Chapter One

INTRODUCTION

CHAPTER ONE: INTRODUCTION

1.1 General introduction

One of the significant breakthroughs in science today is the elucidation of the hereditary mechanisms and the manipulation of deoxyribonucleic acid, or **DNA** - the unit of transmission, recombination and function (Vogel and Motulsky, 1997). After the birth of genetics by Gregor Mendel in 1865, Mendel's laws were rediscovered in 1900 and two years later, Sutton and Boveri formulated the chromosome theory of inheritance. It was not until 1944 that Oswald Theodore Avery and co-workers were able to show that DNA is the genetic material and rapid advances have since been made both in the theoretical and technological aspects of this science. Watson and Crick presented the double helical structure of DNA in 1953 and in 1961 Nirenberg and Matthaei initiated the study to decipher the genetic code. These crucial discoveries were due to the development of a whole new methodology referred to as **Molecular Genetics** and have provided novel means of exploring new frontiers of knowledge.

Today, genetics is a dynamic science, recognised as the very core of modern biology. Molecular genetics has become an important tool in research and application as current interest in human molecular genetics is towards mapping and characterising the entire human genome. Acquiring this information will assist a great deal in understanding the normal and pathological functions of the genetic material, both at the individual as well as at the population level.

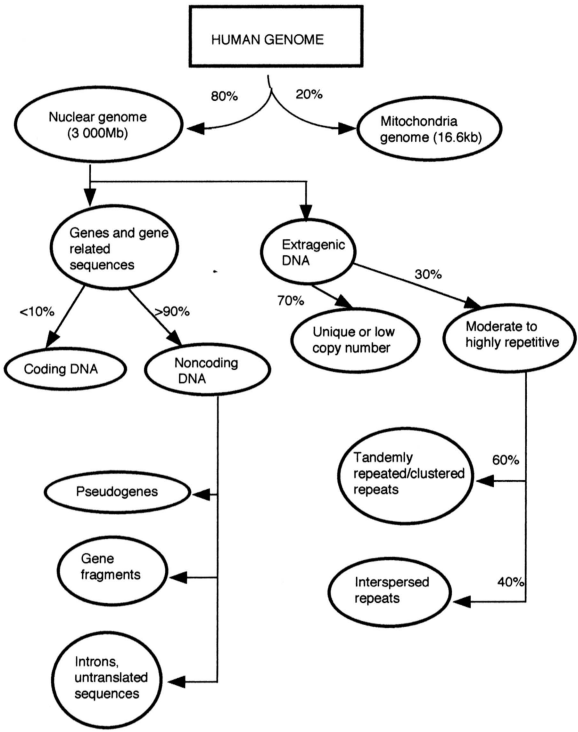


Figure 1: Organization of the human genome. (Modified from Strachan, 1994.)

1.2 Genome organisation

1.2.1 The human genome

The human genome consists of a linear arrangement of nucleotides — a basic repeat unit made up of a five-carbon sugar, a nitrogenous base and a phosphate group — forming the DNA molecule. The functional DNA molecule comprises two polydeoxyribonucleotide strands, intertwined to form a double helix that is tightly packed into structures called chromosomes in the cell nucleus. There are four types of nucleotides found in DNA and each nucleotide carries a different base, namely, adenine (A), guanine (G), cytosine (C) and thymine (T). Genetic information lies within the DNA strand as codes in the form of three-base signals. The information is transcribed, translated and undergoes several other processes within the cell to form a polypeptide or, in other words, a functional protein. DNA segments that contain information or encode specific RNA molecules and proteins are thus referred to as genes. Genes are situated in discrete regions, commonly referred to as genetic loci, on the 46 chromosomes. A single genetic locus may have different alleles but each individual can have no more than two alleles at a given locus. However, a population may have multiple alleles at any given locus and this variation is termed genetic polymorphisms, which have been the crucial basis of DNA typing systems.

The entire nuclear genome of a human haploid cell has approximately $3x10^9$ bp, which are distributed among 22 autosomes and a sex chromosome. It is estimated that only 2-5% of the $3x10^9$ nucleotide pairs represent coding sequences specifying for about 65,000 - 80,000 genes (Fields *et al.*, 1994). Only a small fraction of the human genome is directly involved in specifying a polypeptide or a functional RNA product, such as, mRNA, rRNA and tRNA (Figure 1). The great

majority (97-98%) of DNA in the genome are noncoding and extragenic DNA, and for most of them, their functions are not yet known (Strachan, 1994). Some noncoding sequences are part of transcriptional units; for example, the nontranslated flanking regions of the mRNA, or the introns that are spliced during mRNA processing. Others might be involved in regulatory mechanisms, providing signals recognised by proteins, such as, promoter and enhancer elements for regulating transcription. In addition, certain extragenic and noncoding DNA sequences possess specific chromosomal function and examples of these are telomeric and centromeric sequences. Telomeres are the initiation sites for pairing of homologous chromosomes and they are also involved in ensuring that replication of DNA at the chromosome termini is complete, while centromeres ensure accurate disjunction of the chromosomes into daughter cells during mitosis and meiosis.

