RESTRICTION FRAGMENT LENGTH POLYMORPHISMS IN SOUTHERN AFRICAN POPULATIONS AND A SEARCH FOR LINKAGE TO TYROSINASE-POSITIVE OCULOCUTANEOUS ALBINISM

Ruth Alice Heim

A Thesis submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfillment of the requirements for the Degree of Doctor of Philosophy.

Johannesburg, 1988

DECLARATION

I declare that this thesis is my own, unaided work. It is being submitted for the degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg, and has not been submitted for any degree or examination in any other University.

R Heim

Ruth Alice Heim

March 1988

To my parents, Peter and Evelyn Heim

and to my sister, Helen de Jager.

ABSTRACT

In this study, new polymorphic genetic markers were identified and investigated in three major southern African populations, the San, Caucasoids and Negroids. They will contribute, it is proposed, towards the construction of a human linkage map, and will provide useful anthropological information.

RFLPs were screened for with single-copy probes isolated from human genomic libraries. The probes λRH52, λMR7 and λMR22 identified base substitution-type RFLPs, and λMR3 identified a locus at which there was a 300 bp insertion/deletion, giving rise to a re-arrangement-type RFLP. The probes were tentatively assigned by somatic cell hybridization and by linkage with serological markers, to chromosomes 9, 8, 6 and 17 respectively. All RFLPs identified by these probes were bi-allelic and the alleles were inherited in an autosomal codominant fashion. Frequency data and statistics summarising the patterns of frequency variation found were obtained. The locus showing the greatest variation was that identified by λMR3.

Genomic heterozygosity was calculated in the Negroid population to be 0.0025, so that 2 or 3 variants would be expected in every 1 000 nucleotides.

Twenty polymorphic markers, including three isolated during the study, were investigated for linkage to TPOCA. There was no suggestion of linkage to MNS, ABO, PGM_1, 6PGD, ACP_1, GPX_1, GLO_1, GPT_1, PEP A, Tf, Gc, α_1-AT, Hp, λRH52, λMR3, or λMR22. Close linkage to the TPOCA locus was excluded for the Rh, Gc, and β-globin loci. The latter finding indicates that the β-globin and albino syntenic group found in lower mammals may not have been evolutionarily conserved in man, as indirect evidence had previously suggested. The lod score for Bf versus TPOCA is 'interesting' and suggests the possible assignment of TPOCA to chromosome 6. These preliminary results form the basis for an ongoing linkage study.

PUBLICATIONS

Publications arising from this study:

HEIM RA, DUNN DS, CANDY SE, ZWANE E, KROMBERG JGR, JENKINS T (1988). The tyrosinase-positive oculocutaneous albinism locus is not linked to the β-globin locus in man. *Hum Genet* (in press).

HEIM RA, DUNN DS, CANDY SE, ZWANE E, KROMBERG JGR, JENKINS T (1988). A search for linkage to tyrosinase-positive oculocutaneous albinism. (in preparation).

MORRIS DM, HEIM RA, CONN V, JENKINS T (1988). Study of 24 DNA markers in three southern African populations. (in preparation).

ACKNOWLEDGEMENTS

Many individuals willingly donated blood samples for DNA extraction. Without their co-operation the studies reported here could not have been initiated. I am indebted to them all, and especially to Sister Esther Zwane, who located individuals affected with albinism and travelled widely to collect samples from their families.

Financial support for these studies is gratefully acknowledged. Bursaries and scholarships were provided by the Medical Research Council, the University of the Witwatersrand, and the South African Institute for Medical Research. The Director of the South African Institute for Medical Research, Professor J. Metz, kindly made facilities in the Department of Human Genetics available for the research described in this thesis.

I am indebted to and thank the following people:

Janice Gallant for diagrams and Helen de Jager for effort and time spent in assisting with the figures and tables, Mel Anderson, Martha Ulrich and Gill Martin for the photographic prints, and Vazu Naidoo for the lithographs.

