko T, Shigeto Y, Ohue T, Furuyama J, Higashino K (1991) Inactivation of the cholinesterase gene by Alu insertion - possible mechanism for human gene transposition. Proceedings of the National Academy of Sciences of the United States of America 88: 11315-11319

Murray J, Buard J, Neil DL, Yeramian E, Tamaki K, Hollies C, Jeffreys AJ (1999) Comparative sequence analysis of human minisatellites showing meiotic repeat instability. Genome Research 9: 130-136

Nag DK, Scherthan H, Rockmill B, Bhargava J, Roeder GS (1995) Heteroduplex DNA formation and homolog pairing in yeast meiotic mutants. Genetics 141: 75-86

Nagel S, Borisch B, Thein SL, Oestreicher M, Nothiger F, Birrer S, Tobler A, Fey MF (1995) Somatic mutations detected by minisatellite and microsatellite DNA markers reveal clonal intratumor heterogeneity in gastrointestinal cancers. Cancer Research 55: 2866-2870

Narita N, Nishio H, Kitoh Y, Ishikawa Y, Minami R, Nakamura H, Matsuo M (1993) Insertion of a 5' truncated L1 element into the 3' end of exon-44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. Journal of Clinical Investigation 91: 1862-1867

Neil DL, Jeffreys AJ (1993) Digital DNA typing at a second hypervariable locus by minisatellite variant repeat mapping. Human Molecular Genetics 2: 1129-1135

Neil Dl (1994) Allelic variation and mutation at human hypervariable minisatellite loci. PhD Thesis, University of Leicester

Nelson DL (1993) Six human genetic disorders involving mutant trinucleotide repeats: Similarities and differences. In: Davies KE, Warren ST (eds) Genome Rearrangement and Stability, vol 7. Cold Spring Harbor Laboratory Press, pp 1-24

Nevo E, Moseman JG, Beiles A, Zohary D (1984) Correlation of ecological factors and allozymic variations with resistance to *Erysiphe graminis hordei* in hordeum-spontaneum in Israel - patterns and application. Plant Systematics and Evolution 145: 79-96

Nicholls RD, Fischelghodsian N, Higgs DR (1987) Recombination at the human alpha-globin gene cluster - sequence features and topological constraints. Cell 49: 369-378

Nicolas A (1998) Relationship between transcription and initiation of meiotic recombination: toward chromatin accessibility. Proceedings Of the National Academy Of Sciences Of the United States Of America 95: 87-89

Ninio J (1996) Gene conversion as a focusing mechanism for correlated mutations: a hypothesis. Molecular & General Genetics 251: 503-508

Novick GE, Novick CC, Yunis J, Yunis E, Martinez K, Duncan GG, Troup GM, Deininger PL, Stoneking M, Batzer MA, Herrera RJ (1995) Polymorphic human specific Alu insertions as markers for human identification. Electrophoresis 16: 1596-1601

Nystrom-Lahti M, Kristo P, Nicolaides NC, Chang SY, Aaltonen LA, Moisio AL, Jarvinen HJ, Mecklin JP, Kinzler KW, Vogelstein B, Delachapelle A, Peltomaki P (1995) Founding mutations and Alu-mediated recombination in hereditary colon cancer. Nature Medicine 1: 1203-1206

Oettinger MA (1996) Cutting apart V(D)J recombination. Current Opinion in Genetics & Development 6: 141-145

Orkin SH, Alter BP, Altay C, Mahoney MJ, Lazarus H, Hobbins JC, Nathan DG (1978) Application of endonuclease mapping to the analysis and prenatal diagnosis of thalassemias caused by globin gene deletion. New England Journal of Medicine 299: 166-172

Orr HT, Chung MY, Banfi S, Kwiatkowski TJ, Servadio A, Beaudet AL, McCall AE, Duvick LA, Ranum LPW, Zoghbi HY (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type-1. Nature Genetics 4: 221-226

Osman F, Subramani S (1998) Double strand break induced recombination in eukaryotes. Progress In Nucleic Acid Research and Molecular Biology 58: 263-299

Ottolenghi S, Giglioni B (1982) The deletion in a type of δ^{o}-β^{o}-thalassemia begins in an inverted AluI repeat. Nature 300: 770-771

Page DC, Brown LG, Delachapelle A (1987) Exchange of terminal portions of X-chromosomal and Y-chromosomal short arms in human-XX males. Nature 328: 437-440

Palmer MS, Collinge J (1993) Mutations and polymorphisms in the prion protein gene. Human Mutation 2: 168-173

Pardue M, Gall J (1970) Chromosomal localisation of mouse satellite DNA. Science 168: 1356-1358