1.2.2 Repetitive DNA

While some of these nonfunctional DNA segments are unique sequences, occurring only once per haploid genome, other DNA stretches are present as repeated segments or repetitive sequences. Repetitive DNA regions are very common in most vertebrate and plant genomes and change in the copy number of repeats generally occurs at a much higher rate than change due to point mutations (Hoelzel, 1995). Britten and Kohre had first discovered the existence of repetitive noncoding sequences in eukaryotic genomes in the 1960s (Wolfe, 1993). They developed a method, now known as reassociation kinetics, for detecting repetitive DNA sequences in bulk DNA samples. They concluded that all eukaryotes have three classes of DNA sequence elements: unique sequences occurring in only one copy per haploid genome, moderately repetitive sequences in a few to hundred thousand copies and highly repetitive if they occur in hundreds of thousands to

millions of copies (Wolfe, 1993). Nonfunctional repeated DNA sequences can occur either as individual repeat units interspersed within other DNA sequences or as arrays of tandem repeats. The latter can be sub-divided according to the average size of the tandem repeats into satellite, minisatellite and microsatellite DNA (Table 1).

Table 1: Classes of tandemly repetitive DNA sequences (adapted from Strachan, 1994).

Class	Size of repeat unit (bp)	Total number of copies of repeat units	Major chromosomal location(s)
Tandem repeats			
Satellite DNA			Heterochromatin
Simple sequence	5-25[a]	?	-of 1q, 9q, 16q, Yq
Alpha (alphoid DNA)	171[a]	8×10^5	-of centromeres
Beta (*Sau*3AI family)	68[a]	5×10^4	-of 9, 13, 14, 15, 21, 22
Minisatellite DNA			
Telomeric family	6	$2 \times 10^4 - 3 \times 10^4$	-Telomeres
Hypervariable family	9-64	3×10^4	-All chromosomes, often near telomeres
Microsatellite DNA			
$(A)_n/(T)_n$	1	10^7	-All chromosomes
$(CA)_n/(TG)_n$	2	7×10^6	-All chromosomes
$(CT)_n/(AG)_n$	2	3×10^6	-All chromosomes

[a]Higher order periodicities may also be observed.

Satellite DNA is a highly repetitive DNA. The repeated unit is relatively homogenous within each species, but major differences are observed between related repetitive DNAs in different species, even of the same genus (Smith, 1976).

In recent years, minisatellites and microsatellites have been of particular interest in molecular genetic studies. Comparison of the repeat unit sequences revealed that there is quite a significant level of similarity between repetitive regions (Smith, 1976). These regions are thus prone to errors such as unequal recombination between misaligned satellites (Smith, 1976), or by slippage at replication forks

during cell division (Kunkel, 1993). A misalignment of DNA strands will cause the molecule to be repaired in such a way that may result in a gain or loss in the array of repeated elements. Consequently, high variation of repeat length between alleles may be observed, making some of these loci highly polymorphic. Such regions are termed hypervariable regions (HVR), and in almost all cases, the resulting polymorphisms are due to variable number of tandem repeats, or VNTR for short (Nakamura *et al.*, 1987).

VNTR are regions of hypervariable DNA made up of tandem repeat sequences and are multiallelic in nature. The hypervariability of VNTR produces a high level of heterozygosity in the population, and VNTR are thus useful genetic markers.

Several functions have been suggested for minisatellite DNA sequences, but their role *in vivo* remains to be clarified. Some minisatellite sequences have been shown to bind protein *in vitro* and promote homologous recombination, while others, most notably microsatellites such as trinucleotide repeats, have been implicated in human diseases (Kunkel, 1993). Correlations have been found between the instability of trinucleotide repeats and the onset of genetic disorders such as fragile X syndrome and myotonic dystrophy (Kunkel, 1993). For myotonic dystrophy, normal individuals are expected to have 5-25 repeats in the untranslated 3' portion of the myotonin protein kinase gene; mild or marginal cases have up to 100 repeats, and beyond that the full mutation is observed. The greater the expansion, the worse the disease (Brook *et al.*, 1992).

1.2.3 Hypervariable loci

As mentioned earlier, several regions displaying high levels of DNA polymorphism have been identified in various parts of the genome. They are known

as HVR which promise to be highly informative for genetic analysis. The roles they play in the regulation of the human genome are not readily defined. These loci are termed hypervariable owing to the fact that in any given population, there is a very large number of distinct form of genes, or alleles, present at a given locus (Devlin *et al.*, 1990). Polymorphisms detected in these HVR are mainly due to variability in the number of tandemly repeated units as opposed to base or sequence modifications. Investigators have identified more than 2,500 polymorphic loci in the human genome (Lee *et al.*, 1994).

Wyman and White (1980) gave the first description for polymorphisms of this nature. They identified an arbitrary region showing significant frequency of DNA sequence variation. With the use of recombinant repetitive DNA probes to screen a library of 15 to 20 kb segments of the human genome, they successfully defined at least 16 distinct allelic combinations per chromosome (Wyman and White, 1980; Jeffreys *et al.*, 1985b; Balazs *et al.*, 1989). Subsequently, several other highly variable regions were discovered in which the frequencies of heterozygosity may be in the order of 80 to 100%, thus providing invaluable markers for genetic analyses (Jeffreys, 1987; Jeffreys and Morton, 1987).