Sally Candy for helping to set up the database, David Dunn for invaluable and always good-humoured assistance with computer analyses, and John de Jager for assistance with the printing.

Maria dos Santos and Isobel Marques for their work towards creating a somatic cell hybrid panel, Drs U. Francke, T. Mohandas and A.R. Retief for generously donating hybrid cell lines, and David Dunn, Coleen Morgan and Stephen Hart for the serogenetic data used in the linkage study.

Lastly, special thanks are due to Professor Trefor Jenkins for his supervision and critical evaluation of this manuscript, and to my colleagues for their help and friendship, especially Shirley Alper, Sally Candy, David Castle, Vera Conn, David Dunn, Stephen Hart, Amanda Krause, Jennifer Kromberg, Stephen Lambert, Tony Lane, Douglas Lester, Coleen Morgan, Deborah Morris and Michele Ramsay.

CONTENTS

CHAPTER 3

CHAPTER 4

LIST OF TABLES

LIST OF FIGURES

CHAPTER 5

LIST OF ABBREVIATIONS

ACD	Acid citrate dextrose
bp	Base pair
cDNA	Complementary deoxyribonucleic acid
cM	centiMorgan
cpm	Counts per minute
dCTP	Deoxycytidine-5'-triphosphate
DNA	Deoxyribonucleic acid
EDTA	Ethylene-diamine-tetra-acetic acid
g	Gram
HBV	Hepatitis B virus
hepes	N'-2 ethanesulfonic acid
kb	Kilobase pair
mA	milliAmps
mg	Milligram
ml	Millilitres
mM	milliMolar
n.m	Nanometres
OCA	Oculocutaneous albinism
pfu	Plaque forming units
poly A	Polymeric deoxyadenosine mon phosphate
pol III	DNA polymerase III
PPi	Pyrophosphate disodium dihydrogen
PVP	Polyvinylpyrrolidone
RFLP	Restriction fragment length polymorphism

RNA	Ribonucleic acid
SDS	Sodium dodecyl sulphate
TCA	Trichloroacetic acid
TNOCA	Tyrosinase-negative oculocutaneous albinism
TPOCA	Tyrosinase-positive oculocutaneous albinism
Tris	Tris(hydroxymethyl)methylamine
µg	Microgram
µl	Microlitre
UV	Ultra violet
VNTR	Variable number of tandem repeats

CHAPTER 1

1. INTRODUCTION

1.1 Background

The human haploid genome is thought to comprise 3 x 10^9 nucleotides (Bishop, 1974). This amount of DNA is 1 000 times that of the prokaryotic genome, *Escherichia coli*, which contains 3.2 x 10^6 nucleotide pairs (Cairns, 1963). Most of the DNA in eukaryotes has been termed 'junk' (Doolittle and Sapienza, 1980; Orgel and Crick, 1980) and is found within and between genes. The present study is concerned with DNA analysis that may help to elucidate the relationship between the structural organization of the genome and its function.

The aims of the study were conceived during a period of excitement in human genetics. A concerted effort to create a linkage map of the human genome had been initiated some years previously (Botstein *et al.*, 1980), and this goal seemed both worthwhile and attainable. A precise knowledge of the human gene map is a pre-requisite for understanding each gene's function and mode of expression, and also for revealing the genetic history of man. Furthermore, a map yields new approaches to genetic disease and has considerable medical and social significance. One of the primary aims of the present study was to contribute new poly-

morphic genetic markers towards the construction of a human linkage map.

There are unique opportunities in southern Africa for studying indigenous populations, and the investigation of genetic variation within and between populations was included in the broader aims of the study. The initial search for new polymorphic markers was to be carried out in the southern African Bantu-speaking Negro population and then extended to the San and southern African Caucasoids.

Oculocutaneous albinism (OCA) is a genetic disorder known to be common in the local Negroid population. Psychosocial and genetic aspects of this disorder have been investigated (Kromberg, 1985), but the molecular basis of albinism is not known and no OCA loci have been placed on the human genetic map. This lack of knowledge motivated the writer to initiate a search for linkage to the locus for the tyrosinase-positive form of OCA, which is one of the commonest inherited conditions in the South African Negroid. Genetic markers used for the search would include those isolated during the study.