Pattinson JK, Millar DS, Grundy CB, Wieland K, Mibashan RS, Martinowitz U, McVey J, Tan-Un K, Vidaud M, Goossens M, Sampietro M, Krawczak M, Reiss J, Zoll B, Whitmore D, Bradshaw A, Wensley R, Ajani A, Mitchell V, Rizza C, Maia R, Winter P, Mayne EE, Schwartz M, Green PJ, Kakkar VV, Tuddenham EGD, Cooper DN (1990) The molecular genetic analysis of haemophilia A: a directed search strategy for the detection of point mutations in the human factor VIII gene. Blood 76: 2242-2248

Pauly M, Kayser I, Schmitz M, Ries F, Hentges F, Dicato M (1995) Repetitive DNA sequences located in the central region of the human MDR1 (multidrug-resistance) gene may account for a gene fusion event during its evolution. Journal Of Molecular Evolution 41: 974-978

Petrij-Bosch A, Peelen T, vanVliet M, vanEijk R, Olmer R, Drusedau M, Hogervorst FBL, Hageman S, Arts PJW, Ligtenberg MJL, MeijersHeijboer H, Klijn JGM, Vasen HFA, Cornelisse CJ, vantVeer LJ, Bakker E, vanOmmen GJ, Devilee P (1997) BRCA1 genomic deletions are major founder mutations in Dutch breast cancer patients (vol 17, pg 341, 1997). Nature Genetics 17: 503

Pettijohn D, Hanawalt P (1964) Evidence for repair - replication of ultraviolet damaged DNA in bacteria. Journal of Molecular Biology 9: 395-410

Polymeropoulos MH, Rath DS, Xiao H, Merril CR (1992) Tetranucleotide repeat polymorphism at the human β-actin related pseudogene H-β-AC-ψ-2 (ACTβψ2). Nucleic Acids Research 20: 1432

Porter SE, White MA, Petes TD (1993) Genetic evidence that the meiotic eecombination hotspot at the HIS4 locus of *Saccharomyces cerevisiae* does not represent a site for a symmetrically processed double strand break. Genetics 134: 5-19

Pousi B, Hautala T, Heikkinen J, Pajunen L, Kivirikko KI, Myllyla R (1994) Alu-Alu recombination results in a duplication of seven exons In the lysyl hydroxylase gene in a patient with the type-VI variant of Ehlers-Danlos syndrome. American Journal Of Human Genetics 55: 899-906

Prades C, Laurent AM, Puechberty J, Yurov Y, Roizes G (1996) SINE and LINE within human centromeres. Journal Of Molecular Evolution 42: 37-43

Puget N, Torchard D, SerovaSinilnikova OM, Lynch HT, Feunteun J, Lenoir GM, Mazoyer S (1997) A 1 kb Alu-mediated germline deletion removing BRCA1 exon 17. Cancer Research 57: 828-831

Purandare SM, Patel PI (1997) Recombination hot spots and human disease. Genome Research 7: 773-786

Radman M, Wagner R (1993) Mismatch recognition in chromosomal interactions and speciation. Chromosoma 102: 369-373

Rapp M, Therman E, Denniston C (1988) Non-pairing of the X and Y chromosomes in the spermatocytses of BDF1 mice. Cytogenetics and Cellular Genetics 19: 85-93

Rayssiguier C, Thaler DS, Radman M (1989) The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch repair mutants. Nature 342: 396-401

Reiter LT, Murakami T, Koeuth T, Pentao L, Muzny DM, Gibbs RA, Lupski JR (1996) A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. Nature Genetics 12: 288-297

Reiter LT, Hastings PJ, Nelis E, DeJonghe P, VanBroeckhoven C, Lupski JR (1998) Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. American Journal Of Human Genetics 62: 1023-1033

Richards RI, Sutherland GR (1994) Simple repeat DNA is not replicated simply. Nature Genetics 6: 114-116

Ripley LS (1982) Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. Proceedings of the National Academy of Sciences of the United States of America - Biological Sciences 79: 4128-4132

Roeder GS (1997) Meiotic chromosomes: it takes two to tango. Genes and Development 11: 2600-2621

Rose O, Falush D (1998) A threshold size for microsatellite expansion. Molecular Biology and Evolution 15: 613-615

Ross LO, Maxfield R, Dawson D (1996) Exchanges are not equally able to enhance meiotic chromosome segregation in yeast. Proceedings Of the National Academy Of Sciences Of the United States Of America 93: 4979-4983