1.3 Genetic polymorphisms and genetic studies

1.3.1 Conventional source of analysis

The basis of almost all genetic analyses is genetic markers. Classically, the presence of genes has been recognised by the variability of their effects, that is, only polymorphic traits (genes) are readily detectable. Researchers exploit human biochemical markers which can be detected serologically or by gel electrophoresis. These are mainly polymorphic protein markers, e.g., the ABO blood group

polymorphisms first used in 1900 (Bernstam, 1992). Other examples of such markers are polymorphisms in red blood cells (e.g., haptoglobin, HLA antigens), red blood cell enzymes (e.g., adenylate kinase, phospho-glucomutase, adenosine deaminase), white blood cell antigens and plasma, which have been shown to exist in allelic forms transmitted according to the Mendelian laws of inheritance (Bernstam, 1992). However, progress is limited owing to lack of informative polymorphic protein compounds for study.

1.3.2 Restriction fragment length polymorphisms (RFLP)

In 1970, Smith and co-workers isolated the first restriction endonuclease — a bacterial enzyme having the property of cutting DNA into fragments at specific recognition sites called restriction sites, probably as a defence mechanism against invading foreign DNA (Gardner *et al.*, 1991). Digesting human DNA with the enzyme will reveal restriction fragments whose lengths are considerably variable in the population, termed restriction fragment length polymorphisms (RFLP). These polymorphisms are normally due to base substitution, microdeletion, insertion, or rearrangement that will create, eliminate or rearrange a cleavage site.

RFLP were first used as a tool for genetic analysis in 1974 (Botstein *et al.*, 1980). The fragments produced by the enzymes can be separated with high resolution by electrophoresis in agarose or polyacrylamide gels (Southern, 1975). Molecular analysis of chromosomal DNA by using restriction endonuclease digestion revealed that RFLP occur frequently (Jeffreys *et al.*, 1993) and the informative potential of DNA polymorphisms by far exceeds that of protein polymorphisms. This is due to the fact that protein polymorphisms arise from the

coding DNA regions which occupy only a small portion of the genome, whereas DNA polymorphisms are distributed throughout the entire genome (Jeffreys, 1987).

The early survey of the incidence of RFLP in human β-globin gene (Jeffreys, 1979) together with the work of Kan and Dozy (1978) showed that RFLP are common in human DNA. RFLP can also provide a unified approach to developing unlimited numbers of human genetic markers and may offer a new approach to linkage analysis and anthropological studies. This idea was then elaborated by Botstein *et al.* (1980) who defined the strategy and feasibility of global linkage mapping by using RFLP.

By early 1980s, it was clear that RFLP were not only difficult to be detected but also expensive to type, as well as being uninformative in most pedigrees (almost all were diallelic systems) (Jeffreys *et al.*, 1993). A more definitive system with higher levels of heterozygosity and informativeness is needed as markers for genetic analysis.

Numerous restriction endonucleases have been recognised to date — with most having a unique, sequence specific cutting site — and this has greatly facilitated RFLP analysis of the human genome. Nevertheless, as these restriction site dimorphisms can only have two allelic states, and as the mean heterozygosity of human DNA is low (~ 0.001 per base pair), the chance of any restriction enzyme to detect an RFLP at a given locus is slim. Although the probability of detection can increase with the use of enzymes such as *Msp*I and *Taq*I (whose recognition sequences contain the mutable CpG doublet), the informativeness of the dimorphism is limited to a maximum heterozygosity of 50% (Cooper *et al.*, 1985; Jeffreys *et al.*, 1985b).

Through a major international collaboration (NIH/CEPH Collaborative Mapping Group 1992) (Jeffreys *et al.*, 1993), RFLP have made possible the construction of linkage maps of the X-chromosome and the discovery of markers linked to disease loci such as Huntington's chorea and cystic fibrosis (Kunkel, 1993).

1.3.3 HVR as an alternative typing system

Some repetitive regions are highly variable in length and are able to generate high levels of heterozygosity. Moderately sized arrays of tandemly repeated DNA sequences are called minisatellites by virtue of their structural similarity, on a small scale, to classical satellite DNA. Furthermore, VNTR markers show that most minisatellites cluster towards the ends of the linkage map, while some do occur at interstitial locations (Armour *et al.*, 1989).

The length variation can be detected by using any restriction endonuclease that cleaves at sites flanking the repeat elements and not within the repeat sequence. Arising presumably by mitotic or meiotic unequal recombination or by DNA slippage during replication, these repeat units vary considerably in size but share a common core sequence. Unlike minisatellites, the smaller sized repeat units such as microsatellites or short tandem repeats (STR, e.g., dinucleotide and trinucleotide tandem repeats), which are equally highly polymorphic, have been found to be abundant throughout the human genome. The frequency of heterozygosity may reach more than 95%, thus providing a vast source of highly polymorphic alleles ideal as markers for genetic analysis (Wong *et al.*, 1987).

1.4 Minisatellites: application to forensic studies

Before the discovery of polymorphisms at the DNA level, individual identification is based upon serological and biochemical tests that identify protein polymorphisms. However, the advent of DNA polymorphisms has contributed greatly to research in human genetics as genetic polymorphisms can now be detected directly at the DNA level. The diagnostics offer more accurate results compared to protein polymorphisms. DNA polymorphism is then used as an alternative.