1.2 Genetic variation

Genetic variation can be measured by the extent of polymorphism within genomes and between populations. The complexity of the human genome is evident in the organisation of repeated and unique DNA sequences, and it is often repetitive sequences which underly DNA sequence polymorphism.

1.2.1 Repetitive DNA

The repeated nature of nucleotide sequences in eukaryote genomes was first described by Waring and Britten (1966). That single-copy DNAs are interspersed with repetitive DNA sequences (Schmid and Deininger, 1975) is now known to be a general feature of organization in the eukaryote genome. Several kinds of repeat families have been reported, some of which appear to have been at least partly conserved in evolution (Maio *et al.*, 1981; Haynes and Jelinek, 1981; Shafit-Zagardo *et al.*, 1982; Singer *et al.*, 1983), suggesting that they perform an important evolutionary role.

Repetitive DNAs can be broadly grouped into interspersed, multigene and tandem repeat families.

Interspersed repeat families

The origin and evolution of interspersed repeat families is only beginning to be understood. Sequences in these families are colinear with RNA transcripts, are followed by a variable (A)-rich or oligo(A) tail (Houck *et al.*, 1979; Singer and Skowronski, 1985), and are flanked by a repeat of 8-19 bp which represents a duplication of a putative insertion site: The term retroposon has been suggested for such sequences, on the supposition that an RNA intermediate is involved in their mechanism of transposition (Jagadeeswaran *et al.*, 1981; Rogers, 1985). Occasional 5' truncation may reflect premature termination of reverse transcription. The sequences also have a tendency to cluster by insertion into pre-existing retroposon tails, and there is evidence that they undergo frequent superposition and rearrangement (reviewed in Rogers, 1985). Hwu *et al.* (1986) suggest that a large number of events of insertion and/or deletion of these DNA sequences has occurred during higher primate evolution.

Interspersed repeat families comprise either long (LINES) or short (SINES) sequences. The well-characterised *Kpn*I or LINE1 family is now known as L1Hs, where L1 stands for the first family of LINES and Hs (*Homo sapiens*) is the animal species in which it occurs or from which a probe has been derived (Singer and

Skowronski, 1985). The family was first studied in human DNA (Adams *et al.*, 1980), and has been extensively studied in primate genomes. It is widely conserved and a related family occurs in mouse DNA (Fanning, 1983; Singer *et al.* 1983).

Alu or SINE 1 is found with very high frequency in human and other primate DNA (Houck *et al.*, 1979). It is about 300 nucleotides long and includes a *Pol* III promoter. *Alu* sequences are not randomly distributed. They have been associated with characterized genes in variable frequencies: five, for example, within 26 kb of the α-like globin cluster (Fritsch *et al.*, 1980) and two flanking the human insulin gene (Bell *et al.*, 1980).

The total number of copies of the L1Hs and *Alu* families in humans has been measured by an accurate titration method (Hwu *et al.*, 1986); there are 910 000 copies of the *Alu* family and 107 000 copies of the L1Hs family. The function of LINES and SINES is unknown. It has been suggested that L1Hs is a multigene family comprising a large number of pseudogenes and some functional genes (Singer and Skowronski, 1985), possibly functioning as transcription enhancers. A putative L1 transcript has been described (Skowronski and Singer, 1985). *Alu* sequences may function as origins of DNA replication (Jelinek, 1978) or have a function

expressed through an RNA product (Haynes and Jelinek, 1981).

Multigene families

Some sequences known to have a coding function are also repeated in the human genome. Examples include the genes coding for the 45 s precursor of ribosomal RNA, repeated 150 to 300 times per haploid genome (Gaubatz *et al.*, 1976; Wellauer and Dawid, 1979); the genes coding for 5 s rRNA and tRNA, repeated at least 2 000 and 1 300 times respectively (Hatlen and Attardi, 1971); and histone genes, repeated 30 to 40 times (Wilson and Melli, 1977).