Rousseau F, Heitz D, Biancalana V, Blumenfeld S, Kretz C, Boue J, Tommerup N, Van der Hagen C, DeLozier-Blanchet C, Croquette M-F, Gilgenkrantz S, Jalbert P, Voelckel M-A, Oberle I, Mandel J-L (1991) Direct diagnosis by DNA analysis of the fragile X syndrome of mental retardation. New England Journal of Medicine 325: 1673-1681

Rouyer F, Simmler MC, Page DC, Weissenbach J (1987) A sex chromosome rearrangement in a human XX male caused by Alu-Alu recombination. Cell 51: 417-425

Royle NJ, Clarkson RE, Wong Z, Jeffreys AJ (1988) Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. Genomics 3: 352-360

Rubin CM, Vandevoort CA, Teplitz RL, Schmid CW (1994) Alu repeated DNAs are differentially methylated in primate germ cells. Nucleic Acids Research 22: 5121-5127

Rubnitz, Subramani (1984) The minimum amount of homology required for homologous recombination in mammalian cells. Molecular and Cellular Biology 4: 2253-2258

Rüdiger NS, Gregersen N, Kiellandbrandt MC (1995) One short well conserved region of Alu-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. Nucleic Acids Research 23: 256-260

Rupert CS, Goodgal SH, Herriott RM (1958) Photoreactivation *in vitro* of ultraviolet inactivated *Haemophilus influenzae* transforming factor. Journal of General Physiology 41: 57-67

Russanova VR, Driscoll CT, Howard BH (1995) Adenovirus type-2 preferentially stimulates polymerase-III transcription of Alu elements by relieving repression - a potential role for chromatin. Molecular and Cellular Biology 15: 4282-4290

Saffer JD, Thurston SJ (1989) A negative regulatory element with properties similar to those of enhancers is contained within an Alu sequence. Molecular and Cellular Biology 9: 355-364

Saiki RD, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N (1985) Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230: 1350-1354

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning, a laboratory manual. Cold Spring Harbour Laboratory Press

Sanderson RJ, Mosbaugh DW (1998) Fidelity and mutational specificity of uracil-initiated base excision DNA repair synthesis in human glioblastoma cell extracts. Journal of Biological Chemistry 273: 24822-24831

Sanger R, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences Of the United States Of America 74: 5463-5467

Sawadogo M, Vandyke MW (1991) A rapid method for the purification of deprotected oligodeoxynucleotides. Nucleic Acids Research 19: 674

Scherthan H, Weich S, Schwegler H, Heyting C, Harle M, Cremer T (1996) Centromere and telomere movements during early meiotic prophase of mouse and man are associated with the onset of chromosome pairing. Journal of Cellular Biology 134: 1109-1125

Schmid CW, Jelinek WR (1982) The Alu family of dispersed repetitive sequences. Science 216: 1065-1070

Schmid CW (1996) Alu: Structure, origin, evolution, significance, and function of one tenth of human DNA. Progress in Nucleic Acid Research and Molecular Biology 53: 283-319

Schmid CW (1998) Does SINE evolution preclude Alu function? Nucleic Acids Research 26: 4541-4550

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, Doniskeller H (1985) Detection of more than five hundred single copy RFLPs by random screening. Cytogenetics and Cell Genetics 40: 739

Schwacha A, Kleckner N (1994) Identification of joint molecules that form frequently between homologs but rarely between sister chromatids during yeast meiosis. Cell 76: 51-63

Schwartz A, Chan DC, Brown LG, Alagappan R, Pettay D, Disteche C, McGillivray B, delaChapelle A, Page DC (1998) Reconstructing hominid Y evolution: X homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. Human Molecular Genetics 7: 1-11

Seeburg E, Eide L, Bjoras M (1995) The base excision repair pathway. Trends in Biochemical Sciences 20: 391-397

Shannon M, Weigert M (1998) Fixing mismatches. Science 279: 1159-1160

Shapiro JA (1969) Mutations caused by the insertion of genetic material into the galactose operon of *Escherichia coli*. Journal of Molecular Biology 40: 93-106

Shen MR, Batzer MA, Deininger PL (1991) Evolution of the master Alu gene(s). Journal Of Molecular Evolution 33: 311-320

Sherman SL, Petersen MB, Freeman SB, Hersey J, Pettay D, Taft L, Frantzen M, Mikkelsen M, Hassold TJ (1994) Nondisjunction of chromosome-21 in maternal meiosis I - evidence for a maternal age-dependent mechanism involving reduced recombination. Human Molecular Genetics 3: 1529-1535

Shimada K, Weissberg R (1973) *E. coli* mutants produced by the insertion of bacteriophage λ. Genetics 73 (Suppl.): 81-83