In addition to the significant increase of informativeness presented by DNA polymorphisms, sources of DNA for analyses are almost unlimited. Assisted by the advent of the polymerase chain reaction (PCR) (see Section 1.4.3), biological samples can be analysed with greater sensitivity and confidence. DNA polymorphisms have been the basis of genetic variation analysis mainly in the aspects of legal medicine, forensic investigation, parentage analysis, the study of evolution and anthropology, animal (Burke and Bruford, 1987; Jeffreys and Morton, 1987) and plant sciences (Kirby, 1990).

The possibility of using DNA polymorphisms as a basis of individuality was highlighted by Alec Jeffreys in an identification technique he termed "DNA fingerprinting" (Jeffreys *et al.*, 1985c).

1.4.1 DNA fingerprinting

In his work on the evolution of human gene families, Jeffreys stumbled upon a short tandem repeat region in an intron of the human myoglobin gene (Jeffreys, 1979). Sequence analysis of this minisatellite revealed some similarities to other known minisatellites (Weller *et al.*, 1984). This suggested that minisatellite

sequences are related and not random. Further studies were conducted, which involved the preparation of a probe from the human myoglobin minisatellite to detect other cross hybridising regions of the human DNA. A pure repeat probe, constructed by head to tail ligation of 23 copies of a 33 bp repeat element, was used against human DNA digested with restriction endonuclease *Hin*fI or *Hae*III. Multiple DNA fragments of variable lengths were detected following hybridisation and autoradiography. These minisatellite length variants were observed to be transmitted in a Mendelian fashion, in that each polymorphic band in the daughter can be identified within one or the other parent (Jeffreys *et al.*, 1985a).

Extensive studies were done by further hybridising the 33 bp repeat probe to a library of 10-20 kb *Sau*3A fragments of human DNA cloned in phage λL47.1 (Jeffreys *et al.*, 1985b). In that study, 40 clones containing various minisatellite regions were identified. Eight were randomly selected and subsequently subcloned into M13 cloning vectors for further analysis. Probes were prepared from each minisatellite clone and Southern hybridisation was performed against a panel of 14 unrelated human DNA samples digested with *Hin*fI. Four of the clones showed considerable variability and three of them were observed having between 5-8 alleles per locus.

Sequence analysis of hybridising regions showed that each of the selected clones contains a minisatellite of variable copy number but sharing a 10-15 bp core region of almost invariant sequence (Jeffreys *et al.*, 1993). Most minisatellites contain integral copy number and none of them are flanked by direct repeats thus excluding transposition as the origin of minisatellites. The function of the shared core sequence is unknown, but it has been suggested that it may have a role similar to that of the χ sequence in *Escherichia coli* — a signal for recombination and

possibly promotes the formation of minisatellites (Jeffreys *et al.*, 1985b). Despite the ambiguity of the role of minisatellites, it is now known that this sequence similarity can be used as a basis to detect other hypervariable regions.

Simultaneous detection of many hypervariable loci will generate an individual-specific DNA 'fingerprint'. This offers a wide range of applications in human genetic analysis, particularly in the establishment of family relationships and forensic identification, as well as applications to non-human species (Jeffreys *et al.*, 1993). Each individual pattern is unique with the exception of a DNA fingerprint shared by monozygous twins (because they originate from a single fertilisation). It is estimated that the probability of 2 unrelated individuals sharing the same 36 bands is $2x10^{-22}$. Similarly, the chance that these two individuals have identical DNA fingerprints can be estimated at 4×10^{-30}. Furthermore, the probability that two first degree relatives, for example siblings, are identical is about 3×10^{-14} (Jeffreys, 1987).

The DNA fingerprinting technique of individual identification has already been established and accepted in court cases concerning forensic investigations, family analysis and immigration casework. In 1987, Cellmark Diagnostics was established in the UK and USA to provide commercial DNA fingerprinting diagnostic services to the world.

1.4.2 DNA profiling

One of the drawbacks of DNA fingerprinting is that the multiple band patterns are often too complex and difficult to interpret. To overcome the complexity of DNA fingerprinting, Jeffreys and co-workers introduced single locus DNA profiles, aptly termed DNA profiling, in 1987 (Jeffreys, 1987). Single locus

DNA profiling offers a better detection system with greater sensitivity and results of this method are easier to interpret.

The probability of two unrelated individuals sharing the same DNA profile at any one site might be in the order of 5 in 100. The chance occurrence of a specific genotype will diminish significantly with each additional locus tested (Lee *et al.*, 1994). These patterns are not individual specific but most of these loci are sufficiently variable that they enable the determination of alleles inherited from each parent, thus providing highly informative genetic markers for linkage and segregation analyses.

Another advantage of DNA profiling is that the probe sensitivity is increased since each probe will hybridise to a single locus. In addition, DNA profiling is more applicable towards degraded DNA because the probes are locus specific, instead of whole genome hybridisation involved in DNA fingerprinting. Biological samples for forensic casework such as blood stains, semen swabs or hair roots are sometimes recovered in less than ideal conditions, which may result in difficulty to obtain fully resolved fingerprints. With the advent of the polymerase chain reaction (see below), even more minute quantities of DNA can be analysed if the target sequences are amplified prior to analysis (Lewontin and Hartl, 1991).