Tandem repeat families

Satellite DNA sequences constitute many short tandem repeats (Brutlag, 1980) and are concentrated essentially in constitutive heterochromatin. Other tandem repeats include the long tandem repeats found in alphoid sequences (Wu and Manuelidis, 1980; Willard, 1984), and the 319 bp repeats of the *Hinf* family (Shimizu *et al.*, 1983). Some of these sequences have been called evolutionary 'hitchhikers' (Orgel and Crick, 1980). Functional roles have also been ascribed, including a role in discriminating chromosomes from one another (Lee and Singer, 1982; Willard

et al., 1986), a role in nucleosome arrangement (Wu *et al.*, 1983), and a role in homologous chromosome recognition (Brutlag, 1980).

Very short clusters of non-satellite tandem repeats have also been characterised in the human genome. At some loci the number of copies of the short repeat varies within the population, giving rise to minisatellites or hypervariable regions termed VNTRs (Variable Number of Tandem Repeats) (Nakamura *et al.*, 1987b). It is possible that such sequences encode 'hotspots' for recombinational activity, leading to their origin and mutability through unequal exchange. A 'core' sequence (GGGCAGGAXG) seems to be common to nearly all VNTR sequences found by Jeffreys *et al.* (1985a,b) and similar 'core' sequences, (GGG--GTGGGG or the almost invariant sequence G----TGGG) have been found by Nakamura *et al.* (1987b).

Some sequences appear to have a relationship with the *Chi* sequence of phage lambda (GCTGGTGG), which has been implicated as a hotspot for *rec*A-mediated recombination in *E coli* (Lam *et al.*, 1974). Nakamura *et al.* (1987b) have suggested that similarity to *Ch* sequences is due to functional constraints, or viral or extrachromosomal origins for VNTRs. It is interesting that sequences from the X-gene region of HBV can identify VNTR loci, because this region is thought to be the viral site of

integration into the human genome. Perhaps the region is recombinogenic in a similar way to the long terminal repeats of retroviruses (Mizusawa *et al.*, 1985; Yaginuma *et al.*, 1985).

Vassart *et al.* (1987) have now reported that a sequence in the protein III gene of the bacteriophage vector M13 detects a set of hypervariable minisatellites in human and animal DNA. This finding raises the question of the role of hypervariable regions and their maintenance during evolution, since sequences showing cross-hybridization with the 14 bp M13 repeat are widely conserved (in horses, cows, cats, dogs and man). A highly conserved minisatellite associated with transposable elements has also been reported (Liebermann *et al.*, 1986).

Other tandem repeats have been termed midisatellite DNA (Nakamura *et al.*, 1987a). These repeats constitute 250 - 500 kb of repetitive DNA clustered at a single locus, and show a highly polymorphic pattern in the population.

Tautz *et al.* (1986) have examined a DNA sequence library and found significantly high levels of 'cryptic simplicity' in both coding and non-coding regions. Cryptically simple regions are biased in nucleotide composition and consist of permutations of repetitive

motifs that differ within and between species. This finding suggests that ubiquitous slippage-like mechanisms are a major source of variation in all regions of the genome.

The functional and evolutionary significance of repetitive DNAs has not been established, but that they do not constitute 'junk' DNA seems increasingly likely.

1.2.2 Polymorphism

Cavalli-Sforza and Bodmer (1971) have defined polymorphism as "a condition in which the population contains at least two phenotypes (and presumably at least two genotypes) neither of which is rare - that is, neither of which occurs with a frequency of less than, say, one per cent". This implies that at least two genotypic variants are present if there is polymorphism; and this can be extended to variants in both coding and non-coding regions.

Polymorphism was originally a term applied to phenotype, but is now also applied to DNA. In this instance it refers to the difference between the base pair sequences of two chromosomes, at some locus. Many protein polymorphisms had been described and catalogued by the 1970s (Harris and Hopkinson, 1972), but the detection of nucleotide sequence variation awaited the

advent of new technologies with which DNA could be analysed directly.