Shimamura M, Nikaido M, Ohshima K, Okada N (1998) A SINE that acquired a role in signal transduction during evolution. Molecular Biology and Evolution 15: 923-925

Shine J, Seeburg PH, Martial JA, Baxter JD, Goodman HM (1977) Construction and analysis of recombinant DNA for chorionic somatomammotropin. Nature 270: 494-499

Sia EA, Jinks Robertson S, Petes TD (1997) Genetic control of microsatellite stability. Mutation Research - DNA Repair 383: 61-70

Singer BS, Westlye J (1988) Deletion formation in bacteriophage-T4. Journal of Molecular Biology 202: 233-243

Smit AFA (1996) The origin of interspersed repeats in the human genome. Current Opinion In Genetics & Development 6: 743-748

Smith KN, Nicolas A (1998) Recombination at work for meiosis. Current Opinion In Genetics & Development 8: 200-211

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

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. Journal of Molecular Biology 98: 503-517

Sparrow AH (1950) Tolerance of *Tradescantia* to continuous exposures of gamma radiation from colbalt 60. Genetics 35: 135

Speyer JF (1965) Mutagenic DNA polymerase. Biochemical and Biophysical Research Communications 21: 6-8

Stadler LJ (1928) The rate of induced mutation in relation to dormancy, temperature and dosage. Anatomical Record 41: 97

Stahl F (1996) Meiotic recombination in yeast: coronation of the double strand break repair model. Cell 87: 965-968

Stoppa-Lyonnet D, Tosi M, Laurent J, Sobel A, Lagrue G, Meo T (1987) Altered C1 inhibitor genes in type-I hereditary angioedema. New England Journal Of Medicine 317: 1-6

Stoppa-Lyonnet D, Carter PE, Meo T, Tosi M (1990) Clusters of intragenic Alu repeats predispose the human C1-inhibitor locus to deleterious rearrangements. Proceedings Of the National Academy Of Sciences Of the United States Of America 87: 1551-1555

Stoppa-Lyonnet D, Duponchel C, Meo T, Laurent J, Carter PE, Aralachaves M, Cohen JHM, Dewald G, Goetz J, Hauptmann G, Lagrue G, Lesavre P, Lopeztrascasa M, Misiano G, Moraine C, Sobel A, Spath PJ, Tosi M (1991) Recombinational biases in the rearranged C1-inhibitor genes of hereditary angioedema patients. American Journal Of Human Genetics 49: 1055-1062

Strachan T (1992) The human genome. Bios Scientific Publishers Limited, Oxford

Strand M, Prolla TA, Liskay RM, Petes TD (1993) Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 365: 274-276

Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M (1967) Frameshift mutations and the genetic code. Cold Spring Harbor Syposium: Quantitative Biology 31: 77-84

Strout MP, Marcucci G, Bloomfield CD, Caligiuri MA (1998) The partial tandem duplication of ALL1 (MLL) is consistently generated by Alu-mediated homologous recombination in acute myeloid leukemia. Proceedings Of the National Academy Of Sciences Of the United States Of America 95: 2390-2395

Sturtevant AH (1913) The linear arrangement of six sex-linked factors in *Drosophila* as shown by their mode of association. Journal of Experimental Zoology 14: 43-59

Sutherland GR, Baker E, Richards RI (1998) Fragile sites still breaking. Trends in Genetics 14: 501-506

Sutton WS (1903) The chromosomes in heredity. Biological Bulletin March Biological Laboratories, Woods Hole 4: 231-248

Swensen J, Hoffman M, Skolnick MH, Neuhausen SL (1997) Identification of a 14 kb deletion involving the promoter region of BRCA1 in a breast cancer family. Human Molecular Genetics 6: 1513-1517

Szostak JW (1983) The double strand break repair model for recombination. Cell 33: 25-35

Tamaki K, Monckton DG, Macleod A, Allen M, Jeffreys AJ (1993) Four-state MVR-PCR - increased discrimination of digital DNA typing by simultaneous analysis of two polymorphic sites within minisatellite variant repeats at D1S8. Human Molecular Genetics 2: 1629-1632

Tamaki K, May CA, Dubrova YE, Jeffreys AJ (1999) Extremly complex repeat shuffling during germline mutation at human minisatllite B6.7. Human Molecular Genetics 85: 879-888

Taylor AL (1963) Bacteriophage-induced mutation in *Escherichia coli*. Proceedings of the National Academy of Sciences Of the United States Of America 50: 1043-1051

Teng SC, Kim B, Gabriel A (1996) Retrotransposon reverse transcriptase-mediated repair of chromosomal breaks. Nature 383: 641-644