In addition to their use to provide evidence in courts, DNA profiling techniques have also been applied in linkage studies of inherited disorders, detection of allele losses from tumours (Vogelstein *et al.*, 1989), and detection of abnormalities from chromosomal aberrations (Malcolm *et al.*, 1991), such as partial trisomy and uniparental disomy (Frezal and Schinzel, 1991). However, their use in general linkage analysis is limited owing to uneven distribution in the genome.

Although minisatellites appear to be equally represented on all human autosomes, most hypervariable loci have been localized to subtelomeric regions.

1.4.3 Polymerase chain reaction (PCR) based approaches

The polymerase chain reaction (PCR) is a technique that enables rapid amplification of a specific region of DNA (Figure 2). The basic steps involve hybridisation of short oligonucleotides (primers) to a template DNA strand to initiate synthesis, followed by primer extension along the template to form a complementary strand, driven by an enzyme, namely DNA polymerase. Repeating the process will result in exponential amplification of the desired DNA region.

Forensic DNA analysis was revolutionised with the advent of PCR driven by *Taq* DNA polymerase by Cetus in 1988 (Mullis *et al.*, 1986; Saiki *et al.*, 1988; Mullis, 1990). This has allowed supersensitive DNA typing to be applied to crime-scene samples with very little or degraded DNA. Examples of crime-scene samples which can be used for forensic casework include single hair root, small blood stain, finger nail, semen samples, and even saliva recovered from postage stamps (Lewontin and Hartl, 1991; Lee *et al.*, 1994). The amount of DNA needed is minimal as PCR can be applied to amplify a DNA segment from just a single cell (Jeffreys *et al.*, 1988b).

Although PCR is generally used to amplify short regions of a few hundred base pairs, it has been shown that the entire human minisatellite can be amplified, both singly and in a multilocus format by multiplex PCR (Jeffreys *et al.*, 1988b).

PCR has other wide applications in addition to forensic casework and constitutes one of the significant developments in molecular biology. PCR has transformed molecular biology, enabling revolutionary advances in many scientific

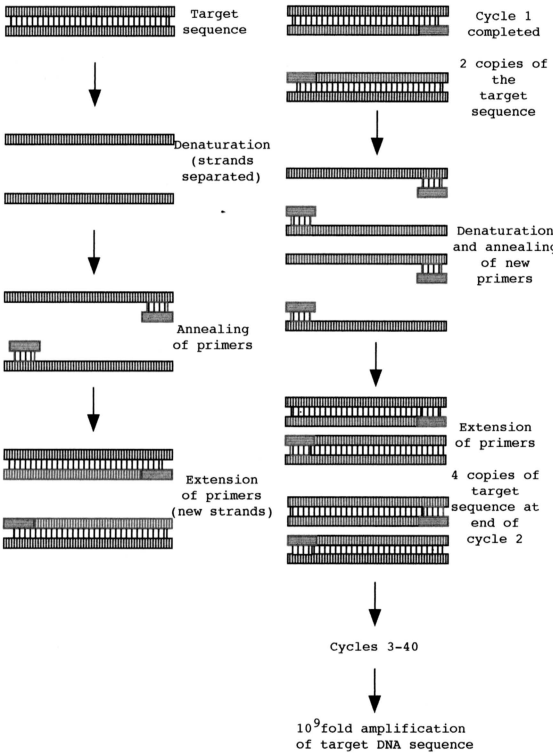


Figure 2: The polymerase chain reaction (PCR). (Modified from Mullis, 1990; Cooper and Krawczak, 1993.)

disciplines, and has proved to be an invaluable tool in biological research as well as in the diagnosis of genetic disorders and infectious diseases (Lee *et al.*, 1994).

1.4.4 Minisatellite repeat unit sequence variant

DNA fingerprinting and DNA profiling are now established techniques, used routinely to assist criminal and civil investigations. DNA fragment length is a mere estimation from comparison of a band position on an autoradiograph with another band or a reference marker ladder. Precise allele identification is impossible, and bands are not readily assigned to their respective loci. The method employed to interpret the resulting DNA band patterns (comparison of band position on a gel) has been the subject of many arguments (Lander, 1989).

This has prompted researchers to carry out detailed analysis of the minisatellite sequence content with the hope to establish a more reliable and accurate DNA typing system. In 1990, Jeffreys and colleagues described the development of a new DNA typing technique, which claimed to combine the variability of minisatellites with the speed and sensitivity of PCR. The technique relied on detecting sequence variations specific to particular alleles, and promises higher accuracy and sensitivity in determining individual fingerprints.

The technique was termed digital DNA typing or minisatellite variant repeat by PCR or MVR-PCR. Since the method of detection is based on allelic variation in the interspersion patterns of variant repeat units along minisatellites, this eliminates possible measurement errors of the conventional DNA typing systems which are based on allele length and subjected to electrophoretic 'band-shift'.

Sequence analysis of cloned minisatellites revealed that repeat units within a minisatellite usually display some level of variation in sequence. Owerbach and

Aagaard (1984) sequenced three different cloned hypervariable locus 5' to the insulin gene and found up to nine different types of repeats per allele, differing slightly from the consensus repeat sequence. Other minisatellite loci analysed, such as the loci near the c-Ha-*ras* oncogene (Capon *et al.*, 1983) and α-globin gene clusters (Jarman and Wells, 1989), also exhibit repeat unit sequence heterogeneity. Therefore, minisatellite variant repeat analysis has provided new insights into the population genetics of hypervariable loci and allows the investigation of a higher level of allelic variability at minisatellite loci (Jeffreys *et al.*, 1991).