DNA sequence polymorphism (DSP) is manifested as variation in single nucleotides or in oligonucleotide sequences. It can be detected in one of two ways: in the course of sequence analysis of cloned DNA fragments, or by recognition of altered restriction enzyme cleavage sites. If a cleavage site has been created or destroyed, the resultant fragment length variation is termed restriction fragment length polymorphism (RFLP). DNA polymorphisms are valuable as genetic markers, since every allele behaves as a codominant Mendelian trait and heterozygotes can be distinguished.

1.2.3 Estimates of polymorphic variation

Initial estimates of the amount of variation in the human genome came from analysis of protein variants, reflecting variation in the coding regions of DNA only, and giving no indication of the distribution of polymorphism. Cloned DNA segments have now been used as RFLP-detecting probes, and their analysis has permitted estimation of single-copy DNA sequence heterozygosity.

Jeffreys (1979) was the first to calculate genetic heterozygosity in the human genome at the DNA level by extrapolating from the frequency of RFLPs around the

γβ-globin loci. Other investigators, notably Murray *et al.* (1983) and Antonarakis *et al.* (1982), applied the same formula to the flanking regions of the albumin locus and to a larger region around the β-globin complex than that investigated by Jeffreys. All three studies demonstrated heterozygosity of at least one variable site per 100 nucleotides. Ewens *et al.* (1981) pointed out that Jeffreys' calculations produced an over-estimate of the frequency of polymorphism and recalculated a frequency of 1 in 200.

A less biased estimate of single-copy DNA sequence heterozygosity was given by Cooper and Schmidtke (1984) as 2 or 3 variable sites per 100 nucleotides, using cloned DNA segments selected at random with respect to their coding potential as probes. Cooper *et al.* (1985) later revised this to 3 or 4 per 100.

DNA estimates of heterozygosity are an order of magnitude higher than previous estimates from protein data. Jeffreys (1979) and, later, Neel (1984) suggested that discrepancies between data from DNA and from protein studies can be explained by a non-random distribution of RFLPs along the DNA. This seems likely. Poncz *et al.* (1983) compared published data on the 16.5 kb region encompassing the δβ-globin gene cluster and pointed out that 2 polymorphic sites were detected in the 885 nucleotides comprising 6 exons

(0.22%), whereas there were 36 such sites in the remaining 6 216 nucleotides (0.58%). If mutation is more or less randomly distributed throughout the DNA, then such an uneven distribution of RFLPs must reflect the action of selection, acting so that coding sequences are more conserved than non-coding sequences.

There are several sources of error in estimations of DNA heterozygosity, to date. Sample sizes have been small with respect to both the number of DNA regions examined and the number of base pairs within each region. Single-copy probes do not measure heterozygosity associated with sequences flanking repetitive elements in the genome, and this may account for observed interchromosomal differences (Aldridge *et al.*, 1984; Cooper *et al.*, 1985). Lastly, the majority of information from RFLPs has been derived from a relatively small number of enzymes and this may have introduced further bias.

1.2.4 Inter-population variation

RFLPs on the Y-chromosome can give an indication of the source of the male contribution to a particular population, whereas because of its maternal origin, polymorphisms of the mitochondrial (mt) DNA will give an indication of the female contribution to the gene pool of a particular population. Both mt- and Y-

derived RFLPs are useful for studying gene flow and the direction of such flow. The contribution of RFLPs to population genetic and anthropological studies is discussed in Section 1.3.3.

1.3 DNA sequence polymorphism (DSP)

1.3.1 Aetiology

The majority of DSPs reported have been variations in single nucleotides (Cooper and Schmidtke, 1985).

More complex DSPs are structural variants arising from the rearrangement of an oligonucleotide sequence. The rearrangement may be an insertion, deletion, inversion or duplication of DNA, giving rise to fragment lengths which vary by a constant amount if the generating enzyme cuts the DNA 5' and 3' to the rearrangement. Should the enzyme have a recognition site within the rearrangement, the DSP will be equivalent to the single base change type.

