Chapter One Introduction

1.1 Types of repetitive elements

In the human genome, a large number of different types of repetitive elements are found. Three main categories are now recognized: (i) a highly repetitive class including sequence families with more than 10^5 copies per haploid genome, (ii) a middle repetitive sequence class (10^2-10^5 copies), and (iii) a low repetitive class whose members possess between 2 and 100 copies per haploid genome (reviewed by Jelinek and Schmid, 1982; Hardman, 1986; Vogt, 1990; Cooper and Krawczak, 1993).

Repetitive sequences found in the human genome can be in dispersed manner, like the *Alu* elements, belonging to the short interspersed repeat elements (SINEs) (Schmid and Jelinek, 1982), L1 elements of LINEs (long interspersed repeat elements) (Singer and Skrowronski, 1985), or in tandemly repeated arrays (Armour *et al.*, 1993b).

1.2 Tandemly repetitive DNA

Tandemly repetitive DNA comprises satellite DNA, midisatellite DNA, minisatellite DNA, and microsatellite DNA (Cooper and Krawczak, 1993).

1.2.1 Satellite DNA

Satellite DNA comprises the majority of heterochromatin and is clustered in tandem arrays of up to several megabases in length. A number of different families [e.g., simple sequence (5-25 bp repeats), alphoid (171 bp repeat), *Sau*3A (~68 bp repeat)] have been identified (reviewed by Vogt, 1990). The major alphoid satellite sequences, which exist in chromosome-specific subclasses, between them account for about 5% of the human genome (Waye and Willard, 1986). The blocks of alphoid satellite are composed of tandem arrays up to 5 Mb in length (Willard, 1991), and appear to correspond to the functional part of human centromeres (Willard, 1990).

1.2.2 Midisatellite DNA

Midisatellite DNA, which consists of 250-500 kilobases of repetitive DNA, is clustered at a single locus and shows a highly polymorphic pattern in the population (Nakamura *et al.*, 1987b). The core sequences of the midisatellite locus bear some

homology to those of the repetitive sequence of the insulin gene (Bell *et al.*, 1982) and the zeta-globin pseudogene (Proudfoot *et al.*, 1982). Midisatellite is speculated to have an important chromosomal function, such as the recognition of a homologous region at meiotic pairing (Nakamura *et al.*, 1987b).

1.2.3 Minisatellite DNA

Tandem arrays of intermediate size, the minisatellite loci (Jeffreys *et al.*, 1985a,b), have a total length usually in the range 0.5-30 kb, composed of short repeated units. The hypervariable minisatellite sequences (about 10^4 copies/genome) share a core consensus sequence (GGTGGGCAGARG, where R = purine), which is reminiscent of the *Escherichia coli* Chi element known to be a signal for generalized recombination (Jeffreys, 1987). These minisatellites exhibit substantial copy number variability in terms of the number of constituent repeat units, probably due to the process of unequal recombination and slipped mispairing (Cooper and Krawczak, 1993).

1.2.4 Microsatellite DNA

Microsatellite DNA families are simple sequence repeats, the most common being (1) (A)n/(T)n and (2) (CA)n/(TG)n and (CT)n/(AG)n types. These two different sub-families each accounts for between 0.2 and 0.5% of the genome, respectively (Vogt, 1990; Williamson *et al.*, 1991; Beckmann and Weber, 1992). As for the short arrays of dinucleotide repeats (generally up to 60 bp), they are found extremely abundant and widely dispersed in the genome (Litt and Luty, 1989; Weber and May, 1989; Weber, 1990). The microsatellite DNA's high copy number variability and association with a considerable number of different genes has meant that it provides a very valuable source of highly informative markers for indirect diagnosis of human genetic disease (Cooper and Krawczak,1993).

1.3 Minisatellite DNA and DNA polymorphism

Polymorphism in DNA structure provides the basis of genetic analysis (Armour and Jeffreys, 1992a). The first human DNA sequence variants to be analysed directly were

base substitutional polymorphisms affecting the recognition sites for restriction enzymes or restriction fragment length polymorphisms (RFLPs) for DNA sequence adjacent to human β-globin structural gene (Kan and Dozy, 1978). As these restriction-site dimorphisms can only have two allelic states, their informativeness is limited to a maximum heterozygosity of 50% (Armour and Jeffreys, 1992a).