Allelic variation may be observed in the interspersion pattern of subtly different types of repeat unit along the tandem array (Jeffreys *et al.*, 1991, 1993). Repeat unit specific primers designed to prime from variants of the repeat sequence will generate a set of products extending from each repeat unit of a particular type into the flanking DNA. Resolving these products by gel electrophoresis will enable the type of repeat variant within a minisatellite to be determined unambiguously. By assigning a suitable digital code to represent the array of repeat variants, a 'fingerprint' approaching complete individual specificity could be achieved.

MVR-PCR was initially studied at the hypervariable locus D1S8 (MS32) (Jeffreys *et al.*, 1991) followed by the hypervariable locus D7S21 (MS31) (Neil and Jeffreys, 1993) on Caucasian and Japanese populations and the hypervariable locus D16S309 (MS205) on African and non-African populations (Armour *et al.*, 1996). The study of MVR PCR on both hypervariable minisatellite loci MS31 and MS32 simultaneously will overcome the limitation by similarity diploid codes from unrelated individuals. This could reduce the possibility of unambiguous individual identification, particularly from the first few repeat units in badly degraded DNA.

One of the crucial factors to ensure the success of MVR-PCR is the ability to design allele specific primers. For this purpose, much effort has been channelled towards characterising the sequence content upstream of the minisatellite region (5' flanking region).

The identification of polymorphic positions upstream of D7S21 was described by Neil and Jeffreys (1993). Two sites of base substitutional polymorphisms were revealed in their study. The first is an A/G transition, located 4 bp 5' to the first minisatellite repeat, which gives rise to an *Alu*I RFLP. The second is a C/G transversion 220 bp from the minisatellite generating an *Hga*I polymorphic site (see Section 1.5).

MVR-PCR analysis provides evidence that most alleles with the same length show different internal structures, and suggested that unequivocal allele identification will then be possible (Figure 3). Conversion of allele specific MVR-PCR banding pattern into binary codes will generate informative unambiguous digital codes from human DNA (Jeffreys *et al.*, 1993). From the preliminary data available, it has been suggested that at least 10^8 different alleles are identifiable from within the MS32 locus alone. Thus, perhaps for the first time ever, the true level of allelic variability can be achieved.

This finding is a major breakthrough in allelic variability research, considering only a mere 100 or so alleles distinguishable by allele length. It has great potential for application in forensic medicine as well as allowing indepth exploration of allelic variability and mutation process at minisatellites (Jeffreys *et al.*, 1991, 1993; Monckton *et al.*, 1993).

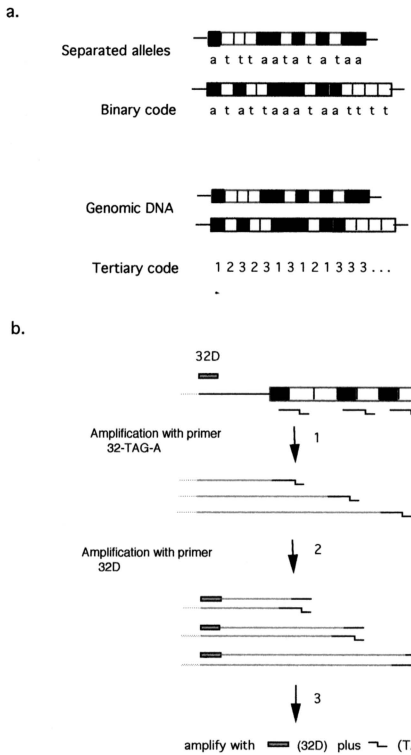


Figure 3: The principle of minisatellite repeat coding. a. Two, a-type and t-type, variant repeat units. Individual alleles can be encoded as a binary string extending from the first repeat unit. In total genomic DNA a corresponding tertiary code of both alleles superimposed can be generated. b. The principle of MVR-PCR, illustrated for a single allele amplified by using primer 32-TAG-A. TAG primer is used to prevent internal priming. (Modified from Jeffreys *et al.*, 1991.)

1.4.5 Other PCR based DNA typing methods

Several other DNA typing systems have been established to complement fingerprinting and profiling techniques. They include human leukocyte antigen (HLA)-DQα reverse dot blot analysis, amplified fragment length polymorphism (AMP-FLP) system, STR (short tandem repeats) typing, mitochondrial DNA analysis (Lee *et al.*, 1994) and expression PCR (Thornton and Rashtchian, 1992).