Another type of DNA structural rearrangement, tandem repetition of an oligonucleotide sequence, is found in 'hypervariable' regions (see Section 1.2.1). The length of the restriction fragment is a function of the number of copies of the tandem repeat present within the fragment.

derived RFLPs are useful for studying gene flow and the direction of such flow. The contribution of RFLPs to population genetic and anthropological studies is discussed in Section 1.3.3.

1.3 DNA sequence polymorphism (DSP)

1.3.1 Aetiology

The majority of DSPs reported have been variations in single nucleotides (Cooper and Schmidtke, 1985).

More complex DSPs are structural variants arising from the rearrangement of an oligonucleotide sequence. The rearrangement may be an insertion, deletion, inversion or duplication of DNA, giving rise to fragment lengths which vary by a constant amount if the generating enzyme cuts the DNA 5' and 3' to the rearrangement. Should the enzyme have a recognition site within the rearrangement, the DSP will be equivalent to the single base change type.

Another type of DNA structural rearrangement, tandem repetition of an oligonucleotide sequence, is found in 'hypervariable' regions (see Section 1.2.1). The length of the restriction fragment is a function of the number of copies of the tandem repeat present within the fragment.

The first hypervariable locus was described by Wyman and White (1980) and further characterized some years later (Wyman *et al.*, 1984, 1986; Balazs *et al.*, 1986). The locus is unlike the minisatellite region (Jeffreys *et al.*, 1985a) in sequence or length. Tandem repeat systems have been observed at other loci including those for the insulin gene (Bell *et al.*, 1981, 1982), the Harvey-*ras* oncogene (Capon *et al.*, 1983), the zeta-globin pseudogene (Proudfoot *et al.*, 1982) and the myoglobin gene (Jeffreys *et al.*, 1985a). New alleles are thought to be generated by alteration in the repeat number, either by slippage during DNA replication or by unequal exchange driven by the 'core' sequence, a putative recombination signal (Jeffreys *et al.*, 1985a).

A marker locus based on a single base pair variant has only two alleles. Markers with multiple alleles are obviously more useful: these are usually found in hypervariable regions or, if several two-allele markers are close to one another, haplotyping can convert a 'locus' into a highly informative marker, as has been the case for the β-globin gene cluster (Antonarakis *et al.*, 1982).

1.3.2 Detection

The standard procedures for detecting DNA sequence polymorphisms are variations of the technique developed by Southern (1975). High-molecular-weight DNA is digested with a restriction endonuclease and size fractionated by agarose gel electrophoresis. The DNA is transferred *in situ* to a solid support, which is then hybridized to a radio-labelled probe and auto-radiographed.

Statistically, there is a good chance that a given undefined DNA fragment will identify a DNA polymorphism. Based on the number of sequence polymorphisms detected in the β-globin cluster of man, it can be estimated that the human haploid genome would contain 3×10^7 sequence variants, or approximately 10^6 per chromosome (Jeffreys, 1979). Even if only ten per cent of such polymorphisms could be recognised by a restriction enzyme, there would still be 10^5 possible polymorphisms per chromosome that could be used as markers.

Single-copy probes are used to detect RFLPs, because repetitive sequences hybridise to any region of homology and could therefore reveal fragments from several loci.

The first screening strategy to develop single-copy RFLP-detecting probes was pioneered by R. White and colleagues (Wyman and White, 1980) for screening human genomic libraries (Lawn *et al.*, 1978). The strategy was extended to flow-sorted chromosomes (Davies *et al.*, 1981), and to somatic cell hybrids containing single or few human chromosomes (Davies *et al.*, 1981; Gusella *et al.*, 1980; Kao *et al.*, 1982). A probe isolated from an unknown, probably non-coding region, was referred to as 'anonymous', 'arbitrary' or 'random'.