Therman E (1986) Human chromosomes: structure, behaviour, effects, Second Edition edn. Springer-Verlag, New York

Tusie-Luna MT, White PC (1995) Gene conversions and unequal crossovers between CYP21 (steroid 21-hydroxylase gene) and CYP21p involve different mechanisms. Proceedings of the National Academy of Sciences of the United States of America 92: 10796-10800

Tvrdik T, Marcus S, Hou S-M, Falt S, Noori P, Padlutskaja N, Hanefeld F, Stromme P, Lambert B (1998) Molecular characterisation of two deletion events involving Alu sequences, one novel base substitution and two tentative hotspot mutations in the hypoxanthine phophoribosyltransferase (HPRT) gene in five patients with Lesch-Nyhan syndrome. Human Genetics 103: 311-318

van Endert PM, Lopez MT, Patel SD, Monaco JJ, NcDevitt HO (1992) Genomic polymorphism, recombination, and linkage disequilibrium in human major histocompatibility-encoded antigen-processing genes. Proceedings of the National Academy of Sciences of the United States of America 89: 1629-1632

Vanin EF, Henthorn PS, Kioussis D, Grosveld F, Smithies O (1983) Unexpected relationships between four large deletions in the human beta-globin gene cluster. Cell 35: 701-709

Vansant G, Reynolds WF (1995) The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. Proceedings Of the National Academy Of Sciences Of the United States Of America 92: 8229-8233

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

Vidaud M, Vidaud D, Siguret V, Lavergne JM, Goossens M (1988) Mutational insertion of an Alu sequence causes haemophilia. American Journal of Human Genetics 45: A226

Viswanathan A, Doetsch PW (1998) Effects of nonbulky DNA base damages on *Escherichia coli* RNA polymerase-mediated elongation and promoter clearance. Journal of Biological Chemistry 273: 21276-21281

Vnencak-Jones CL, Phillips III JA (1990) Hot spots for growth hormone gene deletions in homologous regions outside of Alu repeats. Science 250: 1745-1748

Vogel F, Motulsky AG (1986) Human genetics - problems and approaches, Second Edition edn. Springer, Berlin

Voytas DF (1996) Retroelements in genome organization. Science 274: 737-738

Walker GC (1995) SOS-regulated proteins in translesion DNA synthesis and mutagenesis. Trends in Biochemical Sciences 20: 416-420

Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover AM, Collins FS (1991) A *de novo* Alu insertion results in neurofibromatosis type 1. Nature 353: 864-866

Watnick TJ, Gandolph MA, Weber H, Neumann HPH, Germino GG (1998) Gene conversion is a likely cause of mutation in PKD1. Human Molecular Genetics 7: 1239-1243

Watson JD, Crick FHC (1953) A structure for deoxyribose nucleic acid. Nature 171: 737-738

Waugh O'Neil R, O'Neill MJ, Marshall Graves JA (1998) Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature 393: 68-72

Weaver DT, DePamphilis ML (1982) Specific sequences in native DNA that arrest synthesis by DNA polymerase a. Journal of Biological Chemistry 257: 2075-2086

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Human Molecular Genetics 2: 1123

Weir BS (1990) Genetic data analysis. Sinauer Associates Inc., Sunderland, MA

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M (1992) A second-generation linkage map of the human genome. Nature 359: 794-801

Wolff RK, Nakamura Y, White R (1988) Molecular characterisation of a spontaneously generated new allele at a VNTR locus: No exchange of flanking DNA sequence. Genomics 3: 347-351

Wolff RK, Plaekte R, Jeffreys AJ, White R (1989) Unequal crossing over between homologous chromosomes is not the major mechanism involved in generation of new alleles at VNTR loci. Genomics 5: 382-384

Wong Z, Wilson V, Patel I, Povey S, Jeffreys AJ (1987) Characterization of a panel of highly variable minisatellites cloned from human DNA. Annals of Human Genetics 51: 269-288

Woods-Samuel P, Kazazian HH, Antonarakis SE (1991) Non-homologous recombination in the human genome: deletions in the human factor VIII gene. Genomics 10: 94 -101

Wu CI, Maeda N (1987) Inequality in mutation rates of the two strands of DNA. Nature 327: 169-170

Yandava CN, Gastier JM, Pulido JC, Brody T, Sheffield V, Murray J, Buetow K, Duyk GM (1997) Characterization of Alu repeats that are associated with trinucleotide and tetranucleotide repeat microsatellites. Genome Research 7: 716-724

Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends in Genetics 13: 335-340

Zoghbi HY (1996) The expanding world of ataxins. Nature Genetics 14: 237-238