HLA-DQα reverse dot blot is a technique of individual identification based on the analysis of nucleotide polymorphisms present within the HLA DQ-alpha gene (Helmuth *et al.*, 1990). In this technique, allele-specific variation is detected at the level of single nucleotide substitutions by a 'reverse dot-blot' method. There are eight possible variants of this gene in humans and every individual carries two of these variants, one inherited from each parent. The success of this method very much depends on the ability to differentiate the types of HLA-DQα between samples that are under investigation. Allele-specific oligonucleotide (ASO) probes are used in this assay, and they are bound to nylon membrane strips. The PCR products are produced by amplification with primers tagged with biotin and are then hybridised to immobilised ASO probes. A reaction with streptavidin-horseradish peroxidase conjugate and peroxidase will reveal a blue precipitate. The DQα genotype is determined by the pattern of blue dots that developed (Helmuth *et al.*, 1990; Koh and Benjamin, 1993). Initially, the eight HLA-DQα alleles (1.1, 1.2, 1.3, 2, 3, 4.1, 4.2, and 4.3) were detected as six different alleles (1.1, 1.2, 1.3, 2, 3, and 4) by the DNA amplification and typing kit from Cetus/Perkin Elmer. Now, the new DNA amplification and typing kit from Perkin Elmer differentiates the eight HLA-DQα alleles into seven different alleles (1.1, 1.2, 1.3, 2, 3, 4.1, and 4.2 or 4.3).

AMP-FLP analysis is a PCR-based DNA-typing strategy that is conceptually analogous to RFLP. The AMP-FLP loci contain a large array of alleles that differ in the number of tandem repeats they possess. The analysis does not require restriction endonuclease digestion. The alleles amplified are considerably shorter than the fragments generated from the standard RFLP procedure and are often resolved on polyacrylamide gels. However, electrophoresis in high-percentage agarose gels is also possible. Among the chromosomal loci that are used for forensic application include D1S80, D17S5 and the 3' hypervariable domain of ApoB (Lee *et al.*, 1994).

In STR or microsatellite analysis, short repeated DNA sequence is analysed. STRs exhibit substantial polymorphism owing to variation in the number of repeated core elements. More than 1,300 STR with heterozygosities of >70% have been identified (Lee *et al.*, 1994). STR alleles are relatively short (generally, 100-500 bp) and hence are easily amplified by PCR and identified on polyacrylamide gels. Consequently, for improved discrimination between individuals, several STR are simultaneously amplified in the same reaction (multiplexing) and the PCR products resolved in a single lane, provided that the alleles from different loci do not overlap in size. The advantage of multiplexing is that it can greatly reduce the cost and time for genotype determination. The amplification products can be radiolabeled or stained with ethidium bromide or silver (Lee *et al.*, 1994). In recent years, fluorescently labelled PCR products are produced, separated, and automatically analyzed by the appropriate software.

Mitochondrial DNA (mtDNA) variation is often used in studies involving human evolution, degenerative diseases, and aging. The mtDNA exists extrachromosomally and it presents in multiple copies per cell. It is an alternative

system to nuclear DNA. The mitochondrial genome is similar to prokaryotic DNA and it is maternally derived and does not recombine during meiosis. Most mtDNA-typing systems exploit the highly polymorphic displacement loop (D-loop) region, where investigators have successfully typed the mitochondria's only microsatellite locus – a dinucleotide repeat motif from the D-loop region by PCR amplification and capillary gel electrophoresis. Mitochondrial DNA typing has also been achieved by direct DNA sequencing or by a dot-blot method with allele-specific oligonucleotide probes (Lee *et al.*, 1994).

Expression PCR is a methodology that permits rapid analysis of gene products (protein) without the need for *in vitro* cloning and expression. In this procedure, the transcription and translation initiation signals are incorporated into the amplified product during PCR. The resulting PCR-amplified DNA can then be used for *in vitro* transcription with T7 RNA polymerase followed by *in vitro* translation. The amplification products may be rapidly cloned for future reference and more detailed analyses (Thornton and Rashtchian, 1992).

1.4.6 MS31

The minisatellite locus D7S21 was initially characterised by Wong and co-workers (1987), who found that this locus is very polymorphic. The very high level of allele length polymorphism is the result of variability in the number of repeat elements, producing a heterozygosity level estimated to reach 99%. As previously mentioned, the polymorphic alleles are transmitted in Mendelian inheritance and can be confirmed by segregation analysis.

This hypervariable minisatellite has been mapped within the 7p22-pter region (Figure 4). The original MS31 clone, obtained by *Sau*3AI digestion,

contained the D7S21 locus in addition to 405 bp of 5' flanking DNA, followed by a large number of minisatellite repeats (MS31A) separated by 15 bp from an adjacent minisatellite (Figures 5 and 6). Almost all of the length variation at D7S21 is due to repeat copy number variation at MS31A and all repeat units so far characterised are 20 bp long (Jeffreys *et al.*, 1988a; Royle *et al.*, 1988).

The earlier MVR-PCR experiments required the separation of single alleles by electrophoresis. However, the technique was improved by designing allele specific primers that would allow allele-specific MVR-PCR (Neil and Jeffreys, 1993). Analysis of the 5' flanking DNA revealed two sites of base substitutional polymorphism that would suit this purpose. The first is an A/G transition, located 4 bp 5' to the first minisatellite repeat (-4A/G, Figure 5), and this gives rise to an *Alu*I RFLP. The second is a C/G transversion, 220 bp from the minisatellite (-220G/C, Figure 5), generating an *Hga*I polymorphic site.