A feature of these approaches was that most human DNA sequences of the size found in genomic libraries contain repetitive sequences, and it became necessary to isolate or subclone single-copy sequences. Helentjaris and Gesteland (1983) used cDNA probes to search for RFLPs, since cloned cDNAs are relatively free of repetitive sequences (Crampton *et al.*, 1981). cDNA probes are small, however, and can screen only a limited proportion of the genome. They also do not seem to be as efficient as genomic clones in detecting RFLPs, but may be more useful if the clones are members of gene families (Helentjaris and Gesteland, 1983).

Genomic libraries screened for the first RFLP-detecting probes were constructed in *Rec*A$^-$ hosts and it is now known that repetitive sequences are not stable in such hosts (Leach and Stahl, 1983; Wyman *et al.*, 1984; Wyman

et al., 1985; Donlon *et al.*, 1986). Since the most useful genetic markers are highly polymorphic, new strategies have been designed to detect RFLPs with a large number of morphs.

One approach to the search for multiple allelic loci used cosmids from a human genomic library as the source of probes (Litt and White, 1985). The key to the approach lay in the observation that high frequency base pair substitution polymorphisms were detected by the restriction enzymes *Msp*I and *Taq*I. It had been previously estimated that 1/20 to 1/10 of these sites are polymorphic in man (Barker *et al.*, 1984). Cosmid vectors containing 35- to 45-kb inserts (Collins and Hohn, 1978) were hybridized to Southern transfers after prehybridization with a vast excess of non-radioactive sonicated total human DNA. This strategy prevented regions of probes homologous to repetitive genomic DNA from hybridyzing. By using a long, effectively single-copy probe, more potentially polymorphic sites could be examined. Litt and White (1985) and Litt *et al.* (1986a) used this strategy to isolate three single-copy probes from one cosmid that collectively identified seven linked polymorphic loci.

Another strategy uses the same 'core' sequence from the myoglobin hypervariable region that reveals genetic fingerprints in total human DNA to screen DNA libraries

and to identify clones representing unique loci (Wong *et al.*, 1986b). Similar candidate and synthetic sequences have been used as probes of genomic libraries and new unique VNTR loci have been identified (Nakamura *et al.*, 1987a, 1987b).

Detection of DSPs requires the selection of appropriate restriction enzymes as well as probes. It has been suggested that RFLPs are detected at a higher frequency in restriction enzyme recognition sites that contain the dinucleotide CpG (Barker *et al.*, 1984). The human genome is heavily methylated at this dinucleotide (Van der Ploeg and Flavell, 1980; Cooper, 1983) and 5-methylcytosine is subject to frequent replacement by thymidine due to the de-amination of the methylated base (Coulondre *et al.*, 1978). This explains both the high frequency of the C to T transition (Vogel, 1972; Vogel and Kopun, 1977) and the resulting CpG deficiency observed in vertebrate genomes (Salser, 1977; Bird, 1980).

Other workers have disagreed with this hypothesis and Wijsman (1984) argues that some of the increased frequency is accounted for by the expected distribution of fragment lengths. Cooper and Schmidtke (1984) suggest that because enzymes with CpG in their recognition sites are used more often they are found to detect RFLP more often. Gitschier *et al.* (1985) has

however found an increased frequency of polymorphism with *Taq*I and *Msp*I, and it seems reasonable that enzymes with recognition sites containing CpG should be used in any search for RFLP.

1.3.3 Usefulness

DNA polymorphisms have revolutionised human genetic analysis and have found general use in the mapping of human linkage groups, in the indirect localization of genetic disease loci by linkage, in prenatal and preclinical diagnosis, and in anthropological studies.

Gene mapping studies

RFLPs were first used by Grodzicker *et al.* (1974) as a means of physically mapping temperature-sensitive mutations of adenoviruses. Their usefulness in mapping the human genome was recognised by Botstein *et al.* (1980), who proposed the construction of a linkage map such that every locus can be linked to an RFLP. Potentially large numbers and the codominant expression of the 'alleles' in DNA polymorphisms make this task feasible.