The first hypervariable minisatellite sequence was fortuitously isolated as a random clone from human chromosome 14 (Wyman and White, 1980). This demonstrated that hypervariable regions (HVRs) or variable number of tandem repeats (VNTRs) exist in human DNA, although the variable DNA region itself has only been cloned in 1985 (Wyman *et al.*, 1985). Subsequently, other highly polymorphic sequences were discovered near a region 5' to the human insulin gene (Bell *et al.*, 1982), another 3' to the c-Ha-*ras*-1 oncogene (Capon *et al.*, 1983), in and around the α-globin gene cluster (Higgs *et al.*, 1981), and other locations throughout the genome (Jeffreys *et al.*, 1985a).

Among the minisatellite arrays, there is a range of variability. Some loci appear to be monomorphic in the population studied, while others have been isolated which show hypervariability (Armour *et al.*, 1990). For those human minisatellite loci that show multi-allelic polymorphism, the polymorphism is due to allelic variation in the number of tandemly repeated units, and hence results in variation in the length of an array (Wyman and White, 1980; Higgs *et al.*, 1981; Jeffreys *et al.*, 1985a, b; Wong *et al.*, 1986, 1987; Nakamura *et al.*, 1987a).

The polymorphism at highly variable human minisatellites derives ultimately from an extremely high mutation rate to new length alleles in the germline; this germline mutation rate is so high at some loci that it can be measured directly by inspection of pedigrees (Jeffreys *et al.*, 1988), and can be as high as 15% per gamete (Vergnaud *et al.*, 1991). While unequal sister chromatid exchange and unequal recombination (or gene conversion) between alleles have long been implicated as mechanisms of mutational change in tandemly repeated blocks of DNA (Smith, 1976; Dover, 1982), other mechanisms, such

as replication slippage (Tautz *et al.*, 1986; Levinson and Gutman, 1987) and deletion by intramolecular recombination, are also plausible mechanisms for mutation processes altering the number of tandem repeats (Armour *et al.*, 1993b).

Structural analyses of mutant MS32 alleles have revealed the involvement of complex conversion-like events. Processes such as unequal sister chromatid exchange or replication slippage, inter-allelic transfer, insertion of a donor segment within recipient allele, intra-allelic reduplications interrupted by segment(s) from the other allele, and anomalous repeats acquired by the recipient allele from no obvious origin in either progenitor allele were observed (Jeffreys *et al.*, 1994).

1.4 Hypervariable minisatellite MS32

The hypervariable minisatellite MS32 is probably the best characterized human minisatellite to date (Armour *et al.*, 1993b). The minisatellite MS32 (locus D1S8) located interstitially on chromosome 1 at 1q42 to 1q43 (Wong *et al.*, 1987; Royle *et al.*, 1988). The minisatellite consists of a 29 bp repeat sequence varying in number from 12 to more than 600 (Armour *et al.*, 1989).

DNA sequence analysis of MS32 (Wong *et al.*, 1987) showed that approximately half of the repeat units share an A to G transition (termed site-I), giving rise to two major classes of repeat units, designated a-type and t-type (cut or not cut by *Hae*III), which differ by a single base substitution and show highly diverse dispersion patterns within alleles (Jeffreys *et al.*, 1990). MS32 repeats also contain a second polymorphic C-T transition just two base pair 5' to site-I (termed site-II) (Wong *et al.*, 1987; Tamaki *et al.*, 1993). Null repeats were reported, signifying the existence of further repeat sequence variants in MS32 (Jeffreys *et al.*, 1991; Tamaki *et al.*, 1992).

Internal mapping studies have revealed that there is a gradient or polarity of variation within the tandem array of MS32 (Armour and Jeffreys, 1992b). There appears to be a relatively invariant end of the array, at which there is relatively little variation in

internal structure, and an 'ultravariable' end, at which most variation and mutation seems to occur (Armour and Jeffreys, 1992b).