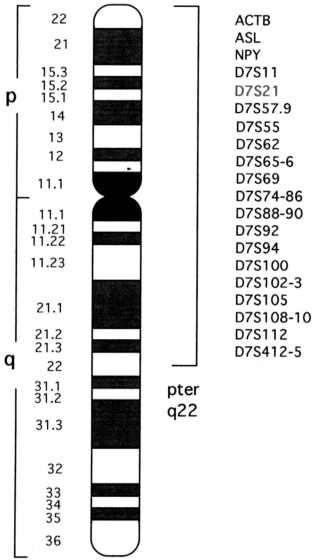
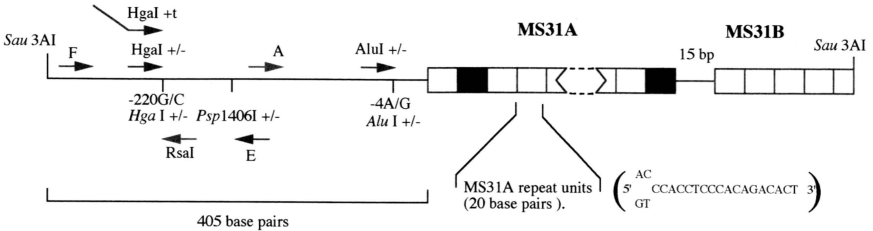


Figure 4 : G-banding pattern of human chromosome 7. (O'Brien, 1990).



PCR primer sequences 5' - 3'

31A	CCCTTTGCACGCTGGACGGTGGCG
31F	CCACTCGGAACCACCTGCAG
31RsaI	GGCAGGAAACCCCAGCAGGCCGTA
31E	TGGACAGCCAAGGCCAGGTCCC

Allele specific primers

31AluI+	CATGAAGGGGACTGGCCTT**A**
31AluI-	CATGAAGGGGACTGGCCTT**G**
31HgaI-	CCTCCCCCACTCAGC**C**
31HgaI+	CCTCCCCCACTCAGC**G**
31HgaI+t	aatacgactcactatagttaCCTCCCCCACTCAGCG

MS31A MVR-PCR PRIMERS

A
3' GGTGGAGGGTGTCTGTGAggcctggtacctgcgtact 5'
G

31-TAG-A

31-TAG-G

TAG PRIMER

3' aggcctggtacctgcgtact 5'

Figure 5: Organization of the D7S21 locus (modified from Neil and Jeffreys, 1993). Variant repeats at MS31A are indicated by filled and empty boxes; dashed lines indicate the presence of different numbers of repeat. PCR primer sites, flanking polymorphic sites, universal and allele-specific primer sequences are also shown.

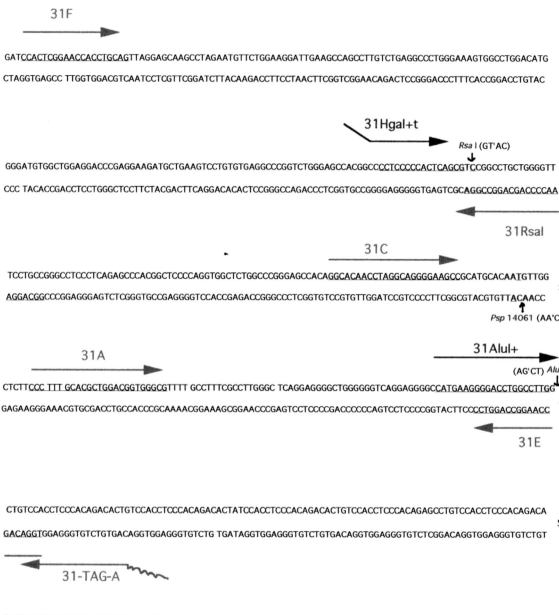


Figure 6: Base sequences for locus D7S21 (MS31A at 5' end) (Neil and Jeffreys, 1993).

1.5 Objective of study

The study of DNA flanking the hypervariable minisatellite at locus D7S21 (MS31A) was initially performed by Neil and Jeffreys in 1993 for Caucasian and Japanese populations. In their study, two RFLP sites located upstream of the minisatellite were investigated. They found that the *Hga*I site and *Alu*I site are common in both populations. Later, the Malaysian population was studied by the same approach (Koh *et al.*, 1993, 1994).

The present research project is a continuation of the study reported by Koh *et al.* (1993, 1994) with a larger population and an additional RFLP *Psp*1406I site located 108 bp to the first minisatellite repeat (-108C/T, Figures 5 and 6).

The method used in the studies mentioned above was based on a simple PCR assay. A suitable primer set was designed to amplify the desired region containing the particular restriction site. Following PCR amplification, the product was subjected to a restriction endonuclease digestion, and the presence or absence of a cutting site was then determined by agarose gel electrophoresis.

Haplotype analysis was carried out for heterozygous samples in both *Hga*I and *Alu*I assays, as well as heterozygous samples in both *Alu*I and *Psp*1406I. This allowed the phase for each polymorphic site to be determined. Statistical analysis was then performed for flanking DNA polymorphism assays and haplotyping assays. Sequencing was done to confirm the size and nucleotide sequence of the fragment in irregular samples from *Hga*I assay.

The significance of this project is to provide information of variations that exist in flanking DNA region of MS31A in the Malaysian population. Later, for heterozygous individuals, single alleles can be analysed directly from genomic

DNA by means of allele-specific MVR-PCR in which allele-specific primers directed to heterozygous sites in the flanking DNA will be used.