The value of RFLPs and other polymorphic markers for human gene mapping lies in the fact that they need not encompass the disease locus in order to be useful, and

that the gene product does not need to be isolated in order to be mapped. Of the nearly 3 500 single gene disorders described in the Sixth Edition of McKusick's catalogues (1983), abnormal enzyme defects have been identified in 213, and abnormalities of non-enzyme proteins in a further 106, representing less than 10 per cent of all diseases. Gene mapping of the majority of disease loci will be achieved by linkage analysis in which RFLPs will play a major role. The mapping of the human genome is reviewed in Section 1.4.

Medical applications

DSPs have been widely used in medical studies. The haemoglobinopathies were the first to be characterised at the molecular level, and a number of RFLPs associated with genes have now been described. The term "reverse genetics" has been coined to describe the isolation of a gene without reference to a specific protein or without any reagents or functional assays in its detection. RFLPs can now be used for the prenatal or preclinical diagnosis of, among others, Duchenne muscular dystrophy (DMD) (Murray *et al.*, 1982; Davies *et al.*, 1983), Hungtington's Chorea (Gusella *et al.*, 1983), phenylketonuria (Woo *et al.*, 1983), retinoblastoma (Benedict *et al.*, 1983), haemophilia A (Gitschier *et al.*, 1985) and B (Camerino *et al.*, 1984), and cystic fibrosis (Knowlton *et al.*, 1985; White *et al.*,

1985; Wainwright *et al.*, 1985). Linkage between a marker on chromosome 11p and manic depressive psychosis in the Amish has recently been postulated (Egeland *et al.*, 1987), and this is the first molecular evidence for an underlying genetic cause of a psychiatric disorder.

Once linkage between polymorphic markers and a disease locus has been established, attempts to clone the gene itself can be made. The use of linked markers in conjunction with techniques such as chromosome walking (Steinmetz *et al.*, 1982) or jumping (Smith and Cantor, 1986b) has enabled the isolation of such loci. A portion of the DMD gene has been cloned (Kunkel *et al*, 1985; Kunkel *et al.*, 1986; Monaco *et al.*, 1986), as has a candidate for the cystic fibrosis gene (Estivall *et al.*, 1987). The strategy for finding the latter locus made use of the close association which has been found between non-methylated, CpG-rich regions (HTF islands) and the 5' ends of expressed genes (Bird, 1986). The most tractable disorders to finding the locus concerned, using molecular approaches, have so far been those highlighted by cytological deletions or translocations, associated with unusually large genes, or frequently disrupted by structural aberrations, as in Wilms' tumour or retinoblastoma.

The applications of RFLPs to human problems with medical import are numerous. RFLPs have been used to analyse the insulin locus (Bell *et al.*, 1981, 1982, 1984) with the intention of correlating the sequence polymorphism with diabetes, but a relationship has not yet been established. Other workers have used a collection of arbitrary loci from a chromosome 13-enriched library (Cavanee *et al.*, 1984; Drija *et al.*, 1984) to elucidate some of the genetic mechanisms involved in retinoblastoma. Yet others (Koufos *et al.*, 1984; Orkin *et al.*, 1984; Reeve *et al.*, 1984; Fearon *et al.*, 1984) have used DNA polymorphisms at both known and arbitrary loci to localise the Wilms' tumour locus and to analyse the role of mitotic non-disjunction and recombination in inherited cancer. X-linked RFLPs can be used to detect the clonal origin of tumours (Vogelstein *et al.*, 1985) and RFLPs are useful in assessing graft-versus-host reactions and the recurrence of leukaemia after bone marrow transplants (Knowlton *et al.*, 1986).

The restriction fragment pattern revealed by the sum of the VNTR loci containing related sequences, scattered throughout the genome, can constitute a somatically stable genetic 'fingerprint' unique to an individual (Jeffreys *et al.*, 1985a). Such fingerprints can be applied directly to problems of human identification including paternity testing and forensic analyses, as