The penultimate repeat at the ultravariable end of the MS32 was found to have extensive substitutions at its 3' end and is 4 bp larger, while the final repeat is also heavily diverged, with a 13 out of 20 mismatch with the first 20 bp of a standard repeat (Dover, 1992).

1.5 Minisatellite variant repeat mapping by the polymerase chain reaction (MVR-PCR)

Digital minisatellite variant repeat analysis by using PCR (MVR-PCR) is a new approach to assessing individual variation in DNA (Jeffreys *et al.*, 1991). The process is simple, rapid and can give unambiguous and easily interpretable information in a digital format ideal for computer databasing and analysis. MVR-PCR examines repeat unit sequence variation within minisatellites and has so far been successfully applied to the hypervariable human minisatellite MS32 (locus D1S8) (Jeffreys *et al.*, 1991; Monckton *et al.*, 1993; Yamamoto *et al.*, 1994a,b), MS205 (locus D16S309) (Armour *et al.*, 1993a), and MS31A (locus D7S21) (Neil and Jeffreys, 1993; Huang *et al.*, 1996).

MVR-PCR uses a PCR primer at a fixed site in the DNA flanking the repeat array, together with primers specific for different repeat variants, to produce ladders of PCR products extending from the flanking DNA to each repeat unit of a given type. If only one site of repeat unit sequence variation is targeted, then "two-state" MVR-PCR can be used to distinguish the a- and t-type repeats, to generate binary codes of the two repeat types interspersed along an allele (Jeffreys *et al.*, 1991, 1995).

More internal structural information can be recovered by analysing additional sites of variability, if they exist, with appropriate MVR-PCR primers. For example, the 29 bp repeat unit of minisatellite MS32 contains two base substitutional polymorphic sites separated by 1 bp (Wong *et al.*, 1987), and simultaneous analysis of both sites by "four-state" MVR-PCR generates a quaternary code (E-, e-, Y-, and y-type repeats) from an

allele and doubles the information recoverable by analysis of either site alone (Tamaki *et al.*,1993; Jeffreys *et al.*, 1995).

MVR-PCR can be applied directly to total genomic DNA to produce a diploid digital code derived from the superimposed maps of the two individual alleles. This diploid digital code is of considerable potential use in forensic identification (Jeffreys *et al.*, 1991; Yamamoto *et al.*, 1994a; Tamaki *et al.*, 1995). For allelic diversity studies, it is necessary to determine codes from individual alleles, most simply by using allele specific MVR-PCR (Armour *et al.*, 1993a; Monckton *et al.*, 1993; Neil and Jeffreys, 1993; Tamaki *et al.*, 1993; Huang *et al.*, 1996).

The extreme variability seen in diploid codes results from extensive variation in the internal maps of individuals alleles. The number of different alleles at MS32 distinguishable by MVR-PCR may be as high as 10^8 for the total world population of 5 x 10^9 individuals (based on known mutation rate and population size) (Jeffreys *et al.*, 1991; Monckton *et al.*, 1993). Analysis of single alleles provides new insights into the evolution of minisatellites and the mutation processes operating at these loci, and could provide a powerful new tool for analysing human population divergence (Jeffreys *et al.*, 1991,1994,1995; Monckton *et al.*, 1993,1994; Armour *et al.*, 1996).

1.6 DNA flanking the minisatellite MS32

DNA sequences flanking the minisatellite MS32 (Armour *et al.*, 1989) revealed the presence of L1 element and long terminal repeat (LTR) of a retrovirus-like element (RTVL-I) (Maeda, 1985) (Figure 1). L1 element was found at the 5' end of the flanking region, which was suggested to be the start of a 5' truncated member of the L1 family of dispersed repeats. Retroviral LTR sequences were found on either side of the minisatellite, in which RTVL-I extends to the boundary of the tandem repeat array, and resumes on the other side. It was suggested that the tandem repeat block may have expanded from within a diverged member of the LTR family (Armour *et al.*, 1989).

Detailed sequence analysis of the DNA flanking the ultravariable end of the minisatellite (Figure 2) has enabled identification of three common polymorphic base substitutions (Hump1, *Hin*fI, and Hump2) within 400 bp of the terminal repeat (Armour *et al.*, 1989; Monckton *et al.*, 1993). A fourth site of variation (O1), a G to C transversion 48 bp upstream of the tandem repeat array and 16 bp upstream of a diverged MS32 repeat which precedes the array, was reported by Monckton *et al.* (1994). These flanking polymorphisms provide another approach to single allele mapping in heterozygotes via the use of allele specific flanking primers in 'knockout' MVR-PCR or allele-specific MVR-PCR (Monckton *et al.*, 1993), facilitating the rapid acquisition of further single allele codes from unrelated individuals. These flanking markers should also assist the investigations into the role of homologous recombination in maintaining MS32 variability (Armour *et al.*, 1993b).

Initial studies of analysing haplotypes at these polymorphic sites have revealed that linkage disequilibrium between the flanking markers does exist, but is not absolute as might be expected for such closely spaced loci (Monckton *et al.*, 1993). Haplotyping of the flanking polymorphisms relative to the minisatellite array shows that in general, groups of aligned alleles share flanking haplotypes, but occasional switching within allele groups is observed, indicative of inter-allelic recombination. However, this type of analysis cannot distinguish true recombinations from shorter 'patches' of gene conversion (Armour *et al.*, 1993b; Monckton *et al.*, 1994).

1.7 Objectives of this study

Before allele-specific MVR-PCR analysis at the minisatellite MS32 can be performed on DNA samples from individuals of a particular population, information on polymorphisms in the 5' flanking region of MS32 should be obtained. For Caucasian and Japanese populations, analyses of the DNA flanking the ultravariable end of MS32 have been reported (Armour *et al.*, 1989; Monckton *et al.*, 1993, 1994). For the Malaysian

population, only *Hin*fI (Koh *et al.*, 1993) and Hump1 (Koh *et al.*, 1994) site variations in the DNA flanking MS32 have been reported.

Hence, the objectives of this study were:

(i) to assay five polymorphic sites (O2, *Hin*fI, Hump1, O1, and Hump2) in the DNA flanking the ultravariable end of MS32 in 210 healthy, unrelated Malaysian individuals (70 Chinese, 70 Indians, and 70 Malays);

(ii) to haplotype these polymorphic sites; and

(iii) to sequence the DNA flanking the ultravariable end of MS32.

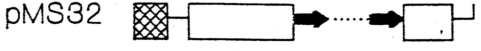
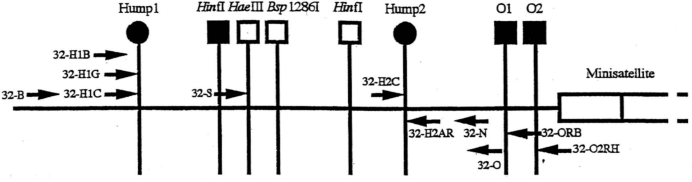


Figure 1 : Dispersed repeat elements flanking the human minisatellite MS32. The tandem repeat arrays have been abbreviated to a pair of arrows separated by a dotted line. Boxes represent regions of dispersed repeat sequence - L1 element as cross-hatched box, retroviral LTR sequences as open boxes (Armour *et al.*, 1989).

A



B

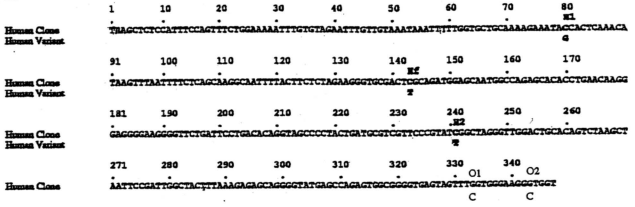


Figure 2 : Polymorphisms in the 5' flanking region of the minisatellite MS32. (A) Map of the region showing polymorphic sites and PCR primers. Filled circles represent polymorphic base substitutions, open squares non-polymorphic restriction sites and filled squares polymorphic restriction sites. Arrows indicate PCR primers. (B) Sequence of the MS32 5' flanking region amplified by PCR primers 32-O and 32-B (modified from Monckton *et al.*, 1993).